52èmes Journées de Statistiques de la Société Française de Statistique (SFdS)

Recueil des soumissions

Table des matières

Analyse en Composantes Principales Topologique

Rafik Abdesselam

Laboratoires ERIC - COACTIS, Université de Lyon, Lumière Lyon 2,
16, quai Claude Bernard 69365 Lyon cedex 07
rafik.abdesselam@univ-lyon2.fr

L'objectif de ce papier est de proposer une approche topologique d'analyse des données qui consiste à explorer, analyser et représenter la structure des corrélation d'un ensemble de variables quantitatives dans un contexte d'analyse en composantes principales. Les mesures de similarité jouent un rôle important dans de nombreux domaines de l'analyse des données. Les résultats de toute opération de structuration, de classification ou de classement d'objets dépendent fortement de la mesure de proximité choisie. Basées sur la notion de graphes de voisinage, certaines de ces mesures de proximité sont plus ou moins équivalentes. La notion d'équivalence topologique entre deux mesures est définie et statistiquement testée selon la description des corrélations entre les variables. Un exemple sur données réelles illustre cette approche topologique.

Mots-clés. Mesure de proximité, graphe de voisinage, matrice d'adjacence, équivalence topologique, corrélation, MDS représentations graphiques.

Abstract. The objective of this paper is to propose a topological approach of data analysis that consists in exploring, analyzing and representing the correlation between a set of quantitative variables in a context of principal component analysis. Similarity measures play an important role in many areas of data analysis. The results of any operation of structuring, clustering or classifying objects depend strongly on the proximity measure chosen. Based on the notion of neighborhood graphs, some of these proximity measures are more or less equivalents. The notion of topological equivalence between two measures is defined and statistically tested according to the description of the correlations between the variables. An example on real data illustrates this topological approach.

Keywords. Proximity measure, neighborhood graph, adjacency matrix, topological equivalence, correlation, MDS graphical representations.

1 Introduction

Le choix d'une mesure de proximité est un problème important en analyse des données topologique. La comparaison d'objets, de situations ou d'idées est une tâche essentielle pour évaluer une situation, classer des préférences ou structurer un ensemble d'éléments. Pour ce faire, nous utilisons des mesures de proximité pour mettre en évidence les similarités ou les dissimilarités entre objets. Nous savons pertinemment que le résultat dépend

de la mesure utilisée. Laquelle est alors la plus utile ? Sont-elles équivalentes ? Comment identifier celle qui est la plus appropriée pour résumer la structure des corrlations d'un ensemble de variables quantitatives ? Selon la mesure choisie, les résultats de cette problématique d'analyse en composantes principales topologique changent.

De nombreux travaux sur les mesures de proximité, l'équivalence topologique entre mesures de proximité (Batagelj et Bren (1995), Rifqi *et al.* (2003), Lesot *et al.* (2009), Zighed *et al.* (2012)), ainsi que sur des approches topologiques d'analyses des correspondances (Abdesselam (2019)) et d'analyse discriminante (Abdesselam (2019)) ont été proposée, mais pas dans un objectif de synthèse topologique des corrélations d'un ensemble de variables quantitatives homogènes.

Nous avons considéré et comparé 15 mesures de proximité les plus utilisées pour des données continues (Warrens (2008)).

2 Equivalence topologique

L'équivalence topologique repose sur la notion de graphe de voisinage. Deux mesures de proximité sont équivalentes si les graphes topologiques induits sur l'ensemble des objets restent identiques. Mesurer la ressemblance entre mesures de proximité revient à comparer leurs graphes de voisinage.

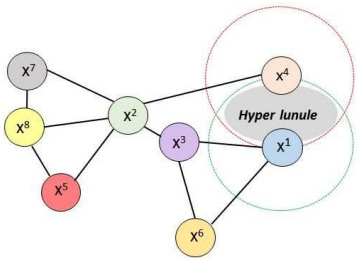
Soit l'ensemble $E = \{x^j \ ; \ j = 1, \cdots, p\}$ à $|E| = p$ objets dans R^n, associé aux p variables quantitatives décrivant un ensemble de n individus. On peut à l'aide d'une mesure de proximité u, définir une relation de voisinage V_u qui sera une relation binaire sur $E \times E$.

Pour une mesure de proximité donnée u, on peut construire un graphe de voisinage sur l'ensemble des objets-variables où les sommets sont les variables et les arêtes sont définies par une relation de voisinage. Il existe de nombreuses définitions pour construire cette relation binaire de voisinage, par exemple, l'Arbre de Longueur Minimale, le Graphe de Gabriel, ou encore le Graphe des Voisins Relatifs (GVR) (Toussaint (1980)), dont les couples de points voisins (x^k, x^l), où k,l = 1,p, vérifient la propriété GVR suivante :

$$\begin{cases} V_u(x^k, x^l) = 1 & si\ u(x^k, x^l) \leq \max[u(x^k, x^r), u(x^r, x^l)] \ ; \ \forall x^k, \ x^l, \ x^r \in E, x^r \neq x^k \ \ et \ \ x^r \neq x^l \\ V_u(x^k, x^l) = 0 & sinon \end{cases}$$

Pour toute mesure de proximité donnée u, on peut lui associer une matrice dite d'adjacence V_u binaire et symétrique d'ordre p. La figure 1 illustre un ensemble de $p = 8$ objets-variables.

Par exemple, pour les variables x^1 et x^4, $V_u(x^1, x^4) = 1$, signifie sur le plan géométrique, que l'hyper-Lunule, intersection des deux hypersphères centrées sur les deux points-variables x^1 et x^4, est vide. Ainsi, si deux variables x^k et x^l vérifient la définition 1, elles sont reliées par une arête directe, les sommets x^k et x^l sont voisins.



V_u	x^1	x^2	x^3	x^4	x^5	x^6	x^7	x^8
x^1	1	0	0	1	0	0	0	0
x^2	0	1	1	0	0	0	0	0
x^3	0	1	1	0	0	0	0	0
x^4	1	0	0	1	0	0	0	0
x^5	0	0	0	0	1	0	0	1
x^6	0	0	0	0	0	1	0	0
x^7	0	0	0	0	0	0	1	1
x^8	0	0	0	0	1	0	1	1

Figure 1: Exemple de GVR avec huit variables - Matrice d'adjacence associe

Comparaison et séléction de mesures de proximité

Pour mesurer l'équivalence topologique entre deux mesures de proximité u_i et u_j, nous proposons de tester si les matrices d'adjacence associées V_{u_i} et V_{u_j} sont différentes ou pas. L'équivalence topologique entre deux matrices d'adjacence est mesurée par la relation de concordance suivante :

$$S(V_{u_i}, V_{u_j}) = \frac{1}{p^2} \sum_{k=1}^{p} \sum_{l=1}^{p} \delta_{kl}(x^k, x^l) \text{ avec } \delta_{kl}(x^k, x^l) = \begin{cases} 1 & \text{si } V_{u_i}(x^k, x^l) = V_{u_j}(x^k, x^l) \\ 0 & \text{sinon.} \end{cases}$$

La mesure de similarité $S(V_{u_i}, V_{u_j}) = 1$ signifie que les deux matrices d'adjacence sont identiques et par conséquent, la structure topologique induite par les deux mesures est la même. Dans ce cas, on parle d'équivalence topologique parfaite entre les deux mesures de proximité. La valeur $S(V_{u_i}, V_{u_j}) = 0$ signifie que la topologie a totalement changé.

On construit la matrice d'adjacence notée $V_{u_\star}$, qui correspond au mieux à la structure de la matrice de corrélations selon le t-test de significativité de la corrélation linéaire de Bravais-Pearson :

$$\begin{cases} V_{u_\star}(x^k, x^l) = 1 & si \ \ p\text{-}value = P[\mid T_{n-2} \mid > t\text{-}value] \leq 5\% \ \ ; \ \ \forall k, l = 1, p \\ V_{u_\star}(x^k, x^l) = 0 & sinon \end{cases}$$

Cette matrice d'adjacence binaire et symétrique V_{u*} dite de référence, est associée à une mesure de proximité de référence, inconnue, notée u_*.

On peut ainsi établir l'équivalence topologique $S(V_{u_i}, V_{u*})$ entre les mesures u_i et $u_\star$, en mesurant la similarité entre les matrices d'adjacence V_{u_i} et V_{u*}.

Pour visualiser les mesures de proximité, on peut par exemple appliquer la technique du thémascope, qui consiste un enchaînement méthodologique d'une méthode de classification sur les résultats d'une méthode factoriel, dans ce cas, une Classification Ascendante Hiérarchique (CAH) selon le critère de Ward (Ward (1963)) sur les facteurs significatifs de l'Analyse en Composantes Principales (ACP) du tableau des dissimilarités $[D]_{ij} = 1 - S(V_{u_i}, V_{u_j})_{i,j=1,15}$. De plus, pour déterminer la classe de mesures de proximité la plus proche de la mesure de référence $u_\star$, cette dernière sera considérée comme élément illustratif dans les analyses, en projetant *a posteriori* le vecteur de dissimilarité $[D]_{\star i} = 1 - S(V_{u_\star}, V_{u_i})_{i=1,15}$.

Test statistique de l'équivalence topologique

Le test exact de Fisher (Fisher (1922)) est bien adapté aux données binaires ou nominales lorsque les tailles d'échantillon sont petites. Ce test ne repose pas sur une statistique dont la loi est connue lorsque n est assez grand mais il calcule, comme son nom l'indique, la p-value exacte directement. Nous utilisons ce test non paramtrique pour mesurer le degré d'équivalence topologique entre deux mesures de proximité à partir de leurs matrices d'adjacence associées, savoir si elles sont statistiquement différentes ou pas. Deux mesures sont statistiquement en équivalence topologique si l'hypothèse nulle H_0 d'indépendance est rejetée.

Il s'agit d'un test de proportion sur deux chantillons indpendants. Ces matrices binaires et symétriques d'ordre p, sont dépliées selon deux vecteurs de composantes appariées, formées des $\frac{p(p-1)}{2}$ valeurs supérieures (ou inférieures) de la diagonale. Le degré d'équivalence topologique entre les deux mesures u_i et u_j est évalué et testé à partir du test exact de Fisher, calculé sur la table 2×2 de contingence formé par les deux vecteurs binaires. Nous testons également l'équivalence topologique entre chacune des 15 mesures de proximité u_i et la mesure de référence $u_\star$ a partir des matrices d'adjacence V_{u_i} et $V_{u_\star}$.

Représentations graphiques

Afin de visualiser les relations topologiques entre les p variables, nous utilisons le positionnement multidimensionnel (MDS) classique de la matrice d'adjacence $V_{u_\star}$ associée à la mesure de proximité $u_\star$, la mesure la plus adaptée aux données considérées, savoir l'ACP du triplet $\{V_{u_\star}\ ;\ M\ ;\ D_p\}$ où $V_{u_\star}$ est la matrice d'adjacence associée à la mesure de proximité $u_\star$, la mesure la plus adaptée aux données considérées, $M = I_p$ et $D_p = \frac{1}{p} I_p$ avec I_p, la matrice identité d'ordre p.

L'ACPT peut être effectuée à partir de n'importe quelle matrice d'adjacence V_{u_i} associée à chacune des 15 mesures de proximité u_i considérées. Quant aux reprsentations des individus actifs, ces derniers sont projetés comme éléments illustratifs.

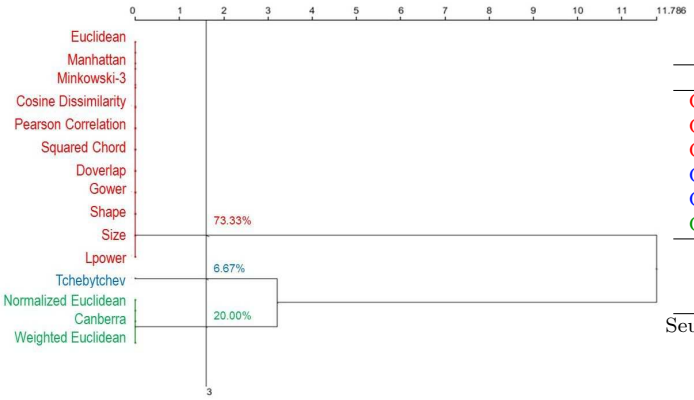
3 Exemple illustratif

Pour illustrer l'approche proposée, nous utilisons ici les données d'eurostat relatives aux finances publiques des 28 pays de l'Union Européenne (UE-28) en 2017, définies par quatre composantes exprimées en pourcentage du PIB : la Dette brute, le Déficit public, les Recettes et les Dépenses totales. L'objectif ici, est de donner une synthèse topologique des finances publiques de l'UE-28.

Dans un contexte métrique et classique, il suffit d'appliquer une ACP sur l'ensemble homogène des 4 caractristiques des finances publiques de l'UE-28.

Dans le contexte topologique considéré, les principaux résultats de l'approche proposée sont donnés dans les tableaux et graphiques ci-dessous.

On a comparé les 15 mesures de proximité consdérées, testé leur équivalence topologique et visualisé la structure des corrélations des finances publiques de l'UE-28.



u_i	u_j	$S(u_i\ ,\ u_j)$	$p-value$
Classe 1	Classe 1	1.0000	0.0083**
Classe 1	Classe 2	0.7500	0.1833
Classe 1	Classe 3	0.7500	0.1833
Classe 2	Classe 2	1.0000	0.0083**
Classe 2	Classe 3	0.5000	1.0000
Classe 3	Classe 3	1.0000	0.0083**
$u_\star$	Classe 1	0.7500	0.1833
$u_\star$	Classe 2	0.6250	0.5000
$u_\star$	Classe 3	0.8750	0.0333*

Seuil de signification α ; $^{**}\alpha \leq 1\%$; $^{*}\alpha \in]1\%; 5\%]$

Figure 2: Arbre hiérarchique - Similarité - Test exact de Fisher

Le dendrogramme de la figure 2 résume les principaux résultats de la partition des 15 mesures de proximité considérées et illustre les trois classes homogènes de mesures retenues. La mesure de référence $u_\star$ projetée en élément supplémentaire, est affectée à la classe 3 constituée des mesures de Canberra et Euclidiennes normalisée et pondérée. Ce sont là les mesures de proximité les plus adaptées pour l'analyse topologique de la structure des corrélations des finances publiques de l'UE-28.

A noter dans tableau de la figure 2, pour les données considérées, les mesures de proximité qui constituent chacune des 3 classes de la partition sont en parfaite équivalence topologique, les similarités sont égales 1, on obtiendrait donc des résultats identiques.

Classe 2 Tchebytchev	Classe 1 : Euclidean $V_{u_2}=0$	$V_{u_2}=1$
$V_{u_1}=0$	2	1
$V_{u_1}=1$	1	6

$S(V_{u_2}\ ,\ V_{u_1}) = 0.75$; $p\text{-}value = 0.1833$

Classe 3 Canberra	Classe 2 : Tchebytchev $V_{u_2}=0$	$V_{u_2}=1$
$V_{u_1}=0$	1	2
$V_{u_1}=1$	2	5

$S(V_{u_3}\ ,\ V_{u_2}) = 0.50$; $p\text{-}value = 1.000$

Classe 1 Euclidean	Classe 3 : Canberra $V_{u_2}=0$	$V_{u_2}=1$
$V_{u_1}=0$	2	1
$V_{u_1}=1$	1	6

$S(V_{u_1}\ ,\ V_{u_3}) = 0.75$; $p\text{-}value = 0.1833$

Mesure Référence	Classe 1 : Euclidean $V_{u_2}=0$	$V_{u_2}=1$
$V_{u_\star}=0$	3	1
$V_{u_\star}=1$	0	6

$S(V_{u_\star}\ ,\ V_{u_1}) = 0.750$; $p\text{-}value = 0.183$

Mesure Référence	Classe 2 : Tchebytchev $V_{u_2}=0$	$V_{u_2}=1$
$V_{u_\star}=0$	2	2
$V_{u_\star}=1$	1	5

$S(V_{u_\star}\ ,\ V_{u_2}) = 0.625$; $p\text{-}value = 0.500$

Mesure Référence	Classe 3 : Canberra $V_{u_2}=0$	$V_{u_2}=1$
$V_{u_\star}=0$	3	1
$V_{u_\star}=1$	0	6

$S(V_{u_\star}\ ,\ V_{u_3}) = 0.875$; $p\text{-}value = 0.0333*$

Seuil de signification α ; $^{**}\alpha \leq 1\%$; $^{*}\alpha \in]1\%; 5\%]$

Table 1: Table 2×2 - Similarité - Test exact de Fisher

Le tableau 1 illustre les tables de contingence 2×2 entre les mesures de proximité de chaque classe : Euclidienne, Tchebytchev, Canberra et la mesure de référence $u_\star$ pour le calcul du test exact de Fisher. Seule l'équivalence topologique entre la mesure de référence et la mesure de Canberra est significative, p$-$value $= 0.0034 < \alpha = 5\%$, l'hypotèse nulle H_0 d'indépendance est rejetée.

La matrice d'adjacence $V_{u_\star}$ associée à la mesure de proximité $u_\star$ adaptée aux données considérées, est construite à partir des p−values du t-test des coefficients de corrélations de la matrice de la figure 3.

Variables	Dette	Déficit	Recettes	Dépenses
Dette	1.000			
Déficit	-0.3403	1.000		
	(0.0764)			
Recettes	0.3071	0.0393	1.000	
	(0.1120)	(0.8428)		
Dépenses	0.3845	-0.2092	0.9689	1.000
	(**0.0434***)	(0.2853)	(**0.0001****)	

Seuil de signification α ; $^{**}\alpha \leq 1\%$; $^{*}\alpha \in]1\%; 5\%]$

$$V_{u_*} = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 \end{pmatrix}$$

Figure 3: Matrice des corrélations (p-value) - Matrice d'adjacence ACPT

La figure 4 présente, à titre de comparaison sur le premier plan factoriel, les corrélations entre les facteurs et les variables d'origine de l'ACPT et de l'ACP. Ces représentations graphiques des variables sont légèrement différentes. Nous pouvons ainsi représenter l'analyse topologique de chacune des 15 mesures de proximité considérées.

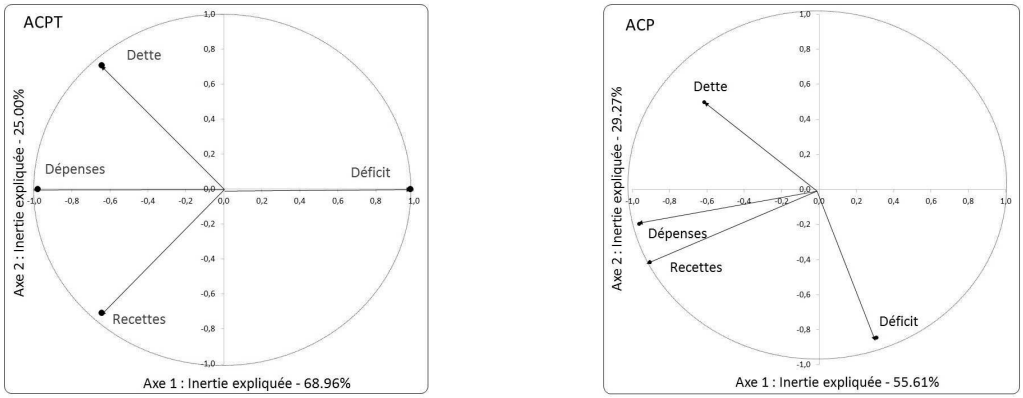


Figure 4: ACPT - ACP : Représentations des finances publiques de l'UE-28

4 Conclusion & Perspective

Ce travail propose une nouvelle approche qui permet de synthétiser et de décrire un ensemble de variables quantitatives dans un contexte topologique. Comme l'ACP, l'ACPT proposée est une méthode exploratoire topologique multidimensionnelle qui peut être utile pour la réduction des dimensions, elle enrichit les méthodes conventionnelles d'analyse de données continues. Il serait intéressant d'étendre cette approche topologique dans le cadre de l'analyse des données évolutives.

Bibliographie

[1] Abdesselam, R. (2019), A Topological Multiple Correspondence Analysis. *Journal of Mathematics and Statistical Science*, Science Signpost Publishing Inc., USA, 5, 8, 175-192.
[2] Abdesselam, R. (2019), A Topological Discriminant Analysis. *In book Chapter, Data Analysis and Applications 2: Utilization of Results in Europe and Other Topics*, ISTE Science Publishing LTD, Wiley, 3, 167–178.
[3] Batagelj, V., Bren, M. (1995) Comparing resemblance measures. *In Journal of classification*, 12, 73–90.
[4] Fisher, R-A. (1922), The Interpretation of χ^2 from Contingency Tables, and the Calculation of P. *Journal of the Royal Statistical Society*, Published by Wiley, 85, 1, 87–94.
[5] Lesot, M-J., Rifqi, M. and Benhadda,H. (2009) Similarity measures for binary and numerical data: a survey. *In IJKESDP*, 1, 1, 63-84.
[6] Rifqi, M., Detyniecki, M. and Bouchon-Meunier, B. (2003) Discrimination power of measures of resemblance. *IFSA'03 Citeseer*.
[7] Toussaint, G. T. (1980), The relative neighbourhood graph of a finite planar set. *In Pattern recognition*, 12, 4, 261–268.
[8] Ward, J-R. (1963) Hierarchical grouping to optimize an objective function. *In Journal of the American statistical association JSTOR*, 58, 301, 236–244.
[9] Warrens, M-J. (2008), Bounds of resemblance measures for binary (presence/absence) variables. In Journal of Classification, *Springer*, 25, 2, 195–208.
[10] Zighed, D., Abdesselam, R., and Hadgu, A. (2012), Topological comparisons of proximity measures. *16th PAKDD 2012 Conference*, Part I, LNAI 7301, Springer, 379–391.

Détection d'anomalies dans des données fonctionnelles multivariées

Martial AMOVIN-ASSAGBA[1,2], Julien JACQUES[2], Irène GANNAZ[3], Frédéric FOSSI[1] & Johann MOZUL[1]

[1] *Arpege Master K, Saint-Priest, France, martial.amovin@masterk.com, frederic.fossi@masterk.com & johann.mozul@masterk.com*
[2] *Univ Lyon, Lyon 2, ERIC EA3083, Lyon, France, julien.jacques@univ-lyon2.fr*
[3] *Univ Lyon, INSA de Lyon, CNRS UMR 5208, Institut Camille Jordan, F-69621 Villeurbanne, France, irene.gannaz@insa-lyon.fr*

Résumé. L'objectif de ce travail est de détecter des anomalies dans les données fonctionnelles multivariées provenant d'appareils de mesure, dans une optique de maintenance prédictive. Des méthodes statistiques comme le clustering fonctionnel et l'estimation linéaire par morceaux ont été testées. Nous montrons l'intérêt de ces méthodes ainsi que leurs insuffisances.

Mots-clés. données fonctionnelles, clustering, modélisation linéaire par morceaux

Abstract. This work aims to detect anomalies in multivariate functional data coming from measuring devices. Statistical methods such as functional clustering and piecewise linear estimation were tested. We show the interest of these methods as well as their lackness.

Keywords. functional data, clustering, piecewise linear modelling

1 Introduction

Nous disposons de mesures temporelles provenant simultanément de divers capteurs. Notre objectif est de détecter des anomalies dans ces données fonctionnelles multivariées. De nombreuses techniques d'apprentissage non supervisé pour la détection des anomalies existent (Chandola et al 2009), mais très peu sont adaptées aux données fonctionnelles (Ramsay et Silverman 2005). Certaines méthodes se basant sur les fonctions de profondeur sont proposées pour les données fonctionnelles univariées (López-Pintado et Romo 2009; Febrero et al 2008), d'autres dans le cadre multivarié (Hubert et al 2015; Dai et Genton 2018). Hubert et al (2015) utilisent le demi-espace de profondeur pour mesurer la "centralité" d'une courbe alors que Dai et Genton (2018) définissent une matrice de décalage en étendant le décalage directionnel.

On distingue plusieurs types d'anomalies quand il s'agit des données fonctionnelles: anomalie de forme, de localisation, etc. Dans l'application qui nous intéresse, nous sommes

confrontés principalement à des anomalies de forme. Si bon nombre d'approches se basant sur les fonctions de profondeur sont fiables dans la détection des anomalies de localisation, elles échouent souvent à identifier ce type de données aberrantes.

Une autre façon de détecter des anomalies pourrait être d'utiliser des techniques de clustering fonctionnelles multivariées (Schmutz et al 2020), qui pourraient permettre d'isoler des clusters de données atypiques.

Afin de détecter des anomalies, nous avons testé sur les données un algorithme de clustering fonctionnel, que nous avons comparé avec une autre approche paramétrique (estimation linéaire par morceaux) basée sur la forme des courbes. Nous présentons les résultats obtenus et leurs limites.

2 Les données

Nous considérons un jeu de données de taille 509, issu d'un matériel comportant 4 capteurs, sachant que nous avons aussi d'autres matériels qui ont plus de 4 capteurs. Une mesure est constituée de 4 courbes. Le nombre de points de mesure des trajectoires diffère d'une observation à l'autre. Il varie entre 2199 et 10675 avec une médiane égale à 5144. Puisque les données sont discrétisées sur des grilles fines, une approche de type données fonctionnelles est privilégiée (Ramsay and Silverman 2005). Nous représentons quelques données sur la Figure 1.

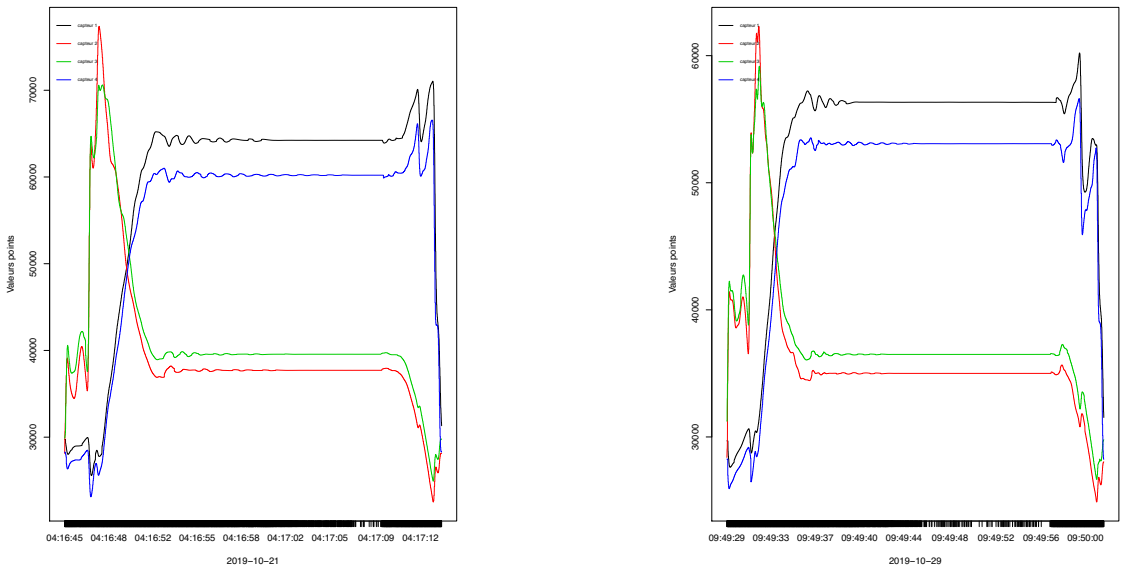


Figure 1: Exemples de courbes des instruments de mesure

D'une donnée normale à l'autre, il y a souvent une différence assez importante en

amplitude et en temps. Cette différence pourrait biaiser les résultats. Par conséquent, nous normalisons au préalable les données en amplitude en les divisant par une valeur moyenne relative aux courbes, et en temps en les ramenant dans l'intervalle $[0, 1]$.

Des analyses d'experts ont permis d'identifier deux mesures comme anormales. La première (notée "Anormale1") provient de la défaillance du Capteur 2 alors que la seconde (notée "Anormale2") est liée à un effet extérieur. Cette dernière n'est pas tout à fait considérée comme aberrante du point de vue de l'expert métier. Notre objectif est de proposer une procédure automatisée permettant sur cette base de données de retrouver ces deux données.

3 Clustering

Afin de détecter un ensemble de données atypiques liées à des défaillances des appareils de mesure, une caractérisation pourra se faire à partir d'une analyse non supervisée des historiques de mesures, de type clustering de données fonctionnelles multivariées. Notre objectif est d'identifier des clusters qui caractérisent les groupes de données atypiques.

Récemment Schmutz et al, ont proposé une nouvelle méthode de clustering fonctionnel multivarié (`funHDDC`) afin de permettre d'identifier des groupes d'individus homogènes. S'agissant des données fonctionnelles, la principale source de difficulté réside dans le fait que les courbes sont censées appartenir à un espace dimensionnel infini, alors qu'en pratique nous disposons d'échantillons observés en un ensemble de points finis. Les auteurs reconstituent la forme fonctionnelle des données en lissant les observations dans une base de fonctions de dimension finie. Leur méthode s'appuie ensuite sur un modèle de mélange latent fonctionnel.

Nous appliquons dans un premier temps cet algorithme dans le cas multivarié où nous prenons en compte toutes les courbes quel que soit l'appareil de mesure. Dans un second temps nous testons le cas univarié où nous ne considérons que les courbes d'un même capteur. Le package `funHDDC` de Schmutz et al propose plusieurs modèles parcimonieux. Il propose également 5 moyens d'initialisation de l'algorithme E-M. Nous avons testé tous les modèles, tout en variant le nombre maximum d'itérations et les initialisations de l'algorithme E-M.

En multivarié, un seul cluster est obtenu à chaque fois: le modèle de clustering n'arrive pas à former des groupes suffisamment distincts les uns des autres. Par contre dans le cas univarié, quel que soit le capteur considéré, le modèle retenu présente au moins 3 clusters. Nous représentons sur la Figure 2 les courbes moyennes des 5 clusters obtenus avec les données du capteur 2, puisque c'est le seul capteur ayant eu une défaillance. Bien qu'identifiant plusieurs clusters, l'algorithme n'arrive pas à distinguer un groupe spécifique de courbes anormales. Aucun cluster n'a de forme atypique d'un point de vue de l'expert métier. Les courbes identifiées auparavant comme anormales sont dans un même cluster que des courbes normales. Ceci est probablement dû au trop faible nombre de données

atypiques. La Figure 3 présente les courbes de ce cluster, avec une distinction en couleur des courbes supposées anormales.

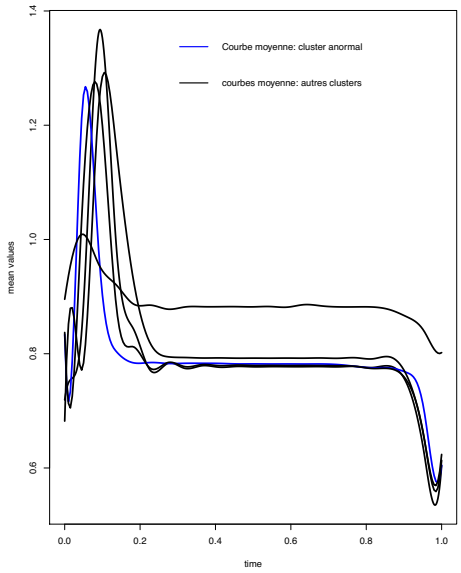


Figure 2: Courbes moyennes des 5 clusters formés par l'ensemble des courbes du capteur 2, cas univarié

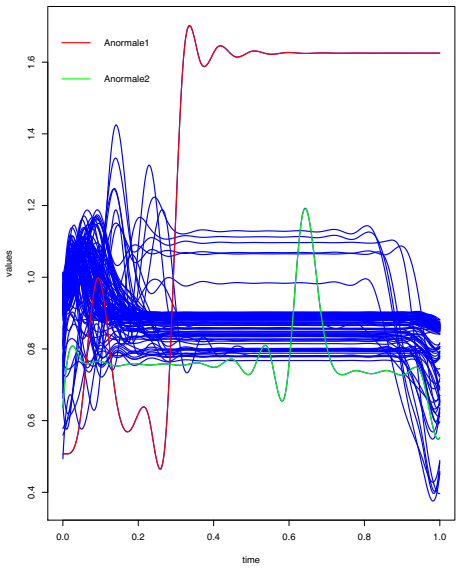


Figure 3: Ensemble des courbes du cluster contenant les données anormales du capteur 2, cas univarié

Avec cette méthode, il n'est donc pas possible de faire une caractérisation des données anormales. Nous ne pouvons donc pas détecter les courbes aberrantes du fait que l'algorithme les met dans un même cluster que les courbes normales. De plus cette approche présente des difficultés du fait que les données sont de tailles différentes. Nous pensons plus tard coupler cette méthode avec une étude de la distribution du temps et de l'amplitude des courbes.

4 Estimation linéaire par morceaux

Dans un second temps, nous avons décidé de nous appuyer sur la forme spécifique des courbes de notre application. Nous avons identifié une allure type pour chaque appareil de mesure. Nous définissons donc un patron pour chaque capteur. Ce patron forme une base de fonctions particulières (linéaires par morceaux, cf. Figure 4) dans laquelle nous lissons les données.

Chaque observation est approchée dans cette base linéaire par morceaux (Muggeo 2017) à l'aide du package R `segmented`. Muggeo définit des modèles de régressions avec

des relations segmentées entre la réponse et la variable tout en estimant les points de rupture. La Figure 4 présente une courbe du deuxième appareil de mesure avec sa reconstruction linéaire par morceaux (patron).

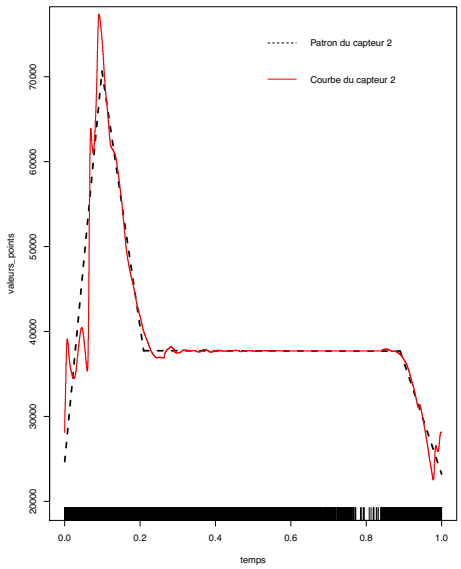


Figure 4: Exemple de patron : cas du capteur 2

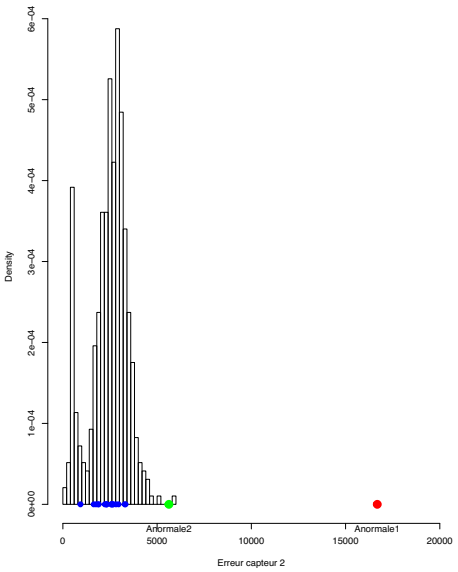


Figure 5: Histogramme des erreurs de reconstruction du capteur 2

Après avoir approché chaque mesure dans la base linéaire par morceaux définissant le patron d'une mesure normale, nous regardons la distribution empirique des erreurs d'approximations (écarts entre les vraies valeurs et les valeurs estimées). Il est alors possible de se baser sur cette distribution empirique pour proposer une évaluation de la probabilité qu'une nouvelle mesure soit atypique.

La Figure 5 présente l'histogramme des erreurs de reconstruction des courbes du capteur 2. Les points en rouge et vert représentent respectivement les erreurs de reconstruction des données défaillantes "Anormale1" et "Anormale2". Visiblement, le point rouge est très éloigné de l'ensemble des points de l'histogramme. La valeur réelle de cette erreur est 16681.9, soit environ 3 fois le maximum des erreurs de reconstruction des courbes normales provenant du capteur 2. Cette méthode nous permet de détecter la donnée atypique issue de la défaillance du capteur 2. Elle permet également de distinguer la seconde anomalie, provenant de la défaillance d'un capteur issue d'une variable exogène ou d'un effet extérieur.

Afin de voir s'il y a un signe précurseur à la défaillance du capteur 2, nous représentons les erreurs de reconstruction de quelques données avant la panne. Ces points (en bleu) sont tous considérés comme normaux suivant la distribution empirique des erreurs. Avec

cette méthode, il ne semble donc pas y avoir de signes précurseurs à la panne, pour cette défaillance.

5 Conclusion

Dans ce travail préliminaire, nous avons appliqué deux méthodes paramétriques pour détecter des anomalies, un clustering fonctionnel multivarié et une méthode basée sur la distribution des erreurs de reconstruction par une base de fonctions linéaires par morceaux. Seule la seconde méthode a permis de détecter la donnée atypique provenant d'un appareil de mesure défaillant. L'algorithme de clustering fonctionnel utilisé nous renvoie un cluster composé d'un mélange de données anormales et normales. Une difficulté en utilisant l'approche fonctionnelle est que les données ne sont pas de même taille. Nous pensons coupler l'approche de clustering fonctionnel avec une étude de la distribution du temps et de l'amplitude de la donnée. Un autre point important est de considérer des métriques basées sur les dérivées. Nous les utilisons déjà avec la méthode FIF : Functional Isolation Forest (Stearman et al 2019).

Bibliographie

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3), 1-58.

Dai, W., & Genton, M. G. (2018). An outlyingness matrix for multivariate functional data classification. *Statistica Sinica*, 28(4), 2435-2454.

Febrero, M., Galeano, P., & Gonzàlez-Manteiga, W. (2008). Outlier detection in functional data by depth measures, with application to identify abnormal NOx levels. *Environmetrics: The official journal of the International Environmetrics Society*, 19(4), 331-345.

Hubert, M., Rousseeuw, P. J., & Segaert, P. (2015). Multivariate functional outlier detection. *Statistical Methods & Applications*, 24(2), 177-202.

López-Pintado, S., & Romo, J. (2009). On the concept of depth for functional data. *Journal of the American Statistical Association*, 104(486), 718-734.

Muggeo, V. M. (2017). Interval estimation for the breakpoint in segmented regression: a smoothed score-based approach. Australian & New Zealand *Journal of Statistics*, 59(3), 311-322.

Ramsay, J. O., & Silverman, B. W. (2005). Functional data analysis. Springer series in statistics.

Schmutz, A., Jacques, J., Bouveyron, C., Cheze, L., & Martin, P. (2020). Clustering multivariate functional data in group-specific functional subspaces. *Computational Statistics*, 1-31.

Staerman, G., Mozharovskyi, P., Clémençon, S., & d'Alché-Buc, F. (2019). Functional Isolation Forest. *arXiv preprint arXiv:1904.04573.*

Modélisation hiérarchique bayésienne pour la prise en compte d'erreurs de mesure partagées dans les études de cohorte. Application en épidémiologie des rayonnements ionisants.

Sophie Ancelet 1 & Sabine Hoffmann [1,2] & Chantal Guihenneuc [3]

[1] *Institut de Radioprotection et de Sûreté Nucléaire, PSE-SANTE/SESANE/LEPID, Fontenay-Aux-Roses; email: sophie.ancelet@irsn.fr*
[2] *Institute for Medical Information Processing, Biometry and Epidemiology, University of Munich, Munich;*
[3] *BioSTM EA 7537, Faculté de Pharmacie de Paris, Université de Paris, Paris;*

Résumé.

Dans les études de cohorte, on s'intéresse généralement à l'association entre le temps jusqu'au décès par une certaine pathologie et l'exposition cumulée à un (ou plusieurs) agent(s) pathogène(s). Dans ce contexte, l'historique d'exposition des individus est généralement estimé rétrospectivement ou mesuré de manière prospective en utilisant différentes stratégies selon la période d'exposition. Cela peut créer des combinaisons d'erreurs de mesure assez complexes, notamment caractérisées par une hétérogénéité dans le temps du type et de la magnitude des erreurs. Par ailleurs, si une erreur est commise sur l'estimation d'un niveau d'exposition supposé commun à un groupe d'individus, cela peut créer des erreurs de mesure dites partagées entre individus. En outre, les conditions d'expositions et les pratiques individuelles peuvent créer des erreurs partagées intra-individus. Bien qu'il soit difficile de prendre en compte des combinaisons d'erreurs de mesure partagées et hétéroscédastiques avec des approches statistiques standard, celles-ci peuvent affecter de façon significative (i.e., biais, perte de puissance, atténuation de la courbe exposition-risque) les inférences statistiques menées dans le cadre d'études épidémiologiques. Dans ce travail, nous avons proposé deux structures hiérarchiques possibles ainsi que des algorithmes Metropolis-Within-Gibbs adaptatifs spécifiques permettant de tenir compte de l'existence d'erreurs de mesure partagées dans un modèle de survie en excès de risque instantané. Ce travail a été motivé par un cas d'étude de cohorte professionnelle en épidémiologie des rayonnements ionisants. L'objectif est d'estimer le risque de décès par cancer du poumon - corrigé des erreurs de mesure partagées sur l'exposition au radon - dans la cohorte française des mineurs d'uranium. Une nette augmentation de l'excès de risque instantané de décès par cancer du poumon a été observée par rapport à une estimation sans prise en compte d'erreurs de mesure ou avec seulement prise en compte d'erreurs de mesure non partagées. Une étude par simulations est actuellement en cours afin d'analyser l'impact d'une mauvaise spécification de modèles sur l'estimation du risque.

Mots-clés. Cancer du poumon, Epidémiologie, Erreurs de mesure d'exposition, Etude de cohorte, Modélisation hiérarchique, Statistique bayésienne

Abstract.

In many cohort studies, one is commonly interested in the association between the time until death by a certain disease and cumulative exposure to one (or several) pathogen(s). In this context, the exposure history of individuals is often estimated retrospectively or measured prospectively using different strategies according to the period of exposure. This may create rather complex patterns of exposure uncertainty, where the type and magnitude of measurement error can vary over time. Moreover, if an error is made in estimating a level of exposure assumed to be the same for a group of individuals, this may imply so-called shared measurement errors between individuals. Additionally, when cumulative exposure is considered and a method of group-level exposure estimation is used, individual's exposure conditions and practices may create intra-individual shared errors. While it is difficult to account for complex combinations of shared and heteroscedastic measurement errors in standard approaches, they pose one of the most important threats to the validity of statistical inference in epidemiological studies (i.e., biased risk estimates, loss in statistical power, attenuation of the exposure-risk curve). In this work, we thus proposed two possible hierarchical structures and adaptive Metropolis-Within-Gibbs algorithms to account for shared exposure measurement error in an excess hazard ratio (survival) model. Our motivation example comes from an occupational cohort study in radiation epidemiology. The aim is to estimate a corrected risk of death by lung cancer due to radon exposure in the French cohort of uranium miners. We observed a marked increase in the excess hazard ratio of death by lung cancer when accounting for shared measurement error in risk estimation compared to the risk estimations obtained without accounting for measurement error or only accounting for unshared errors. A simulation study is under progress to study the impact of model misspecification on risk estimate.

Keywords. Bayesian statistics, Cohort study, Epidemiology, Exposure measurement error, Hierarchical modelling, Lung cancer

1 Introduction

Dans les études de cohorte, on s'intéresse généralement à l'association entre le temps écoulé jusqu'à la mort (ou l'occurrence) d'une certaine pathologie et l'exposition cumulée à un (ou plusieurs) agent(s) pathogène(s). Dans ce contexte, l'historique d'exposition des individus est généralement estimé en utilisant différentes stratégies selon la période d'exposition. Cela peut créer des combinaisons d'erreurs de mesure assez complexes, notamment caractérisées par une hétérogénéité dans le temps du type et de la magnitude des erreurs. Pour les périodes récentes, des mesures prospectives individuelles de l'exposition sont généralement réalisées à l'aide d'appareils de mesure de plus en plus précis. Les erreurs de mesure de type classique engendrées ont des variances qui diminuent dans le temps, pour, le plus souvent, devenir négligeables. Pour les périodes plus

anciennes, les expositions sont généralement estimées rétrospectivement, mais de manière groupée: une même valeur d'exposition est attribuée à plusieurs individus partageant des caractéristiques d'exposition communes. Il s'agit d'une erreur de mesure de type Berkson. Si une erreur classique est commise sur l'estimation de ce niveau commun d'exposition, cela affectera tous les individus du groupe concerné: il s'agit d'une combinaison d'erreurs de Berkson et d'erreurs de type classique partagée inter-individus. En outre, lorsque l'exposition cumulée est considérée et qu'une méthode d'estimation groupée de l'exposition est privilégiée, les pratiques individuelles peuvent aussi créer des erreurs de mesure Berkson partagées intra-individus. Lorsqu'elles ne sont pas ou seulement mal prises en compte, les erreurs de mesure partagées peuvent affecter de façon significative (i.e., biais, perte de puissance statistique, atténuation de la courbe exposition-risque) l'inférence statistique des modèles de survie supposés dans le cadre d'études épidémiologiques (Hoffmann et al. (2018)). Malgré leurs conséquences potentiellement délétères et leur omniprésence dans les études observationnelles, les erreurs de mesure d'exposition sont rarement prises en compte. L'une des principales raisons est que les approches statistiques standard (e.g., régression-calibration, SIMEX) manquent souvent de flexibilité pour prendre en compte des combinaisons d'erreurs de mesure potentiellement partagées et hétéroscédastiques. L'objectif de ce travail est donc de promouvoir l'utilisation de l'approche hiérarchique bayésienne, connue pour sa grande flexibilité. Dans ce travail, nous proposons deux structures hiérarchiques bayésiennes possibles pour tenir compte de telles combinaisons complexes d'erreurs de mesure dans un modèle de survie. Ce travail a été motivé par un cas d'étude de cohorte professionnelle en épidémiologie des rayonnements ionisants.

2 Estimation corrigée de l'association radon/cancer du poumon chez les mineurs d'uranium français

Le radon est un gaz radioactif noble provenant de la désintégration de l'uranium 238 omniprésent dans les sols et les roches. Il a été classé cancérigène pulmonaire chez l'homme par le Centre International de Recherche sur le Cancer en 1988 et représente la 2ème cause de décès par cancer du poumon après le tabac. Afin d'estimer l'association entre une exposition chronique et à faibles doses au radon et la mortalité par cancer du poumon, le laboratoire d'épidémiologie de l'IRSN suit une cohorte de mineurs d'uranium français, qui ont été chroniquement exposés au radon dans le cadre de leur activité professionnelle. Cette cohorte inclut 5086 mineurs d'uranium. La durée moyenne du suivi est de 35 ans. A la date de point du 31 décembre 2007, 211 mineurs étaient décédés d'un cancer du poumon. Des estimations de risque antérieures ont mis en évidence une augmentation statistiquement significative du risque de mortalité par cancer du poumon associée à l'exposition cumulée au radon dans cette cohorte. Néanmoins, ces estimations ne tenaient pas compte de l'existence potentielle d'erreurs de mesure ou ne tenaient compte que de l'existence d'erreurs de mesure non-partagées (Hoffmann et al. (2017)). Aussi, l'objectif

spécifique de ce cas d'étude est d'estimer le risque de décès par cancer du poumon, corrigé des erreurs de mesure partagées sur l'exposition au radon, dans la cohorte française des mineurs d'uranium. Entre 1946 et 1956 (période 1) puis entre 1956 et 1983 (période 2), des estimations groupées de l'exposition au radon ont été réalisées par un groupe d'experts en fonction du type d'activité et de la mine associées à chaque travailleur (période 1) puis en s'appuyant sur des mesures de concentration ambiante en gaz radon réalisées en différents endroits dans les mines (période 2). Couplées à des pratiques individuelles spécifiques des mineurs d'uranium, elles ont donné lieu à une combinaison complexe d'erreurs de Berkson et de nature classique hétéroscédastiques et partagées entre certains travailleurs et sur plusieurs années de suivi d'un même travailleur. A partir de 1983 (période 3), des dosimètres ont été introduits afin de mesurer l'exposition individuelle de chaque travailleur. Les erreurs de mesure de type classique ont un impact négligeable sur l'estimation du risque d'intérêt (Hoffmann et al. (2017)): elles sont négligées dans ce travail.

3 Modèles hiérarchiques et inférence bayésienne

Afin de tenir compte de l'existence d'erreurs de mesure partagées et hétéroscédastiques dans une étude de cohorte, nous proposons 2 modèles hiérarchiques bayésiens alternatifs composés respectivement de 2 et 3 sous-modèles conditionnellement indépendants: le sous-modèle de maladie, le sous-modèle de mesure et le sous-modèle d'exposition. Les 2 modèles ont le même sous-modèle de maladie mais des sous-modèles de mesure différents.

3.1 Le sous-modèle de maladie

Soit T_i le temps écoulé jusqu'à l'évènement d'intérêt, ici: "décès par cancer du poumon" du mineur i. L'échelle de temps considérée est l'âge, impliquant une troncature à gauche à l'âge d'entrée dans l'étude. La variable aléatoire T_i est censurée à droite. Un modèle de survie en excès de risque instantané - standard en épidémiologie des rayonnements ionisants - a été considérée pour modéliser le temps de survie de chaque mineur i: $h_i(t) = h_0(t)(1 + \beta X_i^{cum}(t-5))$ avec $h_i(t)$ le risque instantané de décès par cancer du poumon du mineur i au temps t et $h_0(t)$ le risque instantané de base défini comme la fonction constante par morceaux suivante : $h_0(t) = \lambda_1 1_{t\in]0,40]} + \lambda_2 1_{t\in]40,55]} + \lambda_3 1_{t\in]55,70]} + \lambda_4 1_{t\in]70,85]}$ avec λ_j $j = \{1,2,3,4\}$ quatre paramètres inconnus, suivant les travaux de Hoffmann et al. (2017). Enfin, $X_i^{cum}(t-5)$ désigne l'exposition cumulée au radon du mineur i à l'âge t lagguée de 5 ans afin de permettre un temps de latence entre l'exposition et l'occurrence d'un décès par cancer du poumon qui puisse être associée à cette exposition.

3.2 Les sous-modèles de mesure

Le sous-modèle de mesure décrit le lien entre l'exposition "vraie" inconnue $X_{ij}(t)$ de l'individu i appartenant au groupe j au temps t et l'exposition observée $Z_{ij}(t)$ sujette

à erreur de mesure. Deux sous-modèles de mesure, de forme multiplicative (jugée plus réaliste en épidémiologie professionnelle et environnementale et garantissant la positivité des expositions), ont été considérés pour les périodes 1 (1946-1955) et 2 (1956-1983):

$$\mathcal{M}_1 : \begin{cases} X_{ij}^1(t) = Z_{ij}^1(t).U_i^1 & \text{(période 1)} \\ X_{ij}^2(t) = Z_{ij}^2(t).U_i^2 & \text{(période 2)} \end{cases}$$

$$\mathcal{M}_2 : \begin{cases} Z_j^1 = \xi_j.U_j^1 & \text{(période 1)} \\ X_{ij}^1(t) = \xi_j.T_{ij}(t).U_i^1 & \text{(période 1)} \\ X_{ij}^2(t) = X_i^2(t) = Z_{ij}^2(t).U_i^2 & \text{(période 2)} \end{cases}$$

$T_{ij}(t)$ désigne le nombre de mois travaillés par le mineur i du groupe j à l'année t. Tous les termes d'erreurs de mesure U_i^1, U_i^2 et U_j^1 sont supposés suivre une loi log-normale (conformément à une grande partie de la littérature sur l'incertitude d'exposition au radon) dont la moyenne et la variance à échelle log ont la forme $-\frac{\sigma^2}{2}$ et σ^2 respectivement (mais avec des valeurs supposées distinctes selon le terme d'erreur et la période calendaire pour la période 2). Cette paramétrisation implique que $E(U_i^1) = E(U_i^2) = E(U_j^1) = 1$ et donc que les termes d'erreurs de mesure n'introduisent pas de biais systématiques. Le sous-modèle de mesure $\mathcal{M}_1$ suppose uniquement l'existence d'erreurs de mesure Berkson hétéroscédastiques et partagées à la fois entre les individus d'un même groupe j et sur plusieurs années d'exposition d'un individu pour les périodes 1 et 2. Ainsi, les termes d'erreurs U_i^1 et U_i^2 sont indépendants de j et t. Le sous-modèle de mesure $\mathcal{M}_2$ modélise plus finement l'erreur de mesure sur la période 1 en tenant compte de l'existence additionnelle d'une erreur de mesure classique partagée sur l'estimation moyenne Z_j^1 estimée et attribuée à tous les mineurs d'un même site minier j pour les dix années d'exposition (1946-1955). Pour la période post-1983, l'erreur de mesure classique est négligée.

3.3 Le modèle d'exposition

$\mathcal{M}_2$ nécessite la spécification supplémentaire d'un sous-modèle d'exposition décrivant l'incertitude sur l'espérance ξ_j de l'exposition "réelle" au radon dans le site minier j sur la période 1946-1955. Ceci est indispensable à la modélisation d'une erreur classique. Conformément à de nombreuses études résidentielles et professionnelles qui ont observé une distribution log-normale pour l'exposition au radon, ξ_j suit ici une loi log-normale.

3.4 Lois *a priori* et inférence bayésienne

Une densité *a priori* normale centrée avec une large variance a été considérée pour le paramètre de risque β. Des données externes sur la mortalité par cancer du poumon chez les hommes en France entre 1968 et 2005 ont été utilisées pour spécifier des lois *a priori* Gamma informatives sur les paramètres λ_j ($j = \{1, 2, 3, 4\}$) définissant $h_0(t)$. Les paramètres des composantes d'erreurs de mesure lognormales U_i^1, U_i^2 et U_j^1 ont été fixés selon les travaux de Allodji et al. (2012). Néanmoins, des analyses de sensibilité ont

été réalisées. Les deux modèles hiérarchiques bayésiens proposés ont été inférés à l'aide d'algorithmes de type Metropolis-Within-Gibbs adaptatifs, codés en langage Python.

4 Résultats

Par rapport à une approche sans prise en compte des erreurs de mesure, l'ajustement du modèle hiérarchique bayésien avec sous-modèle de mesure $\mathcal{M}_1$ a mis en évidence une augmentation de 14% de l'excès de risque instantané (EHR) estimé de décès par cancer du poumon associée à l'exposition au radon ainsi qu'un élargissement de l'intervalle de crédibilité à 95%. Celui du modèle incluant $\mathcal{M}_2$ a mis en évidence une augmentation très marquée de 65% de l'EHR estimé associée à une augmentation encore plus importante de l'incertitude d'estimation. Une comparaison avec les estimations de risque obtenues dans la sous-cohorte des mineurs d'uranium embauchés à partir de 1956 montre que $\mathcal{M}_2$ permet la plus forte correction de l'atténuation de la relation exposition-risque observée lors de l'ajustement d'un modèle de maladie sans prise en compte d'erreurs de mesure.

5 Conclusion et perspectives

Ces résultats soulignent l'importance d'une caractérisation minutieuse de toutes les composantes de l'erreur de mesure de l'exposition dans une étude de cohorte. Une étude par simulations est en cours afin d'analyser l'impact d'une mauvaise spécification des sous-modèles proposés sur le risque estimé. Le calcul puis la comparaison du critère d'information de Watabane-Akaike (WAIC) pour les modèles $\mathcal{M}_1$, $\mathcal{M}_2$ et sans prise en compte des erreurs de mesure ainsi qu'une comparaison plus fine des performances prédictives *a posteriori* relatives à ces trois modèles sont prévus.

Bibliographie

Hoffmann, S. et Laurier, D. et Rage, E. et Guihenneuc, C. et Ancelet, S. (2018). Shared and unshared measurement error in occupational cohort studies and their effects on statistical inference in proportional hazards models, *Plos One*, 13(2):e0190792.

Hoffmann, S. et Rage, E. et Laurier, D. et Laroche, P. et Guihenneuc, C. et Ancelet, S. (2017) Accounting for Berkson and classical measurement error in radon exposure using a Bayesian structural approach in the analysis of lung cancer mortality in the French cohort of uranium miners. *Radiation Research*, 187(2):196–209

Allodji, R.S. et Leuraud, K. et Bernhard, S. et Henry, S. et Bénichou, J. et Laurier, D. (2012) Assessment of uncertainty associated with measuring exposure to radon and decay products in the French uranium miners cohort. Journal of Radiological Protection, 32(1):85–100.

Discrimination entre et dans les classes (semi)continues des modèles Tweedie et géométriques Tweedie

Rahma Abid [1] & Célestin C. Kokonendji [2]

[1] *Université de Sfax, Laboratoire de Probabilités et Statistique de Sfax et Université Paris-Dauphine Tunis. rahma.abid@dauphine.tn*
[2] *Université Bourgogne Franche-Comté, Laboratoire de Mathématiques de Besançon. celestin.kokonendji@univ-fcomte.fr*

Résumé. Dans les modèles Tweedie et géométriques Tweedie, le paramètre de puissance commun $p \notin (0, 1)$ est un indicateur de sélection automatique de distribution. Il sépare principalement deux sous-classes de distributions semi-continues ($1 < p < 2$) et positives continues ($p \geq 2$). Nous explorons des outils de diagnostiques basés sur le test du rapport de vraisemblance et le test de Kolmogorov-Smirnov afin de discriminer des distributions très proches dans chaque sous-classe de ces deux modèles selon des valeurs de p. Basés sur l'unique égalité des indices de variation, nous discriminons également les distributions gamma et géométrique gamma avec $p = 2$ des familles Tweedie et géométriques Tweedie, respectivement. Nous effectuons une étude de simulation pour évaluer les procédures de discrimination dans ces sous-classes de deux familles. En se basant sur les probabilités de faire une sélection correcte, les distributions semi-continues ($1 < p \leq 2$) au sens large se distinguent nettement plus que les distributions continues sur-variées ($p > 2$). Pour terminer, deux jeux de données à des fins d'illustration sont étudiés.

Mots-clés. Distance Kolmogorov-Smirnov, Test du rapport de vraisemblance, Probabilité de faire une sélection correcte, Indice de variation, Indice de zéro masse.

Abstract. In both Tweedie and geometric Tweedie models, the common power parameter $p \notin (0, 1)$ works as an automatic distribution selection. It mainly separates two subclasses of semicontinuous ($1 < p < 2$) and positive continuous ($p \geq 2$) distributions. We explore diagnostic tools based on the maximum likelihood ratio test and minimum Kolmogorov-Smirnov distance methods in order to discriminate very close distributions within each subclass of these two models according to values of p. Grounded on the unique equality of variation indices, we also discriminate the gamma and geometric gamma distributions with $p = 2$ in Tweedie and geometric Tweedie families, respectively. We perform a numerical comparison study to assess the discrimination procedures in these subclasses of two families. Based on probabilities of correct selection, semicontinuous ($1 < p \leq 2$) distributions in the broad sense are significantly more distinguishable than the over-varied continuous ($p > 2$) ones. Finally, two datasets for illustration purposes are investigated.

Keywords. Kolmogorov-Smirnov distance, Likelihood ratio test, Probability of correct selection, Variation index, Zero-mass index.

1 Introduction

Les modèles Tweedie et géométriques Tweedie fournissent des familles paramétriques flexibles de distributions pour traiter principalement des données asymétriques à droite et peuvent traiter des données continues avec une masse en zéro (Tweedie, 1984; Jørgensen et Kokonendji, 2011). Le paramètre de puissance commun $p \notin]0,1[$, appelé le paramètre de Tweedie est connecté à l'indice de stabilité (géométrique) commun $\alpha = (2-p)/(1-p)$, joue un rôle intrinsèque dans les deux modèles. En effet, p est un indicateur qui distingue chaque distribution dans chaque famille. La famille Tweedie comprend de nombreuses distributions spéciales, notamment gaussienne, Poisson, gamma décentrée, gamma et inverse gaussienne. La famille géométrique Tweedie, à son tour, provient de sommes géométriques de variables aléatoires Tweedie et peut être considérée comme un mélange exponentiel de la famille Tweedie (Abid et al., 2020). Des distributions particulières représentent la version géométrique de celles de Tweedie.

Comme préliminaires à une procédure de discrimination entre deux distributions, il est nécessaire que les deux distributions aient des caractéristiques communes telles que les supports et les allures des densités. Plus précisément, pour les familles de distributions de Tweedie et de géométriques Tweedie, nous considérerons les indices de zéro masse et de variation récemment introduits par Abid et al. (2020) pour une variable aléatoire non négative Y. Rappelons que l'indice de zéro-masse est défini par $\mathrm{ZM}(Y) := \mathbb{P}(Y \leq y) \in [0,1]$ pour $y \to 0$. Ainsi, $\mathrm{ZM} \to \varrho$ lorsque $y \to 0$ désigne une distribution semi-continue si $\varrho > 0$ et une distribution absolument continue si $\varrho = 0$. Quant à l'indice de variation exprimé par $\mathrm{VI(Y)} = \mathrm{VarY}/(\mathbb{E}\mathrm{Y})^2 \in]0,+\infty[$, il est défini par rapport à la distribution exponentielle standard. En fait, Y est dite sur-, équi- et sous-variée par rapport à la loi exponentielle avec une moyenne $\mathbb{E}Y$ si $\mathrm{VI} > 1$, $\mathrm{VI} = 1$ et $\mathrm{VI} < 1$, respectivement.

L'idée de discriminer deux distributions a été initialement proposée dans le travail pionnier de Cox (1961). Et depuis, plusieurs auteurs ont abordé la discrimination entre deux distributions bien proches. La plupart de ces disriminations sont basés sur le test du rapport de vraisemblance maximale (LRT) et la distance minimale de Kolmogorov-Smirnov (KSD). L'objectif de cet article est de discriminer entre et dans les sous-classes des modèles Tweedie et géométriques Tweedie en utilisant les méthodes maximum LRT et minimum KSD. Ces modèles ont déjà été comparés dans le cadre des modèles linéaires généralisés (Kokonendji et al., 2020). Les sections 2 et 3 présentent certaines caractéristiques des deux modèles avec le cas commun de $p = 2$. La section 4 décrit les procédures proposées de discrimination et la probabilité estimée de faire une sélection correcte (PCS). La section 5 résume les résultats numériques et les axes d'application. Une conclusion est donnée à la section 6.

2 Propriétés de la famille Tweedie

Dans cette section, nous présentons certaines caractéristiques des modèles Tweedie continus et semi-continus. Soit X une variable aléatoire d'une distribution Tweedie, notée $Tw_p(m, \phi)$. Sa fonction de densité est donnée par

$$f_{Tw_p}(x; m, \phi) = a_p(x; \phi) \exp[\{x\psi_p(m) - K_p(\psi_p(m))\}/\phi] \mathbb{1}_{\mathbb{S}_p}(x), \tag{1}$$

où $\phi > 0$ est le paramètre de dispersion, $p \in]-\infty, 0] \cup [1, +\infty[$ est l'indice Tweedie déterminant la distribution, $\mathbb{S}_p$ est le support de la distribution, $a_p(x; \phi)$ est la fonction de normalisation, K_p est la fonction cumulative, ψ_p est la fonction inverse de K_p' et $m = K_p'(\theta)$ est la moyenne de X. Notons que $K_p'(\cdot)$ définit un difféomorphisme entre son domaine canonique Θ_p et son image $M_p := K_p'(\Theta_p)$ qui est son domaine des moyennes. Bien que les densités de Tweedie ne sont généralement pas explicites, leurs fonctions cumulantes sont simples. Les deux ensembles $\mathbb{S}_p$ et M_p dépendent de p. Pour $p = 0, p = 1, 1 < p < 2$ et $p \geq 2$, le support consiste à la droite réelle $\mathbb{R}$, des entiers naturels $\mathbb{N}$, des réelles positives ou nulles $[0, +\infty[$ et strictement positives $]0, +\infty[$, respectivement. Les domaines des moyennes dans ces cas sont les supports convexes de $\mathbb{S}_p$ correspondants. Néanmoins, pour $p < 0$, on a $\mathbb{S}_p = \mathbb{R}$ et $M_p =]0, +\infty[$. Table 1 présente les sous-classes des modèles Tweedie.

Modèles (géométriques) Tweedie	$\alpha = \alpha(p)$	p	$\mathbb{S}_p$	M_p
(Géométrique) Extrême stable	$1 < \alpha < 2$	$p < 0$	$\mathbb{R}$	$]0, +\infty[$
(Laplace asymétrique/) Gaussien	$\alpha = 2$	$p = 0$	$\mathbb{R}$	$\mathbb{R}$
[N'existe pas]	$\alpha > 2$	$0 < p < 1$		
(Géométrique) Poisson	$\alpha = -\infty$	$p = 1$	$\mathbb{N}$	$]0, +\infty[$
(Géométrique) Poisson-gamma-composé	$\alpha < 0$	$1 < p < 2$	$[0, +\infty[$	$]0, +\infty[$
(*Géométrique*) *gamma décentré*	$\alpha = -1$	$p = 3/2$	$[0, +\infty[$	$]0, +\infty[$
`(Géométrique) Gamma`	$\alpha = 0$	$p = 2$	$]0, +\infty[$	$]0, +\infty[$
(Géométrique Mittag-Leffler/) Positive stable	$0 < \alpha < 1$	$p > 2$	$]0, +\infty[$	$]0, +\infty[$
(*Ressel-Kendall/*) *Inverse Gaussien*	$\alpha = 1/2$	$p = 3$	$]0, +\infty[$	$]0, +\infty[$

Table 1: Résumé des modèles Tweedie et de géométriques Tweedie, y compris leur indice de stabilité commun $\alpha = \alpha(p)$, puissance p, support des distributions $\mathbb{S}_p$ et domaine des moyennes M_p.

Étant donnée la moyenne m de $X \sim Tw_p(m, \phi)$, sa variance est ϕm^p. Ainsi,

$$\mathrm{VI}(Tw_p) = \phi m^{p-2} \quad \left(\gtreqless 1 \;\Leftrightarrow\; \phi \gtreqless m^{2-p} \right). \tag{2}$$

Les comportements de $\mathrm{VI}(Tw)$ dans (2) sont des sur-variations pour tous $p \notin]0, 1]$ et une équi-variation pour $p = 2$. Le cas spécial de $\mathrm{VI}(Y) = \phi$ dans (2) correspondant à la distribution gamma ($p = 2$) ne dépend pas de la moyenne m.

3 Propriétés de la famille géométrique Tweedie

Pour cette section, nous nous intéressons aux modèles géométriques Tweedie continus et semi-continus résultant des sommes géométriques des variables de Tweedie. Soit $Z \sim GTw_p(\widetilde{m}, \widetilde{\phi})$ la variable géométrique Tweedie de paramètre de puissance $p \notin]0,1[$, de paramètre de dispersion $\widetilde{\phi} > 0$ et de moyenne $\widetilde{m}$. On a donc la représentation suivante :

$$Z = \sum_{j=1}^{G} T_j,$$

où $T_1, T_2, \ldots$ sont des variables Tweedie indépendantes et identiquement distribuées à $Tw_p(m, \phi)$ et G est une variable aléatoire géométrique, indépendante des T_j, avec la fonction de masse de probabilité $\mathbb{P}(G = g) = q(1-q)^{g-1}$, pour $g = 1, 2, \ldots$ et $q \in]0,1[$. De plus, la famille géométrique Tweedie est intérprétée comme un mélange exponentiel (voir, par exemple, Abid et al., 2019b, Proposition 2.1) et elle est donc exprimée par la formulation hiérarchique suivante

$$X \sim \text{Exponentielle}(1) \quad \text{et} \quad Z|(X = x) \quad \sim Tw_p(x\widetilde{m}, x^{1-p}\widetilde{\phi}).$$

La fonction de densité de $Z \sim GTw_p(\widetilde{m}, \widetilde{\phi})$ est déduite de (3) par

$$f_{GTw_p}(z; \widetilde{m}, \widetilde{\phi}, p) = \int_0^\infty \exp(-x) f_{Tw_p}(z; x\widetilde{m}, x^{1-p}\widetilde{\phi}) dx. \tag{3}$$

Cette densité n'a toujours pas de forme explicite, sauf pour $p \in \{0, 1, 2, 3\}$. La méthode de Monte Carlo fournit une approximation très raisonnable $\widehat{f_{GTw_p}}$ de f_{GTw_p}, grâce à la disponibilité de la densité de Tweedie f_{Tw_p} via la fonction R `dtweedie`.

Etant donnée la moyenne $\widetilde{m}$ de $Z \sim GTw_p(\widetilde{m}, \widetilde{\phi})$, sa variance est $\widetilde{m}^2 + \widetilde{\phi}\widetilde{m}^p$. D'où,

$$\text{VI}(GTw_p) = 1 + \widetilde{\phi}\widetilde{m}^{p-2} \quad \left(\gtreqless 1 \;\Leftrightarrow\; \widetilde{\phi} \gtreqless 0 \right). \tag{4}$$

En considérant la possibilité d'obtenir $\widetilde{\phi}$, les comportements de VI(GTw) dans (4) des modèles géométriques Tweedie étendus sont clairement sur-, équi- et sous-variations pour $p \notin]0,1]$, $p = 2$ et $p \in]-\infty, 0] \cup]1, 2]$, respectivement. Toutefois, la fonction densité associée f_{GTw_p} n'existe pas pour $\widetilde{\phi} < 0$. Notons que, tout comme les modèles Tweedie avec $p = 2$ dans (2), l'indice de variation VI(GTw) dans (4) pour le cas particulier $p = 2$ correspondant à la distribution géométrique gamma est égal à $1 + \widetilde{\phi}$ et ne dépend pas de la moyenne $\widetilde{m}$. Pour $p = 2$ et étant donnés les modèles $\widetilde{m} = m > 0$, les deux indices de variation pour Tweedie (2) et géométrique Tweedie (4) coïncident lorsque leurs paramètres de dispersion diffèrent de +1 au sens de géométrique Tweedie. Plus conventionnellement, on peut écrire $Tw_2(m, \phi) \approx GTw_2(m, 1 + \phi)$ pour $\phi \geq 1$ et tout $m > 0$.

4 Procédures de discrimination

Pour diagnostiquer le modèle d'ajustement approprié parmi deux distributions données pour un jeu de données, deux techniques sont envisagées impliquant le maximum LRT et le minimum KSD comme critères d'optimalité. Considérons un échantillon aléatoire $Y_1, Y_2, \ldots, Y_n$ qui appartient à l'une des distributions parentes $f_p(y; m, \phi)$. Pour $p > 1$ fixé, les estimateurs du maximum de vraisemblance de la moyenne et du paramètre de dispersion sont donnés par

$$\widehat{m} = \frac{1}{n}\sum_{i=1}^{n} Y_i \quad \text{et} \quad \widehat{\phi} = \arg\max_{\phi>0} L_p(\widehat{m}, \phi),$$

où $L_p(\widehat{m}, \phi)$ est la fonction de vraisemblance profilée calculée en $\widehat{m}$. La statistique du rapport de vraisemblance, également appelée la statistique de Cox, est définie par

$$LT_{p_j,p_{j'}} = \log\left(\frac{L_{p_j}(\widehat{m}_j, \widehat{\phi}_j)}{L_{p_{j'}}(\widehat{m}_{j'}, \widehat{\phi}_{j'})}\right). \tag{5}$$

La règle de décision pour discriminer entre deux distributions ayant des densités f_{p_j} et $f_{p_{j'}}$ est de choisir f_{p_j} si $LT_{p_j,p_{j'}} > 0$, et de rejeter f_{p_j} en faveur de $f_{p_{j'}}$ sinon. Notons que, contrairement au LRT, le test KSD peut considérer plus de deux distributions compétitives pour décrire les données. Le KSD est, quant à lui, défini par

$$KS_{p_j} = \sup_{-\infty<y<\infty} |\widehat{F}_{p_j}(y; \widehat{m}_j, \widehat{\phi}_j) - \widetilde{F}(y)|, \quad j \in \{1, \ldots, \ell\}, \tag{6}$$

avec $\ell \geq 2$, $\widehat{F}_{p_j}(\cdot; \widehat{m}_j, \widehat{\phi}_j)$ la fonction de distribution de $f_{p_j}(\cdot; \widehat{m}_j, \widehat{\phi}_j)$ et $\widetilde{F}(\cdot)$ la fonction de distribution empirique calculée directement à partir des données. L'indice du modèle j_0 ayant la distance minimale est donc sélectionné comme modèle gagnant :

$$j_0 = \arg\min_{j\in\{1,\ldots,\ell\}} KS_{p_j}.$$

Les performances des méthodes maximum LRT et minimum KSD sont étudiées par les PCS à partir de simulations. En pratique, nous générons $(Y_n^{(1)}, \ldots, Y_n^{(N)})$, où $Y_n^{(k)}$ sont k-échantillons aléatoires de taille n qui appartiennent à f_p. Nous répétons les deux procédures, LRT et KSD, pour chaque $Y_n^{(k)}$, $k = 1, \ldots, N$. Le PCS qui correspond à la proportion de fois f_p est choisi comme modèle gagnant et peut être évalué par :

$$\widehat{PCS}_p = \frac{1}{N}\sum_{k=1}^{N} \mathbb{1}\{Y_n^{(k)} \text{ est correctement classifie}\}. \tag{7}$$

5 Simulations et applications

Nous appliquerons les méthodes LRT et KSD pour discriminer entre les modèles communs de Tweedie et géométriques Tweedie d'une part et dans chaque sous-classes des modèles Tweedie et géométriques Tweedie d'autre part. D'abord, nous considérons la discrimination entre gamma $Tw_2(m, \phi)$ et géométrique gamma $GTw_2(\widetilde{m}, \widetilde{\phi})$ vérifiant $\widetilde{\phi} = 1 + \phi$. Puis, nous supposons que la distribution parente est Tw_p et les distributions alternatives sont $Tw_{p+\varepsilon}$, avec $\varepsilon > 0$ tel que Tw_p et $Tw_{p+\varepsilon}$ sont de même type (voir Table 1). Ce qui vise à détecter l'évolution de la discrimination entre les distributions pour chaque type: $1 < p < 2$ and $p > 2$. Finalement, nous supposons que la distribution parente est GTw_p et les distributions alternatives sont $GTw_{p+\varepsilon}$, avec $\varepsilon > 0$ tel que GTw_p et $GTw_{p+\varepsilon}$ sont de même type.

Nous comparons dans chaque configuration l'évolution du PCS pour différentes combinaisons de paramètres et de tailles d'échantillons. La méthode LRT s'est avérée plus performante que KSD. De plus, les distributions semi-continues ($1 < p \leq 2$) au sens large sont nettement plus distinguables que celles continues sur-variées ($p > 2$) des deux familles respectives. Nous analysons deux jeux de données. Concernant le premier, les distributions gamma Tw_2 et géométrique gamma GTw_2 sont comparées. Pour le second, les deux sous-classes semi-continues ($1 < p < 2$) de Tweedie et géométrique Tweedie sont considérées en suggérant différentes valeurs de p.

5.1 Temps de défaillance du système de climatisation

Les données sont constituées des temps de défaillance du système de climatisation d'un avion (Linhart et Zucchini, 1986). Le maximum de vraisemblance des paramètres de $Tw_2(m, \phi)$ et $GTw_2(\widetilde{m}, \widetilde{\phi})$ sont $\widehat{m} = 59.60$, $\widehat{\phi} = 1.2317$, $\widehat{\widetilde{m}} = 59.60$ et $\widehat{\widetilde{\phi}} = 0.2380$. Il est à noter que, $\widehat{\widetilde{\phi}} \simeq 1 - \widehat{\phi}$. Les fonctions de log-vraisemblance calculées aux maximum de vraisemblance des paramètres correspondant à Tw_2 et GTw_2 sont -152.1673 et -154.1369, respectivement. Puisque la valeur des fonctions log-vraisemblance correspondant à Tw_2 est légèrement supérieur à celui de GTw_2, le modèle Tw_2 est sélectionné par la méthode LRT pour décrire ces données. Le KSD entre les données et la fonction de distribution Tw_2 estimée est 0.05213, alors que le KSD entre les données et la fonction de distribution GTw_2 estimée est 0.09263. Dans cette perspective, Tw_2 est à nouveau sélectionné avec ce critère.

5.2 Temps de défaillance des pompes

Le deuxième jeu de données concerne le temps de défaillance de 61 pompes sous-marines au cours de la période d'observation entre mai 1987 et décembre 1993 (Dudenhoeffer et al., 1998). Le temps de défaillance montre une présence significative de zéros

($\widehat{\text{ZM}} = 0.1148$), ce qui nous guide pour discriminer dans les sous-classes des familles Tweedie et géométrique Tweedie semi-continus ($1 < p < 2$). Par exemple, neuf valeurs de p entre 1.1 et 1.9 sont considérées. Pour ces données, la moyenne est $23,0327$. Sous l'hypothèse que les données proviennent de Tw_p, les paramètres de dispersion sont estimés pour chaque p donné.

La Table 2 révèle les paramètres de dispersion estimés ainsi que les valeurs de log-vraisemblance et les KSD. Basés sur les valeurs de log-vraisemblance, $Tw_{1,4}$ s'avère être la valeur préférée. De plus, les KSD suggèrent choisir $Tw_{1,6}$. Cependant, il est difficile de décider quel modèle entre $Tw_{1,3}$, $Tw_{1,4}$, $Tw_{1,5}$, $Tw_{1,6}$ ou $Tw_{1,7}$ correspond le mieux aux données correspondantes car la différence dans le sens KSD est assez petite. De même, en supposant que les données proviennent de GTw_p, les valeurs de log-vraisemblance indiquent que $GTw_{1.1}$ est le modèle d'ajustement préféré. Néanmoins, par rapport à la valeur de log-vraisemblance de $GTw_{1.2}$, les deux distributions correspondent bien à ces données. En ce qui concerne les valeurs KSD, $GTw_{1.2}$ et $GTw_{1.3}$ sont le meilleur choix pour les données.

Modèles	$\left(\widetilde{\phi}\right)\widehat{\phi}$	Log-lik	KSD
$(G)Tw_{1.1}$	(1.8000) 5.7238	(−244.7528) −266.0258	(0.0497) 0.1609
$(G)Tw_{1.2}$	(1.5300) 6.1105	(−245.7920) −250.6113	(0.0449) 0.1079
$(G)Tw_{1.3}$	(1.2200) 5.4724	(−250.2459) −246.3995	(0.0449) 0.0791
$(G)Tw_{1.4}$	(2.1000) 4.6439	(−250.5964) −245.9001	(0.0806) 0.0605
$(G)Tw_{1.5}$	(1.2800) 3.8792	(−251.0668) −247.2682	(0.0742) 0.0469
$(G)Tw_{1.6}$	(0.9300) 3.2701	(−252.1867) −250.1159	(0.0672) 0.0379
$(G)Tw_{1.7}$	(0.8600) 2.8560	(−253.9838) −254.8261	(0.0964) 0.0493
$(G)Tw_{1.8}$	(0.6300) 2.7051	(−256.0918) −262.9210	(0.0820) 0.0946
$(G)Tw_{1.9}$	(0.5600) 3.1977	(−260.2073) −280.2501	(0.1076) 0.2092

Table 2: Parmamètres de dispersion estimés, les valeurs du log-vraisemblances (Log-lik) et KSDs pour les alternatives Tw_p and GTw_p avec $1 < p < 2$. Les nombres en parenthèses représentent les résultats à partir des modèles GTw_p.

6 Conclusion

En conclusion, nous avons adopté les méthodes maximum LRT et minimum KSD pour discriminer d'abord les distributions gamma et géométrique gamma ($p = 2$) et suite dans les distributions semi-continues ($1 < p < 2$) et continues ($p > 2$) des familles Tweedie et géométriques Tweedie. Les deux seules distributions discrètes de Tweedie et de géométriques Tweedie sont Poisson et géométrique Poisson avec $p = 1$. Elles sont

incluses dans deux classes de comptage comparables qui sont les modèles Poisson-Tweedie et ses sommes géométriques et méritent d'être discriminées.

Bibliographie

Abid, R., Kokonendji, C. C. and Masmoudi, A. (2020). Geometric-Tweedie regression models for continuous and semicontinuous data with variation phenomenon, *AStA Advances in Statistical Analysis*, 104, 33-58.

Cox, D. R. (1961). Tests of separate families of hypotheses, Proceedings of the Fourth Berkeley Symposium in Mathematical Statistics and Probability, Berkeley, University of California Press, pp. 105-123.

Dudenhoeffer, D. D., Gaver, D. P. and Jacobs, P. A. (1998). Failure, repair and replacement analyses of a navy subsystem: Case study of a pump. *Applied Stochastic Models Data Analysis*, 13, 369-376.

Jørgensen, B. and Kokonendji, C. C. (2011). Dispersion models for geometric sums. *Brazilian Journal of Probability and Statistics*, 25, 263-293.

Kokonendji, C. C., Bonat, W. H. and Abid, R. (2020). Tweedie regression models and its geometric sums for (semi-)continuous data. *WIREs Computational Statistics*, 12, in press (DOI : 10.1002/WICS.1496).

Linhart, H. and Zucchini, W. (1986). *Model Selection*. New York, Wiley.

Tweedie, M. C. K. (1984). An index which distinguishes between some important exponential families. In *Statistics: Applications and New Directions. Proceedings of the Indian Statistical Institute Golden Jubilee International Conference* (J. K. Ghosh and J. Roy, eds.), pp. 579-604, Indian Statistical Institute, Calcutta.

Une approche par noyaux multiples pour l'apprentissage non-supervisé de représentation de données fonctionnelles dans des espaces de Sobolev

Julien Ah-Pine[1,2] & Anne-Françoise Yao[2]

[1] *Université de Lyon, Lyon 2 et ERIC EA3083, 5 Avenue Pierre Mendèes France, 69500 Bron, France; julien.ah-pine@univ-lyon2.fr*
[2] *Université Clermont Auvergne, LMBP UMR6620, 3 place Vasarely, 63170 Aubière, France; anne.yao@uca.fr*

Résumé. Nous appliquons de façon complémentaire l'ACP à noyaux et le k-means à noyaux multiples aux données fonctionnelles. Nous définissons ainsi une approche pour l'apprentissage non-supervisé de ce type de données. Nous supposons que les fonctions appartiennent à un espace de Sobolev et exploitons les fonctions dérivées dans l'analyse selon une approche multi-vue. Notre méthode met en avant l'utilisation de fonctions noyaux permettant de représenter les données fonctionnelles dans des RKHS. Par ailleurs, elle utilise le k-means à noyaux multiples afin de déterminer une combinaison linéaire des fonctions noyaux qui optimise la variance inter-groupe. L'ACP à noyaux permet de réduire en amont les RKHS aux axes principaux et/ou de visualiser, en aval, les résultats du k-means à noyaux multiples.

Mots-clés. Analyse de données fonctionnelles, Fonctions dérivées, k-means à noyaux multiples, ACP à noyaux.

Abstract. We apply kernel PCA and multiple kernel k-means to functional data in a complementary way. We introduce a framework for the unsupervised learning of that kind of data. We assume that the functions belong to a Sobolev space and we emphasize the derivative functions in the analysis following a multi-view approach. Our framework makes it possible to use kernel functions in order to implicitly represent the functions and their derivatives in RKHS. Moreover, it uses a multiple kernel k-means so as to determine a linear combination of the kernel functions that aims at maximizing the variance between the clusters. The kernel PCA can be used to reduce each kernel matrix before clustering and/or to visualize the results of the multiple kernel k-means.

Keywords. Functional data analysis, Derivative functions, multiple kernel k-means, kernel PCA.

1 Contexte et description de l'approche proposée

Les technologies modernes nous permettent d'enregistrer de façon volumineuse des mesures de phénomènes évoluant dans le temps et l'espace. Ces phénomènes peuvent être très

divers allant du domaine de l'environnement au niveau planétaire comme pour le réchauffement climatique, jusqu'aux activités quotidiennnes de chaque individu comme la mesure de la fréquence cardiaque tout au long d'une journée.

Du point de vue formel, ces ensembles discrets de mesures proviennent de fonctions continues que l'on observe en des points dans le temps et/ou l'espace. Ce type de données dépassent le cadre classique multivarié puisque ce dernier ne tient pas compte explicitement de cette dépendance temporelle et/ou spatiale. Or pour l'analyse des phénomènes sous-jacents, il est important d'intégrer ces dépendances comme composante essentielle de la nature même des données que l'on étudie. L'analyse de données fonctionnelles est la branche des statistiques qui s'intéresse à cette problématique [8, 2].

Dans cette communication, les objets à l'étude forment un échantillon de n fonctions $\{x_i\}_{i=1,\dots,n}$. Nous supposons que ce sont des éléments de l'espace de Sobolev $\mathbb{W}^{2,q}$ consistant en des fonctions de $\mathbb{L}^2$ dont les dérivées $\{D^j x_i\}_{i=1,\dots,n;j=1,\dots,q}$ sont également des fonctions de $\mathbb{L}^2$. Nous nous intéressons plus particulièrement au problème de l'apprentissage non-supervisé où il s'agit de déterminer les principales régularités au sein de l'échantillon. Deux applications distinctes mais complémentaires dans ce cadre sont: la réduction de dimension et la classification automatique.

L'approche classique pour la réduction de dimension est l'analyse en composantes principales (ACP) fonctionnelle. Celle-ci repose sur l'analyse spectrale de l'opérateur covariance. En particulier, la fonction covariance est un noyau de Mercer et on peut donc la décomposer dans une base orthonormée de fonctions propres. L'ACP fonctionnelle projette alors les $\{x_i\}_i$ dans le sous-espace engendré par les fonctions propres associées aux valeurs propres les plus grandes afin de conserver au maximum la variance.

Nous proposons d'appliquer une démarche duale qui consiste non pas à étudier l'opérateur de covariance mais la matrice de Gram associée à l'échantillon. En particulier, nous utilisons ici des fonctions noyaux permettant de représenter implicitement les fonctions dans des espaces de Hilbert à noyau reproduisant (RKHS). Nous notons par $\mathbf{K}$ la matrice de Gram de taille $(n \times n)$ et de terme général $\mathbf{K}_{i,i'} = k(x_i, x_{i'})$ où $k : \mathbb{L}^2 \times \mathbb{L}^2 \to \mathbb{R}$ est une fonction noyau symétrique et définie positive. La motivation de cette démarche est similaire à celle au cas multivarié: les $\{x_i\}_i$ peuvent appartenir à des sous-espaces non linéaires de $\mathbb{L}^2$ et dans ce cas leur représentation dans des RKHS pourrait permettre de mieux appréhender cette non-linéarité.

Nous proposons ainsi d'utiliser l'ACP à noyaux [10] pour l'étude de données fonctionnelles. Notons que dans ce cas, il n'est pas nécessaire d'appliquer quelconque modification à la méthode définie dans le cadre multivarié dans la mesure où la structure algébrique de base qu'elle utilise est un espace de Hilbert séparable ce qui est notre cas ici également. Ceci fût déjà discuté dans [9] qui étend les SVM aux données fonctionnelles.

En revanche, nous mettons en avant la nature fonctionnelle des données et supposons en particulier que les fonctions appartiennent à $\mathbb{W}^{2,q}$. Notre hypothèse, classique en analyse de données fonctionnelles, est que les fonctions dérivées peuvent apporter une information pertinente voire cruciale pour l'analyse.

Nous représentons donc un élément x_i de notre échantillon, par la fonction elle-même mais également par ses fonctions dérivées: $(x_i, D^1x_i, \ldots, D^qx_i)$. Ces $q+1$ fonctions sont dans $\mathbb{L}^2$ mais l'utilisation de fonctions noyaux nous permettent également de les représenter dans des RKHS $\{\mathbb{H}^j\}_{j=0,\ldots,q}$ et dans ce cas, on suppose qu'il existe $q+1$ applications $\{\phi^j : \mathbb{L}^2 \to \mathbb{H}^j\}_{j=0,\ldots,q}$ nous permettant d'étudier implicitement et de façon plus large $(\phi^0(x_i), \phi^1(D^1x_i), \ldots, \phi^q(D^qx_i))$. Cette approche permet ainsi un cadre riche pour la représentation des données fonctionnelles en apprentissage automatique.

Il n'en reste pas moins la question de la métrique qui serait la plus avantageuse pour étudier les relations de proximité entre ces fonctions. Dans cette perspective nous supposons le modèle suivant: $\langle x_i, x_{i'}\rangle = \sum_{j=0}^{q} w_j \langle \phi^j(D^jx_i), \phi^j(D^jx_{i'})\rangle_{\mathbb{H}^j}$ où $w_j \geq 0, \forall j = 0, \ldots, q$, pour que la combinaison linéaire donne un noyau valide. Notons que, en appliquant l'astuce du noyau, ceci est équivalent à se donner $q+1$ fonctions noyaux $\{k^j\}_{j=1,\ldots,q}$ et dans ce cas le modèle s'écrit comme suit:

$$k(x_i, x_{i'}) = \sum_{j=0}^{q} w_j k^j(D^jx_i, D^jx_{i'})$$

Notons $\{\mathbf{K}^j\}_{j=0,\ldots,q}$ les $q+1$ matrices de Gram de taille $(n \times n)$. Chaque matrice de Gram peut-être interprétée telle une vue distincte du même élément. Notre problème consiste alors à déterminer $\mathbf{w} = (w_j)_{j=0,\ldots,q}$ tel que $\mathbf{K} = \sum_{j=0}^{q} w_j \mathbf{K}^j$ donne une matrice de Gram efficace pour l'apprentissage non-supervisé c'est à dire qui contribue à faire ressortir à la fois les proximités des éléments formant un groupe homogène et les disparités entre les éléments appartenant à des groupes distincts.

Dans ce contexte, nous soulignons l'importance de standardiser les matrices de Gram afin qu'elles soient mutuellement commensurables. Pour cela, nous divisons chaque matrice de Gram par l'écart-type des valeurs qu'elle contient. De plus, nous proposons d'utiliser une méthode basée sur les k-means à noyaux multiples afin d'estimer $\mathbf{w}$. La méthode consiste à maximiser la variance inter-groupe qui dépend ici de deux variables: $\mathbf{P}$ la partition des éléments en k groupes et $\mathbf{w}$ le vecteur des poids des matrices de Gram donnant le noyau agrégé. Dans les travaux précédents en k-means à noyaux multiples (voir par exemple [11]), plusieurs types de contraintes ont été imposées au vecteur $\mathbf{w}$ afin de borner le problème. Il est à noter que la contrainte $\sum_j w_j = 1$ donne en sortie un vecteur sparse. Dans notre cas, nous choisissons d'utiliser la contrainte $\sum_j w_j^2 = 1$ qui permet également une solution analytique mais qui aboutit à un véritable mélange des matrices de Gram.

Notre approche de k-means à noyaux multiples appliquée aux données fonctionnelles aboutit à deux types de résultats: d'une part, nous obtenons une partition de l'échantillon qui repose sur l'information provenant des fonctions et de leurs dérivées (représentées dans des RKHS ou pas); d'autre part, nous apprenons une combinaison linéaire qui combine les différentes matrices de Gram donnant ainsi un noyau k optimisé.

Enfin, nous proposons d'appliquer l'ACP à noyau sur la matrice de Gram $\mathbf{K}$ issue de la combinaison optimale afin de visualiser l'échantillon dans un espace réduit. Nous

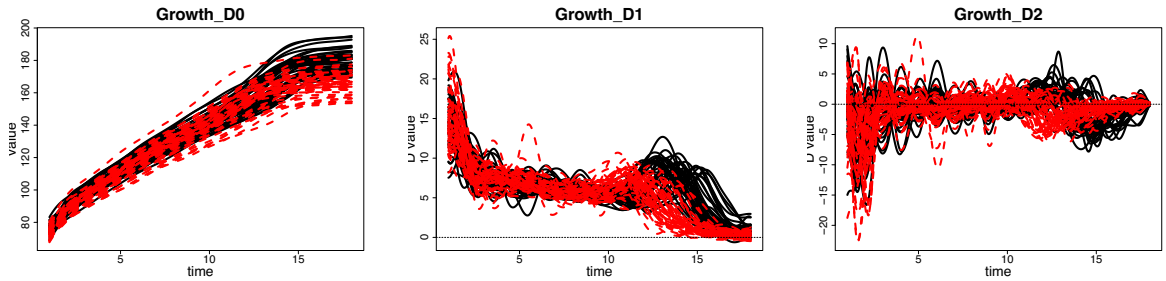


Figure 1: Courbes du jeu de données Growth. De gauche à droite: $\{x_i\}_i$, $\{Dx_i\}_i$, $\{D^2x_i\}_i$. En rouge les filles et en noir les garçons.

montrons également comment il est possible de calculer les contributions des différentes vues à la constitution des axes ce qui permet de mieux apprécier les sources principales de discrimination que notre approche a mise en avant.

2 Illustration des résultats de l'approche proposée

Nous illustrons les résultats de notre méthode à l'aide du jeu de données Growth de Berkeley qui correspond à des mesures de la taille de 93 enfants à plusieurs moments de leur première partie de vie. Nous cherchons à vérifier s'il est facilement possible de différencier de façon non-supervisée les courbes de croissance des filles de celles des garçons.

Dans la Figure 1 nous montrons les courbes des 93 sujets en distinguant en rouge les courbes des filles et en noir celles des garçons. A partir des données brutes qui sont des observations discrètes de chaque courbe, nous reconstituons la forme fonctionnelle des éléments, c'est à dire les $\{x_i\}_i$, en les représentant dans une base de fonctions B-splines. Les fonctions dérivées premières et secondes $\{Dx_i\}_i$ et $\{D^2x_i\}_i$ sont également déterminées dans cette base.

Dans la Figure 2, nous montrons les résultats de notre approche. Nous confrontons la partition obtenue par la méthode k-means à la vérité terrain par le biais de la mesure de l'information mutuelle normalisée[1] (NMI). Cinq représentations sont testées $\{x_i\}_i$, $\{Dx_i\}_i$, $\{D^2x_i\}_i$, $\{(x_i, Dx_i)\}_i$ et $\{(x_i, Dx_i, D^2x_i)\}_i$. Dans le cas des deux dernières, nous utilisons un k-means à noyaux multiples que l'on choisit de même nature:

$$k(x_i, x_i') = \sum_{j=0}^{q} w_j k'(D^j x_i, D^j x_{i'})$$

[1]Plus la mesure est proche de 1, meilleur est le résultat de la classification automatique.

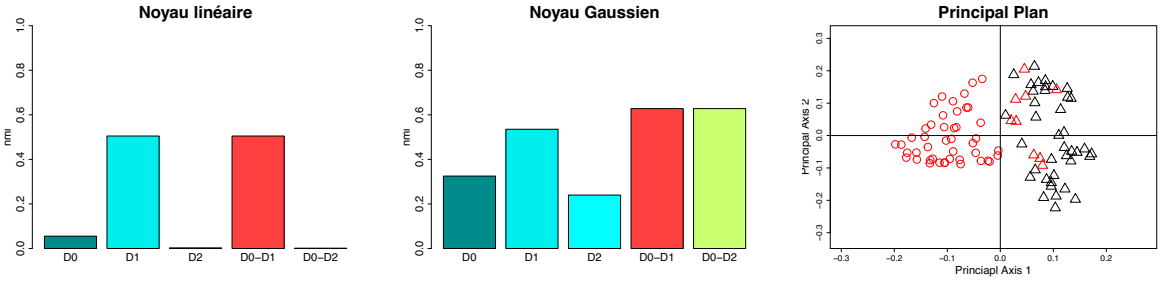


Figure 2: Résultats obtenus par notre approche. Le diagramme de gauche montre les mesures NMI pour les cinq représentations avec un noyau linéaire. Le diagramme du centre correspond aux performances pour le noyau Gaussien. Le nuage de points de droite est le résultat de l'ACP à noyau appliqué au noyau agrégé issu des noyaux Gaussiens avec la représentation $\{(x_i, Dx_i, D^2x_i)\}_i$.

avec $q = 1$ ou $q = 2$ et k' est le noyau linéaire ou (exclusif) Gaussien[2].

Le diagramme en bâtons à gauche de la Figure 2, montre les performances du k-means avec le noyau linéaire. Si nous utilisons une seule vue, clairement, c'est la dérivée première qui donne les meilleurs résultats. Combiner $\{(x_i, Dx_i)\}_i$ donne d'aussi bons résultats mais ce n'est pas le cas pour $\{(x_i, Dx_i, D^2x_i)\}_i$. En effet, malgré l'optimisation du vecteur poids, il semble que la dérivée seconde dans la représentation linéaire, apporte un bruit important qui conduit à un mauvais noyau agrégé.

Le diagramme en bâtons au centre expose les résultats obtenus avec un noyau Gaussien. Il est intéressant de noter que l'utilisation d'une fonction noyau non-linéaire permet ici d'améliorer les performances de chaque vue. Par ailleurs, nous obtenons cette fois-ci de très bons résultats pour le k-means à noyaux multiples en combinant les matrices de Gram des fonctions avec celles des dérivées premières (bâton rouge) puis celles des dérivées secondes (bâton vert).

Le nuage de points à droite de la Figure 2 correspond à la projection des éléments de l'échantillon obtenue par l'ACP à noyau. La matrice de Gram utilisée dans ce cas, est celle obtenue par le k-means à noyaux Gaussiens multiples avec la représentation $\{x_i, Dx_i, D^2x_i\}_i$. La partition à 2 classes qui est solution du k-means à noyaux multiples, peut être visualisée par les symboles cercles *versus* triangles. La vérité terrain est représentée par les couleurs. Ainsi les triangles en rouge à droite du plan sont les erreurs de notre approche non-supervisée (9 sur 93, soit un taux d'erreur de moins de 10%).

[2]Dans ce cas, nous fixons le paramètre σ^2 égale à la médiane des distances $\mathbb{L}^2$ au carré de chaque courbe à son 7ème plus proche voisin. Ce paramétrage est motivé par les résultats empiriques exposés dans [15].

Bibliographie

[1] C. Abraham, P.-A. Cornillon, E. Matzner-Løber, and N. Molinari. Unsupervised curve clustering using b-splines. *Scandinavian journal of statistics*, 30(3):581–595, 2003.

[2] F. Ferraty and P. Vieu. *Nonparametric functional data analysis: theory and practice*. Springer Science & Business Media, 2006.

[3] D. Floriello and V. Vitelli. Sparse clustering of functional data. *Journal of Multivariate Analysis*, 154:1–18, 2017.

[4] M. L. L. García, R. García-Ródenas, and A. G. Gómez. K-means algorithms for functional data. *Neurocomputing*, 151:231–245, 2015.

[5] J. Jacques and C. Preda. Functional data clustering: a survey. *Advances in Data Analysis and Classification*, 8(3):231–255, 2014.

[6] Y. Meng, J. Liang, F. Cao, and Y. He. A new distance with derivative information for functional k-means clustering algorithm. *Information Sciences*, 463:166–185, 2018.

[7] A. Muñoz and J. González. Representing functional data using support vector machines. *Pattern Recognition Letters*, 31(6):511–516, 2010.

[8] J. Ramsay and B. Silverman. *Functional Data Analysis*. Springer Science & Business Media, 2005.

[9] F. Rossi and N. Villa. Support vector machine for functional data classification. *Neurocomputing*, 69(7-9):730–742, 2006.

[10] B. Schölkopf, A. Smola, and K.-R. Müller. Kernel principal component analysis. In *International conference on artificial neural networks*, pages 583–588. Springer, 1997.

[11] G. Tzortzis and A. Likas. Kernel-based weighted multi-view clustering. In *2012 IEEE 12th international conference on data mining*, pages 675–684. IEEE, 2012.

[12] J.-L. Wang, J.-M. Chiou, and H.-G. Müller. Functional data analysis. *Annual Review of Statistics and Its Application*, 3:257–295, 2016.

[13] D. M. Witten and R. Tibshirani. A framework for feature selection in clustering. *Journal of the American Statistical Association*, 105(490):713–726, 2010.

[14] M. Yamamoto. Clustering of functional data in a low-dimensional subspace. *Advances in Data Analysis and Classification*, 6(3):219–247, 2012.

[15] L. Zelnik-Manor and P. Perona. Self-tuning spectral clustering. In *Advances in neural information processing systems*, pages 1601–1608, 2005.

Méthode de comparaison d'aires sous la courbe dans des essais cliniques avec arrêt prématuré du suivi : application aux vaccins thérapeutiques contre le VIH

Marie Alexandre[1] & Mélanie Prague[2] & Rodolphe Thiébaut[3]

Inria SISTM Team, INSERM U1219, Université de Bordeaux, ISPED - France
[1] *marie.alexandre@inria.fr,* [2] *melanie.prague@inria.fr,* [3] *rodolphe.thiebaut@inria.fr*

Résumé. Les interruptions de traitement analytiques (ATI) sont couramment utilisées pour évaluer l'efficacité de nouveaux vaccins thérapeutiques contre le VIH. Ces procédures nécessitent alors la détermination de critères de jugement synthétiques permettant de rendre compte de cette efficacité par comparaison entre différents bras de vaccination tels que l'aire sous la courbe de charge virale moyennée sur le temps de suivi (nAUC). Cependant, dû à la nécessité de remettre les patients à risque sous traitement, l'existence de données manquantes au hasard (MAR) monotone est inévitable au cours d'ATI pouvant mener à des résultats biaisés des tests de comparaison. Cette étude a pour objectif de présenter une méthode évaluant la différence de nAUC entre deux bras de vaccination à partir des dynamiques marginales de groupes estimées par des modèles à effets mixtes. L'application de cette méthode sur des simulations d'essais randomisés à deux bras de vaccination a été menée afin d'en vérifier les propriétés statistiques.
Mots-clés. VIH, efficacité vaccinale, AUC, données manquantes, modèle à effets mixtes, test statistiques

Abstract. Analytic treatment interruption (ATI) are commonly used to evaluate new HIV therapeutic vaccins efficacy. These protocols require the choice of summary endpoints, such as the area under the HIV RNA load curve normalized by follow-up time (nAUC), to assess this efficacy by comparing them between different vaccine arms. However, monotonic missing at random (MAR) data are unavoidable during ATI leading to potential biased results for comparison tests. This study aimed to present a method evaluating the difference of nAUC between two vaccine arms based on marginal dynamics of group level estimated by mixed effects models. In order to evaluate its statistical properties, the method was applied on simulated two-armed randomized vaccine trials where the difference of AUC between the two vaccine arms as well as the missingness were varied.
Keywords. HIV, vaccine efficacy, AUC, missing data, mixed effects models, statistical test

1 Introduction

Le développement de vaccins thérapeutiques est un aspect important dans la recherche de stratégies du contrôle viral du VIH à long terme. Ces derniers ont pour objectif la diminution voir l'élimination complète de l'infection virale jusqu'à présent rendu impossible par l'existence d'un réservoir viral persistant. L'absence de biomarqueurs reconnus capable de prédire le contrôle virologique en l'absence de traitement antiretroviraux (ART) fait des interruptions de traitement analytiques (ATI) l'unique moyen d'évaluer la capacité d'une nouvelle stratégie à contrôler la virémie après arrêt d'ART. Dans ce type d'étude, un choix judicieux de critère de jugement virologique est le nAUC.

1.1 Impact des données manquantes sur la statistique de test

Ces essais sont régis par des critères éthiques stricts afin de minimiser les risques encourus par les patients, tels que des critères de reprise prématurée des ART. Ces reprises précoces de traitement au cours de l'ATI, basées sur des règles définies dans le protocole, notamment pour éviter des niveaux de charges virales trop élevées pour garantir un risque minimal aux patients, sont traitées comme une sortie de l'étude et génère par conséquent des données manquantes monotones, a priori manquantes au hasard (MAR). D'un point de vue statistique, la non-prise en compte des données manquantes dans des tests classiques d'égalité de moyennes de nAUCs mène à de mauvaises propriétés statistiques telles que des erreurs de type-I élevées, des pertes de puissance ou encore un biais sur les résultats du test [1].

2 Objectifs

Nous avons pour objectif de proposer une méthodologie statistique pour tester l'efficacité de ces vaccins en comparant les dynamiques de charge virales entre les différents bras de vaccination. A ces fins, nous nous basons sur un critère de jugement facilement mesurable, précis et interprétable capable de résumer les dynamiques de charge virale, l'AUC normalisée sur la période d'interruption de traitement (nAUC). En 2014, Bell. et al [2] ont mis en évidence l'intérêt d'utiliser des méthodes basées sur le maximum de vraisemblance pour réduire le biais induit par les données manquantes de type MAR et MNAR dans le calcul de l'AUC. En se basant sur ces résultats, nous construisons un test statistique, construit sur les dynamiques marginales de nos données longitudinales estimées par un modèle à effets mixtes, permettant de conclure de l'existence d'une différence de nAUC entre nos deux groupes d'intérêt en présence de données MAR monotones.

3 Méthodes

Nous construisons une méthode permettant de tester l'efficacité d'une stratégie de vaccination thérapeutique se basant sur la comparaison de nAUC de la dynamique de charge virale durant l'ATI entre les différents bras de vaccination, tout en prenant en compte l'existence de données manquantes de type MAR monotone. Dans un premier temps, la méthode se base sur l'utilisation d'un modèle à effets mixtes pour fitter les données individuelles de charge virale. Dans un second temps, les nAUCs à l'échelle des bras de vaccination sont estimées et comparés à partir des estimations des coefficients de régression des dynamiques marginales.

3.1 Le modèle à effets mixtes

On considère un essai clinique comptabilisant N patients répartis au sein de G groupes de vaccination. On note Y_{ij,g_i} les mesures longitudinales de charge virales du patient i, appartenant au bras de vaccination g_i, au $j$ème temps de mesure de ce même groupe, où $i \in \{1, \cdots, N\}$, $j \in \{1, \cdots, m_g\}$ et $g \in \{1, \cdots, G\}$. Le modèle à effets mixtes permettant de modéliser les données Y_{ij,g_i} est construit par la somme d'un intercept γ_0, des effets fixes modélisés par G régressions de type B-splines cubiques et des effets aléatoires décrits par une fonction flexible du temps. Le modèle est donné par (1)

$$Y_{ij,g_i} = \gamma_0 + \sum_{g=1}^{G} \mathbf{1}_{[g_i=g]} \times \sum_{k=1}^{K_g} \beta_k^g \phi_k^g(t_{ij,g}) + h_i(t_{ij,g_i}) + \varepsilon_{ij} \tag{1}$$

où K_g est le nombre de fonctions de bases impliquées dans la courbe B-splines modélisant l'effet de population au sein du groupe g, les termes ϕ_k^g représentent les $k$ème fonctions de bases de la dynamique de population du groupe g où les β_k^g sont leurs coefficients de régression respectifs. Dans le cadre de notre étude, les fonctions h_i modélisant les effets aléatoires sont décrites comme la somme d'un intercept et des fonctions B-splines cubiques définies comme combinaisons linéaires des K_i bases de splines individuelles $(\Psi_k^i)_{1 \leqslant k \leqslant K_i}$ avec les coefficients de régression $(b_{ki})_{1 \leqslant k \leqslant K_i}$, $\forall i \in \{1, \cdots, N\}$, telles que $h_i(t_{ij,g_i}) = b_{0i} + \sum_{k=1}^{K_i} b_{ki} \Psi_k^i(t_{ij,g})$. Les effets aléatoires, $\boldsymbol{B} = (b_{ki}) \in \mathcal{M}_{\max(K_i)+1,N}(\mathbb{R})$, ainsi que les termes d'erreurs sont supposés indépendant et normalement distribués tels que $\boldsymbol{B} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega})$ et $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_e^2)$.

3.2 La statistique de test

On construit une statistique de test permettant de conclure sur la différence de nAUCs entre deux groupes de vaccination. Ces nAUCs sont estimées à l'échelle des bras de vaccination par méthode d'interpolation des trapèzes. L'estimation de nAUC au sein du groupe

g est donnée par

$$\widehat{n\text{AUC}}_g = \sum_{j=2}^{m_g} \frac{(t_{j,g} - t_{j-1,g})}{2} (\hat{\mu}_{j,g} + \hat{\mu}_{j-1,g}) \tag{2}$$

où $\hat{\mu}_{j,g}$ est définie comme l'estimation de la dynamique marginale de la variable Y évaluée au $j$ème temps du mesure au sein du groupe g telle que $\hat{\mu}_{j,g} = \hat{\gamma_0} + \sum_{k=1}^{K_g} \hat{\beta}_k^g \phi_k^g(t_{j,g})$. L'approximation $\widehat{n\text{AUC}}_g$ est ainsi exprimée comme combinaison linéaire des estimations des paramètres γ_0 et $(\beta_k^g)_{\substack{1 \leqslant k \leqslant K_g \\ 1 \leqslant g \leqslant G}}$ du modèle à effets mixtes (1). En se basant sur l'équation (2), on définit la différence des nAUC entre les bras de vaccination g et $\widetilde{g}$ par $\Delta\widehat{n\text{AUC}}_{g-\widetilde{g}} = \widehat{n\text{AUC}}_{\widetilde{g}} - \widehat{n\text{AUC}}_g$. Une statistique de test peut alors être dérivée donnant l'équation suivante

$$T = \frac{\Delta\widehat{n\text{AUC}}_{g-\widetilde{g}}}{\sqrt{\text{Var}\left(\Delta\widehat{n\text{AUC}}_{g-\widetilde{g}}\right)}} \sim \mathcal{N}(0,1)$$

4 Simulations et Résultats

Nous avons testé notre méthode en l'appliquant sur des données simulées à partir du modèle à effets mixtes décrit par (1), où le nombre de noeuds internes des bases de splines à l'échelle des groupes (ϕ_k^g) et individuelle (Ψ_k^i) a été fixé à 2 menant à $K_g = K_i = 5$, $\forall g \in \{1,2\}$ et $i \in \{1, \cdots, N\}$. Les positions de ces noeuds ont été fixées à $(0.25, 5.62)$ semaines pour les bases des effets fixes et à $(2.0, 4.5)$ semaines pour les effets aléatoires. Par ailleurs, nous avons défini la matrice de variance-covariance des effets aléatoires $\mathbf{\Omega}$ comme matrice diagonale telle que $\mathbf{\Omega} = \sigma_b^2 \mathbf{I}_{K_i+1}$.
Afin de vérifier le bon comportement de notre méthode vis-à-vis des conditions de simulation et de sa capacité à gérer les données manquantes, plusieurs jeux de données ont été simulés, sous différentes conditions. Pour chaque condition de simulation, la charge virale au cours de l'ATI a été mesurée à intervalle de temps constant tel que $t = (0, 1, 2 \cdots, 24)^T$ et considérant le nombre de sujets par groupe variant entre $n_g = 20, 50$ et 100. De plus, nous avons fixé la variance des termes d'erreurs telle que $\sigma_e^2 = 0.2$. Les différents jeux de données considérés ont été simulés de manière à faire varier la différence de nAUC entre les deux groupes de traitement telle que $\Delta n\text{AUC} = 0, -0.1$ et -0.25 $\log_{10}$ cp/ml. L'impact de la variabilité des données au sein de chaque groupe sur la robustesse de la méthode a également été évaluée en faisant varier la variance des nAUC. En effet, fixer σ_b^2 aux valeurs 0.01 et 0.08 nous a permis de simuler les données avec des variances de nAUC égales à 0.02 et 0.1. Par ailleurs, en fixant $\gamma_0 = -0.44$ et choisissant différentes valeurs des paramètres de populations $(\boldsymbol{\beta^1}, \boldsymbol{\beta^2})$ nous a permis de faire varier $\Delta n\text{AUC}$. Ainsi, fixer $\boldsymbol{\beta^1} = (-0.55, 4.72, 4.96, 5.18, 4.64)$ pour toutes les simulations et $\boldsymbol{\beta^2} = \boldsymbol{\beta^1}$, $(-0.54, 4.61, 4.85, 5.07, 4.54)$ et $(-0.52, 4.46, 4.69, 4.90, 4.39)$ nous a permis de cibler les différentes valeurs de $\Delta n\text{AUC}$ respectivement souhaitées.

Pour chaque combinaison de n_g, ΔnAUC et Var(nAUC) nous avons testé la méthode en considérant les données complètes, les données censurées à gauche par la présence d'une limite de détection (LOD) fixée à 50 cp/ml, ainsi que les données MAR monotones. Ces données manquantes ont été générées telle que pour tout sujet i au temps j, la variable $Y_{ij,g}$ est considérée comme manquante si $Y_{ij,g} \in \{Y_{ij,g} \mid \exists j^{'} \leqslant j, \{Y_{ij^{'},g} \geqslant \alpha\} \cap \{Y_{ij^{'}-1,g} \geqslant \alpha\}\}$, où α représente le seuil fixe de sortie d'étude. En terme plus littérale, un patient est considéré comme exclu définitivement de l'étude si son niveau de charge virale excède le seuil α au cours de deux mesures consécutives. Trois valeurs du seuil α ont été testé : $\alpha = 100.000, 50.000$ et 10.000 cp/ml (equiv. $5, 4.7$ et $4 \log_{10}$ cp/ml). La considération de ces trois valeurs de seuil a permis en particulier d'évaluer notre méthode pour des pourcentages de patients quittant l'étude allant de 5 à 100% en fonction des conditions de simulations. Contrairement aux données manquantes monotones traitées comme données non disponibles (NA) n'impactant pas littéralement l'estimation du modèle à effets mixtes, l'approximation des paramètres de ce dernier en présence de données censurées à gauche par LOD requiert l'utilisation de méthodes déjà développées incluant la probabilité de données censurées dans le calcul de la vraisemblance ([3, 4, 5, 6]). A ces fins, nous avons utilisé le package R *tlmec* [7] pour estimer nos modèles.

La robustesse de la méthode à estimer la différence d'aire sous la courbe normalisée par le temps de suivi a été évalué au moyen des erreurs de Type-I et des puissances ainsi que par l'estimation du biais de ΔnAUC et de son erreur standard. Par la suite, nous avons comparé ces grandeurs obtenus par notre méthode avec celles obtenues par des méthodes adhoc se basant sur les estimations individuelles des nAUC telles que la méthode LOCF imputant les données manquantes à la dernière valeur connues ou la méthode d'imputation à la moyenne.

La comparaison des résultats entre les différentes méthodes montre des erreurs de Type-I équivalentes et en adéquation avec les valeurs nominales attendues pour toutes les méthodes et pour tous les types de données, à l'exception des données MAR monotones générées par un seuil à 10.000 cp/ml. Alors que notre méthode présente des résultats équivalents aux méthodes adhoc en terme de puissance en l'absence de données manquantes, celle-ci semble plus robuste en présence de données manquantes pour les valeurs de seuils $\alpha \neq 10.000$ cp/ml. Ces résultats sont confirmés par des valeurs de biais et d'erreurs standard de ΔnAUC plus faible dans le cas de notre méthode. En revanche, la considération du seuil de sortie d'étude à 10.000 cp/ml, représentant près de 100% des patients impactés par des données manquantes, montre la limite de notre méthode qui induit alor une sur-estimation du biais et de l'erreur standard conduisant à une sur-estimation de l'erreur de Type-I et des puissances inférieures à celles obtenues par les méthodes basées sur les estimations individuelles.

Bibliographie

[1] John Spritzler, Victor G DeGruttola, and Lixia Pei. Two-sample tests of area-under-the-curve in the presence of missing data. *The international journal of biostatistics*, 4(1), 2008.

[2] Melanie L Bell, Madeleine T King, and Diane L Fairclough. Bias in area under the curve for longitudinal clinical trials with missing patient reported outcome data : summary measures versus summary statistics. *SAGE Open*, 4(2) :2158244014534858, 2014.

[3] Hélène Jacqmin-Gadda, Rodolphe Thiébaut, Geneviève Chêne, and Daniel Commenges. Analysis of left-censored longitudinal data with application to viral load in hiv infection. *Biostatistics*, 1(4) :355–368, 2000.

[4] Rameela Chandrasekhar, Yi Shi, Alan D Hutson, and Gregory E Wilding. Likelihood-based inferences about the mean area under a longitudinal curve in the presence of observations subject to limits of detection. *Pharmaceutical Statistics*, 14(3) :252–261, 2015.

[5] Larissa A Matos, Marcos O Prates, Ming-Hui Chen, and Victor H Lachos. Likelihood-based inference for mixed-effects models with censored response using the multivariate-t distribution. *Statistica Sinica*, pages 1323–1345, 2013.

[6] Florin Vaida and Lin Liu. Fast implementation for normal mixed effects models with censored response. *Journal of Computational and Graphical Statistics*, 18(4) :797–817, 2009.

[7] Larissa Matos, Marcos Prates, and Victor Lachos. *tlmec : Linear Student-t Mixed-Effects Models with Censored Data*, 2012. R package version 0.0-2.

Estimation of univariate Gaussian mixtures for huge raw datasets by using binned datasets

Filippo Antonazzo[1], Christophe Biernacki[2] & Christine Keribin[3]

[1] *Inria, Université de Lille, CNRS, Laboratoire de mathématiques Painlevé 59650 Villeneuve d'Ascq, France, filippo.antonazzo@inria.fr*
[2] *Inria, Université de Lille, CNRS, Laboratoire de mathématiques Painlevé 59650 Villeneuve d'Ascq, France, christophe.biernacki@inria.fr*
[3] *Université Paris-Saclay, CNRS, Inria, Laboratoire de mathématiques d'Orsay 91405 Orsay, France, christine.keribin@universite-paris-saclay.fr*

Résumé. L'intérêt de l'apprentissage non supervisé est magnifié par la croissante constante du nombre d'individus dans les échantillons. C'est en effet l'opportunité de découvrir des informations autrefois inaccessibles. Néanmoins, une importante volumétrie de données pose des difficultés relatives à des temps de calculs rapidement prohibitifs et à la grande consommation d'énergie et des ressources matérielles. L'usage de données regroupées (ou *binned data*) sur une grille adaptative pourrait répondre à ces questions ayant trait à ce qu'on qualifierait aujourd'hui de *green computing*, sans pour autant nuire à la qualité des estimations. Une 1ère approche est menée dans le cadre des mélanges gaussiens univariés, comprenant une illustration empirique et des avancées théoriques.

Mots-clés. Apprentissage non supervisé, données regroupées, big data, green computing.

Abstract. Popularity of unsupervised learning is magnified by the regular increase of sample sizes. Indeed, it provides opportunity to reveal information previously out of scope. However, the volume of data leads to some issues related to prohibitive calculation times and also to high energy consumption and the need of high computational ressources. Resorting to binned data depending on an adaptive grid is expected to give proper answer to such green computing issues while not harming the related estimation issues. A first attempt is conducted in the context of univariate Gaussian mixtures, included a numerical illustration and some theoretical advances.

Keywords. Unsupervised learning, binned data, big data, green computing.

1 Introduction

Assuming observations with values belonging to a real space $\mathcal{X}$, binned data correspond to a reduced dataset only containing the counts of observations in given regions of $\mathcal{X}$. In practice, binned data usually appear as soon as it is impossible to collect data with

infinite precision. Thus, such regions are often imposed by the data collecting process itself.

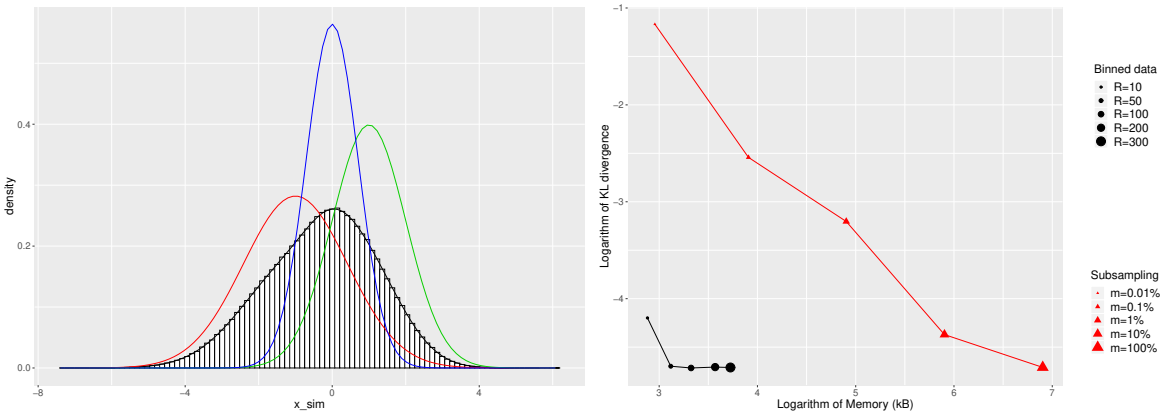
Binned data are so frequent that specific data analysis procedures are designed for them, in particular when regions are too wide to neglect uncertainty they introduce in comparison to the raw (but unavailable) dataset. For instance, in the univariate case ($\mathcal{X} = \mathbb{R}$), McLachlan & Jones (1988) introduced a binned version of the EM algorithm for estimating a univariate Gaussian mixture, whose employment was motivated by an application on red blood cells where only binned and truncated data were available. Induced by a similar problem, Cadez et al. (2002) finally extended this algorithm to the multivariate case.

In this work, we propose to use binned data with a different point of view. We suppose to have a huge amount of raw data and our challenge is to save resources (usually in terms of energy, time and computer memory) while preserving accuracy of the targeting estimation process. The key idea we defend is to group original data in order to obtain *artificially* binned ones. In this way, the size of the resulting dataset is automatically reduced, avoiding too many computing efforts. We focus our attention on the univariate Gaussian mixture estimation in this preliminary work, as a first important step to address more complex situations in the future.

Here is an early numerical example to motivate our proposed "binned" strategy. It illustrates the gain that could be expected in comparison to the classical subsampling strategy usually used for reducing the data size. In this simulation a sample of $n = 10^6$ raw data i.i.d. arises from a univariate Gaussian mixture of three components (Figure 1a), with density

$$f(x;\boldsymbol{\theta}) = 0.6\phi(x;-1,2) + 0.3\phi(x;1,1) + 0.1\phi(x;0,0.5),$$

where $\phi(\cdot;\mu,\sigma^2)$ indicates the univariate Gaussian pdf with mean μ and variance σ^2. Binned data are created through a grid for which a tuning parameter corresponds to its number of finite intervals limits, denoted here by R (more details on the grid will be given later). An EM algorithm was performed respectively with different values of R (thus different candidate binned datasets) and different values of m (thus different candidate subsample datasets).



(a) (b)

Figure 1: (a) Density simulated (black line) with the ones of the three components (red, green and blue lines); (b) Logarithm of Kullback-Leibler divergence from the true parameters for different values of R and m in function of the required computer memory (logarithmic scale).

In Figure 1b it is possible to note that the loss of information (measured by the Kullback-Leibler divergence) induced by binning is much lower than that obtained with subsampling, even negligible if we use a grid moderately dense. This is in addition accompanied by an evident gain in terms of computer memory. Such promising results could be also obtained (but not displayed here) concerning gain in terms of algorithm running time or model selection behaviour.

The outline of the paper focuses on theoretical questionings to be addressed on univariate binned data. It concerns essentially the grid properties: (1) model identifiability, (2) estimates properties and (3) grid selection. We gradually consider these questions firstly in the simplified univariate no mixture Gaussian case (Section 2) and secondly in the univariate mixture Gaussian case (Section 3).

2 Preliminary work: a single univariate Gaussian

2.1 Notations

In the general case, we denote by $\boldsymbol{x} = (x_1, \ldots, x_n)$, with $x_i \in \mathcal{X}$, a raw sample of n observations and by G a grid that divides the space $\mathcal{X}$ into R regions $\mathcal{R}_j$, $j = 1, \ldots, R$. We denote also the resulting binned data vector by $\boldsymbol{y} = (y_1, \ldots, y_R)$, where each component is defined as

$$y_j = \#\{x_i \in \mathcal{R}_j\}, \quad j = 1, \ldots, R.$$

In addition, it is assumed also that the raw sample arises from n continuous i.i.d. random variables with parametric density $f(x;\boldsymbol{\theta})$, $x \in \mathcal{X}$, indexed by a vector of parameter $\boldsymbol{\theta}$, and that $\boldsymbol{y}$ follows a multinomial distribution $M(n,\boldsymbol{p})$ with $\boldsymbol{p} = (p_1,\ldots,p_R)$, where $p_l = \int_{R_l} f(x;\boldsymbol{\theta})dx$.

In the specific case of this section, we suppose that the sample $\boldsymbol{x} \in \mathbb{R}^n$ arises from n i.i.d. univariate Gaussians $N(\mu,\sigma^2)$ with density $\phi(\cdot;\mu,\sigma^2)$. In any univariate context like this, we consider a grid G composed by R points $a_1,\ldots,a_R$ such that we obtain a vector $\boldsymbol{y}$ of $R+1$ binned data where every observation y_j is defined as $(j=0,\ldots,R)$

$$y_j = \#\{x_i : a_j \leq x_i < a_{j+1}\},$$

while setting $a_0 = -\infty$ and $a_{R+1} = \infty$.

We make also two additional hypotheses in Section 2.2.2 and 2.2.3. First, the variance σ^2 is known and equal to 1. Second, the grids considered are equispaced and symmetric around μ. With these last regularity assumptions, the grids will be simply indexed by two parameters which are the number of points R and the "starting" point a_1. Consequently, each grid will be denoted by $G(a_1,R)$.

2.2 Theoretical results

2.2.1 Identifiability

As discussed in Section 1, first of all, we are interested by a fundamental probabilistic property which is identifiability of the Gaussian distribution, related to the binned nature of available information. In that case, thanks to the monotonicity of the Gaussian cdf, it is possible to prove the following proposition, that ensures identifiability under a slight condition on R.

Proposition 2.1 *Binned univariate normal models are identifiable for $R \geq 2$.*

2.2.2 Estimates properties

The second property is statistical. We note $\hat{\mu}^b_{a_1,R}$ the binned maximum likelihood estimate (MLE) of μ obtained from the binned dataset $\boldsymbol{y}$ with an equispaced grid $G(a_1,R)$ symmetric around μ, and $\hat{\mu}^{MLE}$ the MLE of μ obtained from the raw dataset $\boldsymbol{x}$. Good statistical properties of $\hat{\mu}^b_{a_1,R}$ are assured by the following proposition:

Proposition 2.2 *$\hat{\mu}^b_{a_1,R}$ is asymptotically unbiased and* $\lim_{a_1,R\to\infty} Var(\hat{\mu}^b_{a_1,R}) = Var(\hat{\mu}^{MLE})$.

2.2.3 Grid selection

The question of grid selection is fundamental in our work since its main originality is to estimate an optimal one. In this purpose, we are first interested to access the relative

values of the two tuning parameters of the grid (a_1 and R) to obtain the optimal grid from a variance estimates point of view. The next proposition states that a_1 should decrease at least at logarithmic rate with regards to R (and vice versa). Figure 2 graphically illustrates this fact by comparing the established lower bound and the true optimal value.

Proposition 2.3 *The sequence* $a_1^{(R)} = \max_{a_1<\mu} \frac{Var(\hat{\mu}^{MLE})}{Var(\hat{\mu}^b_{R,a_1})}$ *is bounded below by the sequence* $a^{(R)} = -2\log R + \mu$.

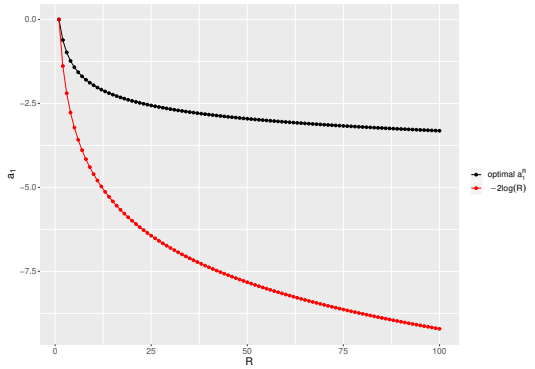


Figure 2: Lower bound for the sequence $a_1^{(R)} = \max_{a_1<0} \frac{Var(\hat{\mu}^{MLE})}{Var(\hat{\mu}^b_{R,a_1})}$ when $\mu = 0$.

The previous statement will be useful to propose sensible grid candidates but the question to select the best grid candidate is open. The following criterion, denoting by $VC^R_{a_1}$, is able to provide the best estimate $\hat{\mu}^b_{a_1,R}$ from the variance point of view, among all the equispaced grids $G(a_1, R)$ symmetric around $\frac{\min(\boldsymbol{x})+\max(\boldsymbol{x})}{2}$ (which is asymptotically equal to μ). Namely, the $VC^R_{a_1}$ criterion is defined by

$$\text{maximize}_{a_1,R} VC^R_{a_1} = \text{maximize}_{a_1,R} \sum_{i=0}^{R} \frac{(\phi(a_i, \hat{\mu}^b_{R,a_1}, 1) - \phi(a_{i-1}, \hat{\mu}^b_{R,a_1}, 1))^2}{\Phi(a_i, \hat{\mu}^b_{R,a_1}, 1) - \Phi(a_{i-1}, \hat{\mu}^b_{R,a_1}, 1)}$$

and its asymptotic property is expressed in the following proposition:

Proposition 2.4 $VC^R_{a_1}$ *criterion is consistent, i.e. the probability of selecting the best* $G(a_1, R)$ *grid tends to* 1 *when* $n \to \infty$.

3 Ongoing work: univariate Gaussian mixtures

3.1 Notations

After having considered a single Gaussian, the next step is to consider the more complex case where univariate Gaussian mixtures are involved. Thus, we assume now that each observation

$x_i \in \mathbb{R}$ $(i = 1, \ldots, n)$ arises from a univariate K-Gaussian mixture of density

$$f(x; \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \phi(x; \mu_k, \sigma_k^2), \;\; \pi_k > 0, \;\; \sum_{k=1}^{K} \pi_k = 1,$$

in which μ_k denotes the mean of the k-th component, σ_k^2 is its variance and $\boldsymbol{\theta}$ is the vector that contains all the parameters, thus $\boldsymbol{\theta} = (\pi_1, \ldots, \pi_K, \mu_1, \ldots, \mu_K, \sigma_1^2, \ldots, \sigma_K^2)$. Moreover, as the observations have real values like in the previous case, we can adopt the same notation for the grids considered.

3.2 Theoretical results

3.2.1 Identifiability

In this general case we are able to set a sufficient condition that assures identifiability, which is a consequence of Proposition 11.5 contained in Valiant (2012). It leads to the following proposition.

Proposition 3.1 *Mixtures of K Gaussian distributions for binned data are identifiable for $R > 4K - 3$.*

3.2.2 Other properties as future work

The previous proposition is only a starting point for our research in this context. In fact we are investigating the theoretical properties of the MLE for binned data and we are researching some criteria allowing to select a grid candidate among a family of sensible grids candidates. We expect that the estimates will have the same behaviour of those founded for the single Gaussian situation, but, due to the more complex form of the densities involved, the mathematical tools to be employed may be more advanced. Finally, once resolved this univariate case we will pass to the multivariate one, where new challenges will appear. In particular, the question of the number of non-empty bins when increasing the dimension will be addressed as a solution for limiting the computer memory impact of binned data even in the multidimensional case.

Bibliography

Cadez, I. V., Smyth, P., McLachlan, G. J. & McLaren, C. E. (2002). Maximum likelihood estimation of mixture densities for binned and truncated multivariate data. Machine Learning, 47(1), 7-34.

McLachlan, G. J. & Jones, P. N. (1988). Fitting mixture models to grouped and truncated data via the EM algorithm. Biometrics, 571-578.

Dempster, A. P., Laird, N. M., Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B (Methodological), 39(1), 1-22.

Valiant, G. J. (2012). Algorithmic approaches to statistical questions (Doctoral dissertation, UC Berkeley).

Estimation alternative des paramètres d'un mélange de régressions binaires

Benjamin Auder [1] & Élisabeth Gassiat [2] Mor-Absa Loum [3]

[1] *Laboratoire de Mathématiques d'Orsay, Université Paris-Saclay, benjamin.auder@universite-paris-saclay.fr* [2] *Laboratoire de Mathématiques d'Orsay, Université Paris-Saclay, elisabeth.gassiat@universite-paris-saclay.fr* [3] *Laboratoire de Mathématiques d'Orsay, Université Paris-Saclay, morabsa.loum5@gmail.com*

Résumé. Considérons un mélange de modèles de régression linéaire généralisée à sorties binaires : $\mathbb{P}(Y = 1|X = x) = \sum_{k=1}^{K} \omega_k g(\langle x, \beta_k \rangle + b_k)$, dont les paramètres sont traditionnellement estimés en maximisant la vraisemblance. Nous proposons une estimation alternative basée sur l'adéquation des moments croisés empiriques avec leurs analogues théoriques. L'identitifiabilité, la consistance ainsi que la normalité asymptotique sont démontrées. Les simulations effectuées sont très encourageantes quant à l'intérêt pratique de la méthode, implémentée dans un package R disponible sur le CRAN.

Mots-clés. Mélange, régression binaire, moments

Abstract. We consider finite mixtures of generalized linear models with binary output: $\mathbb{P}(Y = 1|X = x) = \sum_{k=1}^{K} \omega_k g(\langle x, \beta_k \rangle + b_k)$, which parameters are usually estimated by maximizing the likelihood. We propose an alternative estimation based on the adequation of the empirical cross-moments with their theoretical counterparts. Identifiability, consistency as well as asymptotic normality are proved. Simulations run are very promising concerning the practical interest of the method, implemented in a R package available on CRAN.

Keywords. Mixture, binary regression, moments

1 Introduction

Le modèle de régression logistique (ou de régression linéaire généralisée) est très populaire dans divers domaines. Quand les données étudiées proviennent de plusieurs groupes latents, utiliser des modèles de mélanges est une manière courante de gérer l'hétérogénéité. Beaucoup d'algorithmes ont ainsi été développés pour estimer les paramètres de tels modèles, la plupart basés sur la maximisation de la (log-)vraisemblance via un algorithme E-M. L'objectif de ce travail est d'explorer une approche alternative d'estimation des paramètres basée sur les moments croisés jusqu'à l'ordre 3, dans le cadre des méthodes basées sur les tenseurs − voir par exemple Anandkumar et al (2014).

Soit $X \in \mathbb{R}^d$ le vecteur des covariables et $Y \in \{0,1\}$ la sortie binaire. Selon un modèle de régression binaire, la probabilité conditionnelle $\mathbb{P}(Y = 1|X = x)$ est donnée par $g(\langle \beta, x\rangle + b)$, où $\beta \in \mathbb{R}^d$ est le vecteur des coefficients de régression, $b \in \mathbb{R}$ désignant l'ordonnée en zéro. Le package que nous avons développé permet d'utiliser les liens logit et probit, où g est donnée respectivement par $g(z) = e^z/(1+e^z)$ et $g(z) = \Phi(z)$ avec Φ la fonction de répartition de la loi gaussienne standard $\mathcal{N}(0,1)$.

Si maintenant l'on souhaite modéliser des populations hétérogènes, fixons K le nombre de populations et $\omega = (\omega_1, \cdots, \omega_K)$ leurs poids tels que $\omega_j \geq 0$, $j = 1 \ldots, K$ et $\sum_{j=1}^K \omega_j = 1$. L'expression précédente se généralise naturellement en

$$\mathbb{P}(Y = 1|X = x) = \sum_{k=1}^{K} \omega_k g(\langle \beta_k, x\rangle + b)\,.$$

Notons $\theta = (\omega, \beta, b)$ l'ensemble des paramètres.

Trois résultats théoriques sont démontrés dans Auder et al (2020) : l'identifiabilité (à partir des moments croisés), la convergence et la normalité asymptotique. Cependant, nous préférons ici nous concentrer sur les aspects pratiques de la méthode d'estimation, décrite ci-après. Afin de simplifier les choses et sans perdre de généralité, la covariable X sera supposée suivre une loi gaussienne standard : $X \sim \mathcal{N}(0,1)$. Les résultats peuvent s'étendre à dautre distributions, mais ce n'est pas l'objet de cette communication.

2 Méthode d'estimation

2.1 Idée générale

Commençons par définir les moments croisés : en notant e_j le j^{eme} vecteur de la base canonique de $\mathbb{R}^d$, ceux-ci s'écrivent ainsi :

- $M_1(\theta) := E_\theta[YX]$,
- $M_2(\theta) := E_\theta\left[Y\left(X \otimes X - \sum_{j\in[d]} e_j \otimes e_j\right)\right]$, and
- $M_3(\theta) := E_\theta\left[Y\left(X \otimes X \otimes X - \sum_{j\in[d]} [X \otimes e_j \otimes e_j + e_j \otimes X \otimes e_j + e_j \otimes e_j \otimes X]\right)\right]$.

Le produit tensoriel s'effectue composante par composante, généralisant le produit matriciel. Par exemple $e_j \otimes e_j$ est une matrice et $X \otimes e_j \otimes e_j$ un tenseur d'ordre 3.

Seuls les moments basés sur la vraie valeur du paramètre ($\theta^\star$) sont utiles. Mais comme ce paramètre est inconnu, on approche empiriquement les vrais moments croisés :

$$\begin{aligned}
\widehat{M_1} &= \frac{1}{n}\sum_{i=1}^{n} Y_i X_i \\
\widehat{M_2} &= \frac{1}{n}\sum_{i=1}^{n} \Big[Y_i (X_i \otimes X_i - \sum_{j\in[d]} e_j \otimes e_j) \Big] \\
\widehat{M_3} &= \frac{1}{n}\sum_{i=1}^{n} \Big[Y_i (X_i \otimes X_i \otimes X_i - \sum_{j\in[d]} \big[X_i \otimes e_j \otimes e_j + e_j \otimes X_i \otimes e_j + e_j \otimes e_j \otimes X_i \big]) \Big].
\end{aligned}$$

D'autres part, les moments théoriques peuvent s'exprimer simplement à partir de θ :

$$\begin{aligned}
M_1(\theta) &= \sum_{k=1}^{K} \omega_k E[g'\big(\langle X, \beta_k\rangle + b_k\big)]\ \beta_k\,, \\
M_2(\theta) &= \sum_{k=1}^{K} \omega_k E[g''\big(\langle X, \mu_k\rangle + b_k\big)]\ \beta_k \otimes \beta_k\,, \\
M_3(\theta) &= \sum_{k=1}^{K} \omega_k E[g^{(3)}\big(\langle X, \beta_k\rangle + b_k\big)]\ \beta_k \otimes \beta_k \otimes \beta_k\,.
\end{aligned}$$

Voir l'article pour les détails des calculs. Il est alors naturel d'utiliser un estimateur des moindres carrés, minimisant les écart des $M_i(\theta)$ aux $\widehat{M_i}$.

2.2 Précisions

Considérant la somme de carrés suivante :

$$Q_n(\theta) = \sum_{j\in[d]} \Big\{ \widehat{M_1}[j] - M_1(\theta)[j] \Big\}^2 + \sum_{j,k\in[d]} \Big\{ \widehat{M_2}[j,k] - M_2(\theta)[j,k] \Big\}^2 + \sum_{j,k,l\in[d]} \Big\{ \widehat{M_3}[j,k,l] - M_3(\theta)[j,k,l] \Big\}^2,$$

avec $[d] = [1, \ldots, d]$ l'ensemble des entiers de 1 à d, d étant la dimension des covariables. Une première idée consiste à minimiser $Q_n(\theta)$ directement :

$$\widehat{\theta}_n = \underset{\theta\in\Theta}{\operatorname{argmin}}\ Q_n(\theta). \tag{1}$$

Les termes dans la somme contribuent inégalement, car il y a plus de combinaisons d'indices pour les moments d'ordre supérieur. On peut alors penser à pondérer chaque groupe de termes : cela mène déjà à certaines améliorations. Cependant, on peut aller plus loin en réécrivant le problème de minimisation. Définissons pour tout θ et $i = 1, \ldots, n$

$$\tilde{M}_i(\theta) = Y_i \begin{pmatrix} X_i \\ X_i \otimes X_i - \sum_{j\in[d]} e_j \otimes e_j \\ X_i \otimes X_i \otimes X_i - \sum_{j\in[d]} \big[X_i \otimes e_j \otimes e_j + e_j \otimes X_i \otimes e_j + e_j \otimes e_j \otimes X_i \big] \end{pmatrix} - \begin{pmatrix} M_1(\theta) \\ M_2(\theta) \\ M_3(\theta) \end{pmatrix}$$

comme un vecteur colonne. Posons alors le problème de minimisation suivant (Hansen 1982).

Soit W une matrice symétrique définie positive de taille $d + d^2 + d^3$. Écrivons

$$Q_n^W(\theta) = \left(\frac{1}{n} \sum_{i=1}^{n} {}^t\tilde{M}_i(\theta) \right) W \left(\frac{1}{n} \sum_{i=1}^{n} \tilde{M}_i(\theta) \right),$$

et définissons

$$\widehat{\theta}_n^W = \underset{\theta \in \Theta}{\operatorname{argmin}}\ Q_n^W(\theta). \tag{2}$$

On remarque alors que si W est la matrice identité, on retrouve $Q_n(\theta)$. Si W est diagonale on obtient la première idée mentionnée ci-dessus.

3 Algorithme

À ce stade se pose la question du choix de la matrice W. Hansen (1982) démontre que la matrice W minimisant la variance asymptotique de l'estimateur est donnée par $W(\theta^\star) = (\mathrm{E}[\tilde{M}_i(\theta)^t\tilde{M}_i(\theta)])^{-1}$. Cette matrice optimale n'est pas accessible mais peut être approchée empiriquement par

$$\hat{W}(\hat{\theta}) = \left(\frac{1}{n} \sum_{i=1}^{n} \tilde{M}_i(\hat{\theta})^t\tilde{M}_i(\hat{\theta}) \right)^{-1},$$

avec $\hat{\theta}$ une estimation préliminaire des paramètres, par exemple avec $W = Id$.

L'algorithme consiste alors à obtenir une première estimation des paramètres (avec $W = Id$ par exemple), elle-même initialisée avec les directions de la matrice β estimées depuis les données : $\mu_k = \beta_k / \|\beta_k\|$. On recalcule alors la matrice W selon la formule précédente, puis survient la seconde et dernière itération.

Algorithme d'estimation des paramètres
Entrée: X, Y, K, g
1 : Estimer les directions $\mu_1, \ldots, \mu_K$ via l'algorithme $InitDir$
2 : Minimiser $Q_n^{W_{\texttt{init}}}(\theta)$ en partant partant des directions trouvées en 1.
(S'arrêter ici si $W_{\texttt{init}}$ est considérée assez précise)
3 : Re-calculer W en utilisant les paramètres ω, β et b obtenus
4 : Ré-exécuter l'étape 2.
Sortie: Les paramètres estimés $\widehat{\theta}$

Voir l'article Auder et al (2020) concernant les détails de l'étape d'estimation des directions (un calcul algébrique qui sort du cadre de cette communication).

4 Expériences numériques

Nous comparons l'algorithme du paragraphe précédent avec celui plus classique basé sur la maximisation de la log-vraisemblance implémenté dans le package R flexmix (2019). Des jeux de données simulés sont utilisés à cet effet, de dimension 5 et 10. En dimension 5 nous simulons 2 groupes, puis 3 en dimension 10. Les données sont prises équiréparties dans chaque groupe, et les matrices β contiennent des coefficients aléatoires variant entre -4 et 4. Par exemple dans le cas $d = 10$:

$$\beta = \begin{pmatrix} 1 & 2 & -1 \\ 2 & -3 & 1 \\ -1 & 0 & 3 \\ 0 & 1 & -1 \\ 3 & 0 & 0 \\ 4 & -1 & 0 \\ -1 & -4 & 2 \\ -3 & 3 & 0 \\ 0 & 2 & 1 \\ 2 & 0 & -2 \end{pmatrix}$$

Afin de comparer les erreurs d'estimation, nous affichons dans la table suivante l'erreur L^1 sommée sur les composantes de chaque paramètre. C'est-à-dire que pour ω l'erreur affichée est $\sum_{k=1}^{K} |\omega_k - \hat{\omega}_k|$. De même pour b, et pour β en considérant cette matrice colonne par colonne. Les paramètres sont obtenus en moyennant sur $N = 100$ réplications pour différentes valeurs de n. Ensuite, on retire 2% des plus grandes valeurs, qui sont très rares et biaiseraient trop le résultat visuel − pour les deux méthodes.

Les performances sont comparables, avec un léger avantage à flexmix dans le cas logit, et à notre algorithme − package R morpheus (2020) − dans le cas probit (non montré ici). Les temps d'exécution sont en général meilleurs pour morpheus. Étant donné que l'estimation par maximum de vraisemblance est asymptotiquement optimale, c'est un résultat très encourageant.

	$d=5$				$d=10$				
n	p	β_1	β_2	b	p	β_1	β_2	β_3	b
5.10^3	1.8e-2	2.1e+0	1.6e+0	3.9e-1	4.9e-2	5.2e+0	3.8e+0	1.2e+1	7.8e+0
	6.0e-4	2.7e-1	7.0e-2	6.3e-3	1.3e-1	1.6e+2	1.8e+2	1.4e+0	4.3e+0
10^4	2.5e-2	7.1e-1	9.7e-1	4.6e-1	3.3e-2	3.9e+0	3.4e+0	7.4e+0	2.7e+0
	1.8e-3	3.3e-2	4.2e-2	6.9e-3	1.3e-1	2.2e+0	1.8e+0	4.8e-1	7.3e-2
10^5	5.8e-2	4.1e-1	2.6e-1	3.5e-2	1.7e-2	1.1e+0	6.8e-1	2.0e+0	2.7e-1
	5.9e-5	2.2e-2	1.7e-2	2.4e-3	1.3e-1	4.6e-2	9.6e-2	9.3e-2	5.5e-3
5.10^5	1.9e-2	2.0e-1	8.0e-2	9.4e-3	1.6e-2	9.1e-1	1.0e+0	7.9e-1	7.6e-2
	2.3e-4	5.0e-3	6.1e-3	3.9e-3	1.3e-1	2.8e-2	1.7e-2	1.7e-2	2.5e-3
10^6	7.0e-2	5.3e-1	4.0e-1	1.9e-2	7.5e-3	7.8e-1	8.7e-1	2.7e-1	7.8e-2
	7.0e-5	2.5e-3	7.0e-3	2.5e-3	1.3e-1	5.4e-2	2.0e-2	1.0e-2	3.4e-3

Table 1: Lien *logit*. Somme des erreurs pour notre algorithme (en haut) et flexmix (en bas) moyennées sur $N = 100$ réplications, pour des valeurs croissantes de n.

Bibliographie

A. Anandkumar et al. (2014), Tensor decompositions for learning latent variables models, *Journal of machine learning*, 15, 2773–2832.

B. Auder et al. (2020), Least squares moment identification of binary regression mixtures models, *Arxiv: https://arxiv.org/abs/1811.01714v2*, submitted.

L. P. Hansen (1982), Large sample properties of generalized method of moments estimators, *Econometrica*, 50, 1029–1054.

B. Gruen et al. (2019), flexmix: Flexible Mixture Modeling, *CRAN: https://CRAN.R-project.org/package=flexmix*.

B. Auder et M-A. Loum (2020), morpheus: Estimate Parameters of Mixtures of Logistic Regressions, *CRAN: https://CRAN.R-project.org/package=morpheus*.

Cohérence de matrices aléatoires de grandes dimensions

Distribution asymptotique dans un cadre gaussien dépendant

Maxime Boucher [1*]

[*] *Université d'Orléans, Collegium Sciences et Techniques, Bâtiment de mathématiques, Rue de Chartres, B.P. 6759, 45067 Orléans cedex 2 FRANCE.*
[1] *maxime.boucher@univ-orleans.fr*

Résumé. Dans cet exposé, on introduit la τ-cohérence, notée $L_{n,\tau}$, d'une matrice aléatoire X_n de taille $n \times p$, avec p très grand devant n, définie comme étant le maximum en valeur absolue du coefficient de corrélation empirique de Pearson calculé sur les colonnes de X_n. On s'intéresse au cas où chaque ligne de X_n est une observation indépendante de loi normale dans $\mathbb{R}^p$, centrée et de matrice de covariance réduite Σ. En particulier, on suppose que Σ est définie par bande : une bande centrale de corrélation, une bande de transition asymptotiquement nulle et une partie extérieure d'indépendance. On montre, en utilisant la méthode de Chen-Stein, que l'ajout de cette transition n'impacte pas la distribution asymptotique de la cohérence. On peut montrer, sous certaines hypothèses que la τ-cohérence, correctement corrigée, admet une distribution asymptotique parfaitement définie par sa fonction de répartition.

Mots-clés. Matrice aléatoire, cohérence, corrélation de Pearson, méthode de Chen-Stein, asymptotique, grandes dimensions.

Abstract. In this presentation, we introduce the quantity $L_{n,\tau}$, called τ-coherence of a $n \times p$ random matrix where p is greater than n. It is defined to be the largest magnitude of the Pearson correlation coefficients between the columns of the random matrix. In our study, lines of the observation matrix are i.i.d observations of a p-dimensional centered and reduced Gaussian vector with Σ as correlation matrix. We suppose that Σ is divided into three bands: a central band with correlation coefficients, a transition band with asymptotically null coefficients, and an outside part with null coefficients. Using the Chen-Stein method, and under sufficient hypotheses, we can show that the τ-coherence, with correction terms, has an asymptotic behaviour defined by an explicit distribution function.

Keywords. Random matrices, coherence, Chen-Stein method, Pearson correlation, asymptotic, high dimension.

1 Résumé long

1.1 Modèle

On se propose d'étudier la distribution asymptotique de la τ-cohérence d'une matrice aléatoire. On se donne une matrice $\mathbf{X}$ d'observations de taille $n \times p$ où n et p seront très grands avec $n << p$. Chaque ligne de $\mathbf{X}$ est donc une observation de dimension p, et les lignes seront indépendantes et identiquement distribuées. On définit la τ-cohérence comme étant la valeur :

$$L_{n,\tau} := \max_{1 \leq i < j \leq p, |i-j| \geq \tau} |\rho_{ij}|,$$

où ρ_{ij} est le coefficient de corrélation empirique de Pearson entre les colonnes i et j de $\mathbf{X}$ et $\tau \in \mathbb{N}^*$. En particulier, on se place dans le cadre gaussien : c'est-à-dire que l'on suppose que chaque ligne est une observation d'un p-vecteur gaussien. On a donc le modèle suivant :

$$\left(X_k^1, X_k^2, \ldots, X_k^p\right)_{1 \leq k \leq n} \overset{i.i.d}{\sim} \mathcal{N}_p\left(0, \Sigma\right),$$

où Σ est la matrice de covariance réduite que l'on suppose définie en bande comme suit, avec τ et K deux entiers :

$$\Sigma_{k,j} = \begin{cases} r_{kj} & \text{if} \quad |k-j| < \tau \\ \epsilon_n & \text{if} \quad \tau \leq |k-j| \leq \tau + K \\ 0 & \text{if} \quad \tau + K < |k-j| \end{cases}.$$

Autrement dit, on suppose que deux composantes proches du vecteur (en terme d'indice) sont corrélées, alors que si elles sont suffisamment éloignés elles sont indépendantes. Dans notre modèle, on généralise celui de [CJ11] en ajoutant une bande de transition avec des coefficients $\epsilon_n \underset{n \to +\infty}{\longrightarrow} 0$. Cela permet de considérer une décorrélation progressive entre deux composantes du vecteur, au fur et à mesure qu'elles s'éloignent l'une de l'autre. La distribution asymptotique de L_n a été décrite dans le cas où toutes les observations sont des entrées gaussiennes indépendantes dans [CJ12]. On peut également citer [SZ14] où la convergence en loi de $L_{n,\tau}$ est étudiée sans hypothèse de normalité.

1.2 Résultat

Commençons par définir $\forall \delta \in]0,1[$, $\Gamma_{p,\delta} = \{k \in \{1, \ldots, p\} \text{ tel que } |r_{kj}| > 1-\delta \text{ pour } j \in \{1, \ldots, p\} \text{ et } k \neq j\}$. Nous avons le résultat suivant :

Proposition 1.2.1 *Soit n un entier non nul et $p = p_n$ une suite d'entiers tels que $p \underset{n \to +\infty}{\longrightarrow} +\infty$. On se donne une suite de réels $(\epsilon_n)_{n \geq 1} \in]-1,1[$. Supposons les conditions suivantes :*

Hyp 1 : $\log(p_n) = o(n^{\frac{1}{3}})$ *quand* $n \to +\infty$.

Hyp 2 : $\tau = o(p_n^t)$ *quand* $n \to +\infty$ *pour tout* $t > 0$.

Hyp 3 : $\exists \delta \in]0,1[$ *tel que* $|\Gamma_{p,\delta}| = o(p_n)$.

Hyp 4 : $\epsilon_n \underset{n\to+\infty}{\sim} \gamma\sqrt{\frac{\log(p_n)}{n}}$ *quand* $n \to +\infty$ *avec* $\gamma \in [-2,2]$.

Hyp 5 : $K = K(n) = \mathcal{O}(p_n^\nu)$ *pour* $0 < \nu < c(\gamma,\delta) < 1$.

Sous ces conditions, on peut montrer que :

$$nL_{n,\tau}^2 - 4\log(p_n) + \log(\log(p_n)) \xrightarrow[n\to+\infty]{\mathcal{L}} Z \tag{1}$$

où Z *est une variable aléatoire qui admet pour fonction de répartition* $F(y) = exp\left(-\frac{1}{\sqrt{8\pi}}e^{-\frac{y}{2}}\right)$ *pour tout* $y \in \mathbb{R}$, $c(\gamma,\delta)$ *est une constante dépendant uniquement de* γ *et* δ.

Remarques : En regardant la définition de la matrice Σ, combinée à l'hypothèse 3, on voit que nos composantes peuvent être corrélées si elles sont proches. En revanche, le nombre de composantes fortement corrélées est faible. De plus, avec les ϵ_n, la corrélation de transition est asymptotique nulle. Cela signifie que pour des n suffisamment grands, tous les coefficients ϵ_n passent sous le seuil $1-\delta$. Donc on ne rajoute pas de fortes corrélations pour n grand. Cependant, on gagne en généralité par rapport au modèle de [CJ11] puisque notre bande contenant les ϵ_n peut croître vers $+\infty$ plus rapidement que τ. On constate que l'on conserve la même loi asymptotique. Cela est intuitivement raisonnable car notre modèle, à l'asymptotique, revient au même que le leur.

1.3 Idée de la preuve

Pour étudier la distribution de la τ-cohérence, il est judicieux de choisir une nouvelle variable aléatoire. Ainsi, comme dans [CJ11], on considère une variable aléatoire intermédiaire $\tilde{V}_{n,\tau} := \underset{|i-j|\geq\tau}{max}\left|\sum_{k=1}^{n} X_k^i X_k^j\right|$. L'idée étant de décrire le comportement asymptotique de cette dernière pour ensuite revenir à la τ-cohérence grâce au résultat suivant :

Lemme 1.3.1 *Soit* τ *et* K *deux entiers. Soit* $\mathbf{X}$ *une matrice d'observation de taille* (n,p) *où* $(X^1, X^2, \ldots, X^p)$ *sont les* p *colonnes de* $\mathbb{R}^n$. *Soit* $L_{n,\tau}$, *la* τ*-cohérence de* $\mathbf{X}$ *et* $\tilde{V}_{n,\tau} = \max_{1\leq k<j\leq p, |k-j|\geq\tau} |{}^t X^k X^j|$. *On suppose que* $\log(p_n) = o(n^{\frac{1}{3}})$ *quand* $n \to +\infty$. *Alors,*

$$\frac{n^2 L_{n,\tau}^2 - \tilde{V}_{n,\tau}^2}{n} \xrightarrow[n\to+\infty]{\mathbb{P}} 0 \tag{2}$$

Pour décrire la loi asymptotique de $\tilde{V}_{n,\tau}$, on utilise la méthode de Chen-Stein dont on rappelle l'énoncé (on renvoie également vers [AGG89]):

Lemme 1.3.2 (Chen-Stein method) *Soit I un ensemble d'indices. Soit $\alpha \in I$ et B_α un sous-ensemble de I (i.e. pour tout α, $B_\alpha \subset I$). Soit $(\eta_\alpha)_{\alpha \in I}$ des variables aléatoires. Pour un $t \in \mathbb{R}$ fixé, on définit $\lambda := \sum_{\alpha \in I} \mathbb{P}(\eta_\alpha > t)$. Alors,*

$$\left| \mathbb{P}\left(\max_{\alpha \in I} (\eta_\alpha) \leq t \right) - e^{-\lambda} \right| \leq \min\left(1, \frac{1}{\lambda}\right) . (b_1 + b_2 + b_3) \tag{3}$$

où

- $b_1 = \sum_{\alpha \in I} \sum_{\beta \in B_\alpha} \mathbb{P}(\eta_\alpha > t) \mathbb{P}(\eta_\beta > t)$
- $b_2 = \sum_{\alpha \in I} \sum_{\alpha \neq \beta \in B_\alpha} \mathbb{P}(\eta_\alpha > t, \eta_\beta > t)$
- $b_3 = \sum_{\alpha \in I} \mathbb{E}\left[|\mathbb{E}[\mathbf{1}_{\eta_\alpha > t} | \sigma(\eta_\beta, \beta \in I \backslash B_\alpha)] - \mathbb{E}[\mathbf{1}_{\eta_\alpha > t}]| \right]$

Nous appliquons ce résultat pour $\eta_\alpha = \rho_{ij}$ et pour $I = \{(i,j) \in [\![1,p]\!]^2 : |i-j| \geq \tau\}$. Cela fait donc apparaître la variable $\tilde{V}_{n,\tau}$. Ensuite, il ne nous reste plus qu'à calculer λ_n dont la limite nous donnera la fonction de répartition asymptotique et pour finir contrôler les $(b_i)_{i=1,2,3}$ en montrant qu'ils sont bien tous trois asymptotiquement nuls. La difficulté principale est de gérer le coefficient b_2 qui prend en compte la dépendance entre les colonnes de $\mathbf{X}$.

Notre approche consiste à considérer une nouvelle variable aléatoire $\tilde{V}'_{n,\tau}$ qui est le maximum de la même quantité mais sur un ensemble plus petit que I en retirant de l'étude toutes les quantités de corrélation trop forte. On peut montrer que les deux variables sont bien équivalentes en probabilité pour enfin appliquer la méthode de Chen-Stein à cette dernière.

1.4 Étude Numérique

Pour illustrer notre résultat asymptotique, nous avons effectué des simulations avec le logiciel de statistiques R [R C15]. Pour réaliser nos simulations, correspondant au bon modèle de corrélation, nous avons considéré le schéma numérique suivant :

Nous générons une matrice de gaussienne $(Y_i^j)_{1 \leq i \leq n, 1 \leq j \leq p}$. Ensuite, on construit $\mathbf{X}$ avec le schéma suivant (inspiré par la moyenne mobile en série temporelle) : pour $i \in [\![1:n]\!]$ et $j \in [\![1:p]\!]$:

$$X_i^j = \sum_{k=j}^{j+K-1} \epsilon_n Y_i^k + \sum_{k=j+K}^{j+K+2\tau} r_k Y_i^k + \sum_{k=j+K+2\tau+1}^{j+2\tau+2K} \epsilon_n Y_i^k.$$

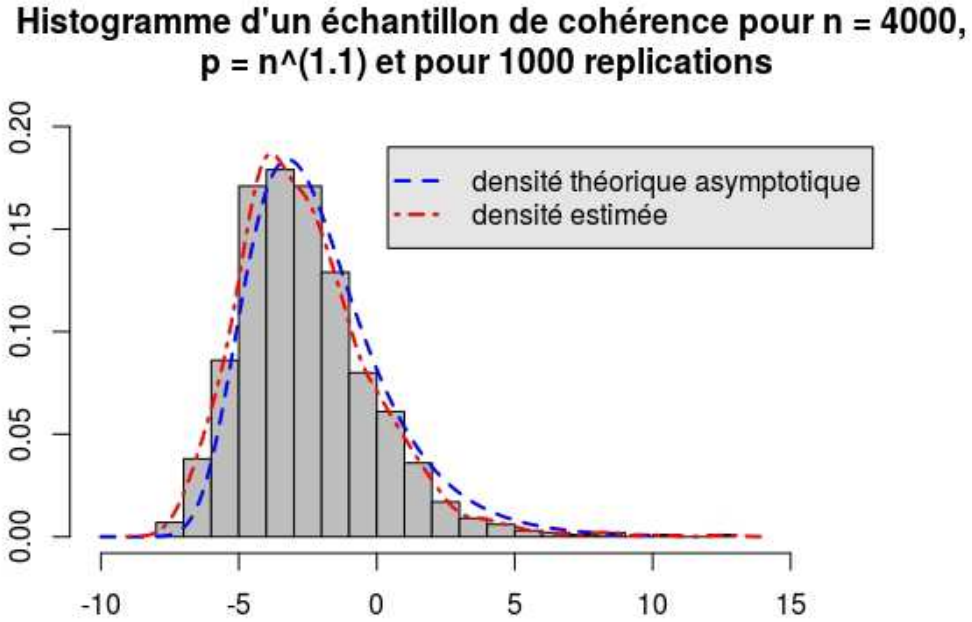


Figure 1: Histogramme, densité estimée et densité asymptotique pour un échantillon de cohérence avec $\tau = 20$, $K(n) = E\left(2 * \log(n)\right)$, $\epsilon_n = \frac{1}{n^2}$.

Pour des p permettant un calcul à l'aide d'un code classique, on obtient les résultats de la Figure 1. On peut également visualiser la convergence en regardant la distance de Kolmogorov entre la distribution empirique et asymptotique, en fonction de la taille n de nos échantillons dans la Figure 2.

Pour aller plus loin, une des difficultés principales est de gérer l'espace mémoire pour de grandes matrices. En effet, notre modèle autorise des p de l'ordre de $exp\left(n^{\frac{1}{3-u}}\right)$ avec $u > 0$. La taille des matrice d'observations ainsi créées peuvent tenir dans une mémoire vive d'un ordinateur classique. Mais les matrices de corrélations dépassent rapidement les dimensions admissible. Nous étudions actuellement des stratégies de calcul distribué afin de pouvoir calculer la cohérence sans avoir avoir à charger entièrement la matrice.

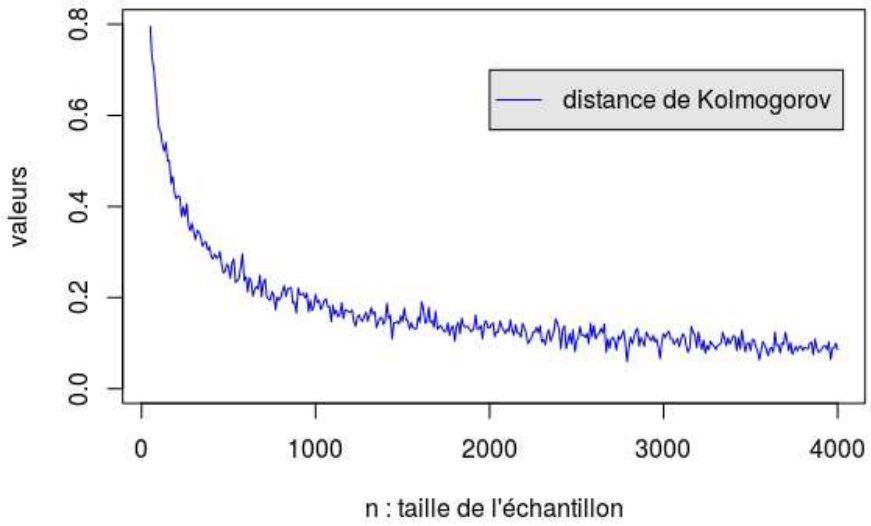


Figure 2: Valeur de la distance de Kolmogorov entre les distributions empiriques et théoriques en fonction de n pour $p = E\left(n^{1.1}\right)$, $\tau = 20$, $K(n) = E\left(2 * \log(n)\right)$, $\epsilon_n = \frac{1}{n^2}$.

References

[AGG89] R. Arratia, L. Goldstein, and L. Gordon. Two moments suffice for Poisson approximations: the Chen-Stein method. *Ann. Probab.*, 17(1):9–25, 1989.

[CJ11] T. Tony Cai and Tiefeng Jiang. Limiting laws of coherence of random matrices with applications to testing covariance structure and construction of compressed sensing matrices. *Ann. Statist.*, 39(3):1496–1525, 2011.

[CJ12] T. Tony Cai and Tiefeng Jiang. Phase transition in limiting distributions of coherence of high-dimensional random matrices. *J. Multivariate Anal.*, 107:24–39, 2012.

[R C15] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

[SZ14] Qi-Man Shao and Wen-Xin Zhou. Necessary and sufficient conditions for the asymptotic distributions of coherence of ultra-high dimensional random matrices. *Ann. Probab.*, 42(2):623–648, 2014.

Advancements in the Markov chain stock model : analysis and inference

Vlad Stefan Barbu [1] & Guglielmo D'Amico [2] & Riccardo De Blasis [3]

[1] *LMRS UMR 6085, Université de Rouen Normandie, Avenue de l'Université, BP.12, F76801 Saint-Étienne-du-Rouvray, France ; barbu@univ-rouen.fr*
[2] *Department of Pharmacy, University G. d'Annunzio of Chieti-Pescara, Italy ; g.damico@unich.it*
[3] *Doctoral School in Accounting, Management and Finance, University G. d'Annunzio of Chieti-Pescara, Italy, and Wollongong University and CMCRC research center ; riccardo.deblasis@unich.it*

Résumé. Dans cette présentation, basée principalement sur Barbu et al. (2017), nous nous intéressons aux applications des techniques statistiques pour les chaînes de Markov en mathématiques financières. L'évolution temporelle du facteur de croissance des dividendes d'un stock est modélisée par une chaîne de Markov. Nous nous intéressons à l'estimation des deux premiers moments du prix d'un stock et aussi à la prédiction du prix du stock dans un horizon de n unités de temps. Ce travail représente une continuation de la modélisation pour les stocks basée sur les chaînes de Markov proposée par Ghezzi et Piccardi (2003). Nous donnons les résultats théoriques sur la consistance et la normalité asymptotique des quantités estimées et nous appliquons nos résultats à des données réelles. Les techniques statistiques pour les chaînes de Markov sont principalement basées sur Sadek et Limnios (2002).

Ces résultats ont été intégrés dans un cadre semi-markovien dans D'Amico (2013), où l'hypothèse semi-markovienne a été considérée et validée sur des données réelles. Une généralisation supplémentaire a été fournie en D'Amico (2016), où un modèle semi-markovien à espace d'état continu a été considéré.

Mots-clés. Modélisation des dividendes, chaînes de Markov, statistique inférentielle, propriétés asymptotiques

Abstract. In this presentation, mainly based on Barbu et al. (2017), we are interested in applications of statistical techniques for Markov chains in financial mathematics. We have modelled through a Markov chain the time evolution of the dividend growth factor of a stock. We were interested in estimating the first two moments of the price of the stock and also in forecasting the price of the stock within n time units. This work represents further advancements of the Markov chain stock model proposed in Ghezzi and Piccardi (2003). We give theoretical results about the consistency and asymptotic normality of the estimated quantities and apply our findings to real data. The statistical techniques for Markov chains are mainly based on Sadek and Limnios (2002).

These results were integrated into a semi-Markov framework in D'Amico (2013), where the semi-Markov hypothesis was assumed and validated on real data. A further generalization was given in D'Amico (2016), where a continuous state space semi-Markov model was considered.

Keywords. Dividend modelling, Markov chains, statistical inference, asymptotic properties

1 The Markov chain dividend valuation model

Let $P(k)$ be the random variable giving the fundamental value of a stock at time $k \in \mathbb{N}$. Let $D(k)$ be the dividend at time $k \in \mathbb{N}$, also assumed to be a random variable, and denote by r one plus the required rate of return on the stock, assumed to be constant. The fundamental valuation analysis states that $p(k) := \mathbb{E}_k[P(k)]$ obeys the equation

$$p(k) = \frac{\mathbb{E}_k[D(k+1) + P(k+1)]}{r}, \tag{1}$$

where $\mathbb{E}_k$ is the conditional expectation given the information available up to time k.

As it is well known, if we assume that

$$\lim_{i \to +\infty} \frac{\mathbb{E}_k[P(k+i)]}{r^i} = 0, \tag{2}$$

then the solution of (1) is expressed by the series

$$p(k) = \sum_{i=1}^{+\infty} \frac{\mathbb{E}_k[D(k+i)]}{r^i}. \tag{3}$$

We analyze the second order moment of the price process. To this end, according to D'Amico (2016), if we set

$$P^2(k) := \left(\frac{D(k+1) + P(k+1)}{r} \right)^2, \tag{4}$$

then, by means of successive substitutions we get

$$\begin{aligned} P^2(k) &= \sum_{i=1}^{N} \frac{D^2(k+i)}{r^{2i}} + 2 \sum_{i=1}^{N} \sum_{j>i}^{N} \frac{D(k+i)D(k+j)}{r^{i+j}} \\ &\quad + 2 \sum_{i=1}^{N} \frac{D(k+i)P(k+N)}{r^{i+N}} + \frac{P^2(k+N)}{r^{2N}}. \end{aligned}$$

Therefore, applying the conditional expectation $\mathbb{E}_k$ we obtain

$$p^{(2)}(k) := \mathbb{E}_k[P^2(k)] = \sum_{i=1}^{N} \frac{\mathbb{E}_k[D^2(k+i)]}{r^{2i}} + 2\sum_{i=1}^{N}\sum_{j>i}^{N} \frac{\mathbb{E}_k[D(k+i)D(k+j)]}{r^{i+j}}$$
$$+2\sum_{i=1}^{N} \frac{\mathbb{E}_k[D(k+i)P(k+N)]}{r^{i+N}} + \frac{\mathbb{E}_k[P^2(k+N)]}{r^{2N}}. \quad (5)$$

In order to guarantee the dependence of the risk measure $(p^{(2)}(k))$ only on the dividend process, it is necessary that both $\lim_{N\to+\infty} \frac{\mathbb{E}_k[P^2(k+N)]}{r^{2N}} = 0$ and $\lim_{N\to+\infty} \sum_{i=1}^{N} \frac{E_k[D(k+i)P(k+N)]}{r^{i+N}} = 0$. In this case, the solution of (5) would be

$$p^{(2)}(k) = \sum_{i=1}^{+\infty} \frac{\mathbb{E}_k[D^2(k+i)]}{r^{2i}} + 2\sum_{i=1}^{+\infty}\sum_{j>i} \frac{\mathbb{E}_k[D(k+i)D(k+j)]}{r^{i+j}}. \quad (6)$$

Formula (6) is the fundamental formula for the risk of the price process.

In order to be able to evaluate (3) and (6), we need to specify a model for the dividend process. In Ghezzi and Piccardi (2003) it is assumed that dividends satisfy the difference equation

$$D(k+1) = G(k+1)D(k), \quad (7)$$

where $\{G(k)\}$ is the dividend growth factor described by a Markov chain.

2 The computation of the moments

For clarity of exposition, we limit ourselves for the moment to the simplest case of a two state Markov chain with state space $E = \{g_1, g_2\}$. The generalization to a general finite state space Markov chain is straightforward.

Let $\mathbf{P} = (p_{ij})_{i,j\in E}$ be the one step transition probability matrix of this Markov chain. The combination of Equations (3) and (7) yields

$$p(k) = d(k)\sum_{i=1}^{+\infty} \frac{\mathbb{E}_k[\prod_{j=1}^{i} G(k+j)]}{r^i} =: d(k)\psi_1(g(k)), \quad (8)$$

where $d(k)$ and $g(k)$ are the values at time k of the dividend process and of the growth dividend process, respectively. The quantity $\psi_1(g(k))$ is the so called price-dividend ratio.

The following assumption will be needed in the sequel

$$\mathbf{A1} : \overline{g} := \max(p_{11}g_1 + p_{12}g_2, p_{21}g_1 + p_{22}g_2) < r. \quad (9)$$

Proposition 1 (see Ghezzi and Piccardi (2003)). *If A1 holds true, the pair* $(\psi_1(g_1), \psi_1(g_2))$ *is the unique and nonnegative solution of the linear system*

$$\begin{aligned} \psi_1(g_1) &= p_{11}\frac{\psi_1(g_1)g_1 + g_1}{r} + p_{12}\frac{\psi_1(g_2)g_2 + g_2}{r} \\ \psi_1(g_2) &= p_{21}\frac{\psi_1(g_1)g_1 + g_1}{r} + p_{22}\frac{\psi_1(g_2)g_2 + g_2}{r}. \end{aligned} \tag{10}$$

In order to compute $p^{(2)}(k)$, we need an additional assumption

$$\mathbf{A2} : \overline{g}^{(2)} := max(p_{11}g_1^2 + p_{12}g_2^2, p_{21}g_1^2 + p_{22}g_2^2) < r^2. \tag{11}$$

Proposition 2. *Assume that hypotheses A1 and A2 hold true. Then, the pair* $(\psi_2(g_1), \psi_2(g_2))$ *is the unique and nonnegative solution of the linear system*

$$\begin{aligned} \psi_2(g_1)\big(r^2 - p_{11}g_1^2\big) - \psi_2(g_2)p_{12}g_2^2 &= p_{11}g_1^2\big(1 + 2\psi_1(g_1)\big) + p_{12}g_2^2\big(1 + 2\psi_1(g_2)\big) \\ \psi_2(g_2)\big(r^2 - p_{22}g_2^2\big) - \psi_2(g_1)p_{21}g_1^2 &= p_{21}g_1^2\big(1 + 2\psi_1(g_1)\big) + p_{22}g_2^2\big(1 + 2\psi_1(g_2)\big). \end{aligned} \tag{12}$$

The results have a straightforward extension to the case of an s-state Markov chain with state space $E = \{g_1, g_2, \ldots, g_s\}$. Note that the assumptions A1 and A2 should be formulated as follows :

$$\overline{g} := \max_{i \in E}\left(\sum_{j=1}^{s} p_{ij}g_j\right) < r \tag{13}$$

$$\overline{g}^{(2)} := \max_{i \in E}\left(\sum_{j=1}^{s} p_{ij}g_j^2\right) < r^2. \tag{14}$$

Let $\mathbf{I}$ be the identity matrix of dimension $s \times s$. For any $r \in \mathbb{R}^* := \mathbb{R} - \{0\}$, we define $\mathbf{I}_r := r\mathbf{I}$ and, more generally, $\mathbf{I}_r^n = \mathbf{I}_{r^n}$ and $\mathbf{I}_r^{-1} = \mathbf{I}_{r^{-1}}$. Moreover, for any $\mathbf{g} = (g_1, \ldots, g_s)^\top$, $\mathbf{g}^n = (g_1^n, \ldots, g_s^n)^\top \in (\mathbb{R}^*)^s$ with $(\,)^\top$ denoting the transpose of a vector, we denote by

$$\mathbf{I_g} \;=\; (I_g(i,j))_{i,j \in E},\; I_{\mathbf{g}}(i,j) = \begin{cases} g_i, \text{ if } i = j, \\ 0, \text{ if } i \neq j. \end{cases} \tag{15}$$

More generally, it results that $\mathbf{I}_{\mathbf{g}}^n = \mathbf{I}_{\mathbf{g}^n}$ and $\mathbf{I}_{\mathbf{g}}^{-1} = \mathbf{I}_{\mathbf{g}^{-1}}$.

Finally let $\mathbf{\Psi}_1 = (\psi_1(g_1), \ldots, \psi_1(g_n))^\top$ and $\mathbf{\Psi}_2 = (\psi_2(g_1), \ldots, \psi_2(g_n))^\top$ be the vectors of the price-dividend ratio of first and second order. Then, the systems (10) and (12) have the following matrix representation :

$$(\mathbf{I}_r - \mathbf{P} \cdot \mathbf{I_g}) \cdot \mathbf{\Psi}_1 = \mathbf{P} \cdot \mathbf{g}, \tag{16}$$

$$\left(\mathbf{I}_r^2 - \mathbf{P} \cdot \mathbf{I}_{\mathbf{g}}^2\right) \cdot \mathbf{\Psi}_2 = \mathbf{P} \cdot \big((\mathbf{g} \diamond \mathbf{g}) + 2\mathbf{\Psi}_1 \diamond (\mathbf{g} \diamond \mathbf{g})\big), \tag{17}$$

where $\cdot$ denotes the usual row by column matrix product and $\diamond$ denotes the Hadamard element by element product. When no confusion is possible, we will omit writing $\cdot$ for the usual row by column matrix product.

Note that, according to Proposition 1, the system (10) (or equivalently (16)) has a unique solution. Consequently, the matrix $(\mathbf{I}_r - \mathbf{P} \cdot \mathbf{I}_g)$ is invertible and the solution is given by

$$\mathbf{\Psi}_1 = (\mathbf{I}_r - \mathbf{P} \cdot \mathbf{I}_{\mathbf{g}})^{-1} \cdot \mathbf{P} \cdot \mathbf{g}. \tag{18}$$

Similarly, according to Proposition 2, the system (12) (or equivalently (17)) has a unique solution. Consequently, the matrix $\left(\mathbf{I}_r^2 - \mathbf{P} \cdot \mathbf{I}_g^2\right)$ is invertible and the solution is given by

$$\begin{aligned} \mathbf{\Psi}_2 &= \left(\mathbf{I}_r^2 - \mathbf{P} \cdot \mathbf{I}_{\mathbf{g}}^2\right)^{-1} \cdot \mathbf{P} \cdot \left(\left(\mathbf{g} \diamond \mathbf{g}\right) + 2\mathbf{\Psi}_1 \diamond \left(\mathbf{g} \diamond \mathbf{g}\right)\right) \\ &= (\mathbf{I}_{r^2} - \mathbf{P} \cdot \mathbf{I}_{\mathbf{g}^2})^{-1} \cdot \mathbf{P} \cdot \left(\left(\mathbf{g} \diamond \mathbf{g}\right) + 2\mathbf{\Psi}_1 \diamond \left(\mathbf{g} \diamond \mathbf{g}\right)\right), \end{aligned} \tag{19}$$

where we used the fact that $\mathbf{I}_{\mathbf{g}}^2 = \mathbf{I}_{\mathbf{g}^2}$ and $\mathbf{I}_r^2 = \mathbf{I}_{r^2}$. It should be remarked that relation (19) gives an explicit formula for the second-order price-dividend ratio that in turn, after multiplication with $d^2(t)$ gives a formula for the second moment of the price process that is expressend in function of the model parameters $\mathbf{P}$ and $\mathbf{g}$.

3 The inferential analysis

For $\mathbf{x} = (x_0, x_1, \ldots, x_m)$ a sample path of the Markov chain observed up to time m, the MLE of p_{ij} is $\widehat{p}_{ij}(m) = \frac{N_{ij}(m)}{N_i(m)}$, where $N_{ij}(m) := \sum_{u=1}^{m} \mathbb{1}_{\{X_{u-1}=i, X_u=j\}}$, $N_i(m) := \sum_{u=0}^{m-1} \mathbb{1}_{\{X_u=i\}}$. The estimator of the price-dividend ratio is

$$\begin{aligned} \widehat{\mathbf{\Psi}}_1(m) &= (\widehat{\psi}_1(g_1; m), \ldots, \widehat{\psi}_1(g_n; m))^\top := \left(\mathbf{I}_r - \widehat{\mathbf{P}}(m)\mathbf{I}_{\mathbf{g}}\right)^{-1} \widehat{\mathbf{P}}(m)\mathbf{g} \\ &= \left(\mathbf{I}_r^{-1} \sum_{n=0}^{\infty} \widehat{\mathbf{P}}^n(m)\mathbf{I}_{\mathbf{g}}^n \mathbf{I}_r^{-n}\right) \widehat{\mathbf{P}}(m)\mathbf{g}. \end{aligned} \tag{20}$$

The following asymptotic results hold true.

Theorem 1. *The estimator of the price-dividend ratio proposed in (20) is*

1. *strongly consistent, as m goes to infinity, i.e.,*

$$\widehat{\mathbf{\Psi}}_1(m) \xrightarrow[m\to\infty]{a.s.} \mathbf{\Psi}_1; \tag{21}$$

2. *asymptotically normal, as m goes to infinity, i.e.,*

$$\sqrt{m}\left(\widehat{\mathbf{\Psi}}_1(m) - \mathbf{\Psi}_1\right) \xrightarrow[m\to\infty]{\mathcal{D}} \mathcal{N}_s(\mathbf{0}, \mathbf{\Sigma}_1), \tag{22}$$

where the covariance matrix $\mathbf{\Sigma}_1$ *has the form*

$$\mathbf{\Sigma}_1 = \mathbf{\Phi}_1' \mathbf{\Gamma} (\mathbf{\Phi}_1')^\top \in \mathcal{M}_{s\times s}, \tag{23}$$

where :

- $\mathbf{\Gamma} \in \mathcal{M}_{s(s-1)\times s(s-1)}$ *is the asymptotic covariance matrix of the vector* $(\sqrt{m}\,(\widehat{p}_{ij}(m) - p_{ij}))_{i=1,\dots,s,j=1,\dots,s-1}$;
- $\mathbf{\Phi}_1' = \left(\frac{\partial \Phi_1^i}{\partial p_{lk}}\right)_{i,l=1,\dots,s,k=1,\dots,s-1} \in \mathcal{M}_{s\times s(s-1)}$ *is the partial derivative matrix of* Φ_1 *with respect to* $(p_{ij}, i = 1,\dots,s, j = 1,\dots,s-1)$;
- *the function*

  $$\Phi_1 = (\Phi_1^1, \dots, \Phi_1^s) : \mathbb{R}^{s(s-1)} \to \mathbb{R}^s$$

  is defined by

  $$\Phi_1(p_{ij}, i = 1,\dots,s, j = 1,\dots,s-1) = (\mathbf{I}_r - \mathbf{P}\mathbf{I}_\mathbf{g})^{-1}\,\mathbf{P}\mathbf{g} = \mathbf{\Psi}_1, \tag{24}$$

  where, for any $i \in E$, *we express* p_{is} *as a function of the arguments of* Φ_1 *in the obvious way,* $p_{is} = 1 - \sum_{j=1}^{s-1} p_{ij}$.

Similar results hold true for the estimation of the second order price-dividend ratio and for the forecasted fundamental prices.

Note that in this presentation we have considered the estimation problem when only one trajectory of the Markov chain is observed. The case when the analyst observes several sample paths of the dividend process can be dealt with using similar techniques. These two different sampling schemes have been deeply studied in the literature of inference of Markov chains.

References

Barbu, V. S., D'Amico, G., De Blasis, R. (2017), Novel advancements in the Markov chain stock model : analysis and inference, *Annals of Finance*, 13 (2), 125–152. doi : 10.1007/s10436-017-0297-9

D'Amico, G. (2013), A semi-Markov approach to the stock valuation problem, *Annals of Finance*, 9, 589–610.

D'Amico, G. (2018), Generalized semi-Markovian dividend discount model : risk and return, *https ://arxiv.org/pdf/1605.02472.pdf*, submitted, 1-37.

Ghezzi, L. L. and Piccardi, C. (2003), Stock valuation along a Markov chain, *Appl. Math. Comput.*, 141, 385–393.

Sadek, A. and Limnios, N. (2002), Asymptotic properties for maximum likelihood estimators for reliability and failure rates of Markov chains, *Comm. Statist. Theory Methods*, 31 (10), 1837–1861.

Debiasing the Elastic Net for models with interactions

Florent Bascou[1,†] & Sophie Lèbre[2,‡] & Joseph Salmon[1,*]

[1] *IMAG, Univ. Montpellier, CNRS Montpellier, France*
[2] *Univ. Paul-Valéry-Montpellier 3, Montpellier, France*
† *florent.bascou@umontpellier.fr,* ‡ *sophie.lebre@umontpellier.fr,*
* *joseph.salmon@umontpellier.fr*

Résumé. Nous présentons un modèle de régression pénalisée et dé-biaisée pour l'estimation, en grande dimension, d'un modèle linéaire parcimonieux avec interactions. L'objectif est d'estimer conjointement le support et les coefficients dé-biaisés associés, en partant d'un estimateur de type Elastic Net. On utilise pour cela un algorithme de descente par coordonnée, qui ne nécessite pas de construire la matrice des interactions. Cette propriété est cruciale sur données réelles sachant que cette matrice peut facilement dépasser les capacités mémoires. Enfin, nous adaptons une méthode de dérivation automatique qui permet d'obtenir simultanément la solution des moindres carrés sur le support, sans avoir à résoudre a posteriori un problème de moindres carrés.

Mots-clés. Lasso, Elastic Net, Interactions, Dé-biasage, Descente par Coordonnée.

Abstract. We present a penalized and de-biased regression model to estimate, in high dimension, sparse linear models with interactions. Our aim is to jointly estimate the support and the associated de-biased coefficients, starting from an Elastic Net type estimator. The main idea is to use a coordinate descent algorithm, which does not require building the interaction matrix. This property is crucial on real data since the design matrix modeling interactions can quickly exceed memory capacities. In addition, we adapt an automatic differentiation method which allows to obtain simultaneously the least squares solution on the support, without having to solve, a posteriori, a least squares problem.

Keywords. Lasso, Elastic Net, Interaction, De-biasing, Coordinate Descent.

1 Introduction

Thanks to their interpretability, linear models are popular for many statistics tasks. Unfortunately, it turns out that the number of variables is frequently larger than the number of samples, so regularization is often required. Sparse regularization techniques leveraging the ℓ_1-norm have led to various popular estimators in the last two decades, including Lasso [Tibshirani, 1996] and Elastic Net [Zou and Hastie, 2005] among the most popular. When targeting feature interactions, such estimator become crucial: even when limited

to quadratic interactions, the number of variables is already (almost) squared, and the number of variables hence created can easily overload computers' memory.

Due to highly correlated variables, we estimate the coefficients using Elastic Net [Zou and Hastie, 2005], which allows to reduce the number of variables thanks to the ℓ_1 penalty, while taking into account the correlation thanks to the ℓ_2 penalty [Tikhonov, 1943, Hoerl and Kennard, 1970]. We adapt a coordinate descent algorithm (popularized by `glmnet` [Friedman et al., 2007, 2010]) so the interaction matrix does not need to be stored.

Finally, it is known that both Lasso and Elastic Net tend to be biased as they shrink large coefficients aggressively. To alleviate this issue, we suggest to compute a de-biased version of the coefficients along with the original coefficients [Deledalle et al., 2017]. We propose an algorithm approaching the LS Elastic Net (Elastic Net followed by a Least Squares step on the support), though in a more stable way as the naive implementation.

2 Elastic Net for interactions

2.1 Model and estimator

In the following, p is the number of features, n the number of samples, and $q = p\,(p+1)/2$ (or $p\,(p-1)/2$ depending whether we include or not the pure quadratic terms) the number of interaction features. The response vector is denoted by $y \in \mathbb{R}^n$. The Elastic Net model now reads :

$$(\mathcal{P}) \qquad \min_{\beta\in\mathbb{R}^p, \Theta\in\mathbb{R}^q} \frac{1}{2n} \|y - X\beta - Z\Theta\|_2^2 + \alpha_1 \|\beta\|_1 + \alpha_2 \|\Theta\|_1 + \frac{\alpha_3}{2} \|\beta\|_2^2 + \frac{\alpha_4}{2} \|\Theta\|_2^2 \quad . \tag{1}$$

where $\alpha_1 > 0, \ldots, \alpha_4 > 0$ are tunning parameters. The parameters α_1 and α_2 control the level of ℓ_1 penalization (resp. for the quadratic features), and the sparsity of β and Θ, while α_3 and α_4 control the level of ℓ_2 penalization and how spread out the signal is among active features.

As previously explained, we can not always handle in memory the design matrix Z (Figure 2), which leads us to reformulate the classical coordinate descent algorithm in this context. Let us remind[1] the main step in the coordinate descent algorithm to solve the Elastic Net problem is the coordinates updates for β_{j_0} and $\Theta_{j_0'}$ (for $j_0 \in [\![1, p]\!]$ and $j_0' \in [\![1, q]\!]$). This requires solving one dimensional problems of the form:

$$\arg\min_{\beta_{j_0}\in\mathbb{R}} \frac{1}{2n}\Big(y - \sum_{j=1}^{p} \beta_j x_j - \sum_{j=1}^{q} \Theta_j z_j\Big)^2 + \alpha_1|\beta_{j_0}| + \frac{\alpha_3}{2}\beta_{j_0}^2 \quad , \tag{2}$$

$$\arg\min_{\Theta_{j_0'}\in\mathbb{R}} \frac{1}{2n}\Big(y - \sum_{j=1}^{p} \beta_j x_j - \sum_{j=1}^{q} \Theta_j z_j\Big)^2 + \alpha_2|\Theta_{j_0'}| + \frac{\alpha_4}{2}\Theta_{j_0'}^2 \quad . \tag{3}$$

[1] See [Friedman et al., 2010] for details.

Proposition 2.1. *We write $\widehat{\beta}^k$ and $\widehat{\Theta}^k$ for the coefficients computed at the k-th pass over the data by the coordinate descent algorithm, and $r^k = y - X\widehat{\beta}^k - Z\widehat{\Theta}^k$ is the associated residuals. The coordinate update rules for the j_0^{th} and $j_0'^{th}$ coordinate reads*

$$\widehat{\beta}_{j_0}^{k+1} = \frac{1}{\|x_{j_0}\|^2 + n\alpha_3} \mathrm{ST}\left(x_{j_0}^\top \left(r^k + \widehat{\beta}_{j_0}^k x_{j_0}\right), n\alpha_1\right) \quad . \tag{4}$$

$$\widehat{\Theta}_{j_0'}^{k+1} = \frac{1}{\|z_{j_0'}\|^2 + n\alpha_4} \mathrm{ST}\left(z_{j_0'}^\top \left(r^k + \widehat{\Theta}_{j_0'}^k z_{j_0'}\right), n\alpha_2\right) \quad . \tag{5}$$

and ST *representing the soft-thresholding, defined for any $x \in \mathbb{R}$ by:*

$$\mathrm{ST}(x, \alpha) = (|x| - \alpha)_+ \operatorname{sign}(x) \quad . \tag{6}$$

In the previous proposition, we need to compute the $z_{j_0'}$ column of Z, which is made possible by a coordinate descent algorithm (Line 6 of Algorithm 2). Thanks to that, we can handle interactions without explicitly storing the interaction (design) matrix.

2.2 De-biasing

Unfortunately, the Lasso and the Elastic Net coefficients are biased (see [Salmon, 2017]): large coefficients are shrunk toward zero. To reduce this effect, one can perform an Least Squares step on the non-zero coefficients obtain by Elastic Net (Naive-LSEnet). Yet, this approach is limited: the interaction design matrix on the support is needed, which for large datasets could not be stored.

2.2.1 Sequentially de-biasing Elastic Net

Algorithm 1: Naive-LSEnet

Input : $[X, Z]$, y, α

1 $\widehat{\beta}, \widehat{\Theta} \leftarrow$ `Enet`([X, Z], y, α)

`// supp. estimat.`

2 $\mathrm{supp}_{\widehat{\beta}} \leftarrow$ where $(\widehat{\beta} \neq 0)$

3 $\mathrm{supp}_{\widehat{\Theta}} \leftarrow$ where$(\widehat{\Theta} \neq 0)$

4 $\widetilde{\beta}, \widetilde{\Theta} \leftarrow$ `LS`([X, Z], y, $\mathrm{supp}_{\widehat{\beta}}$, $\mathrm{supp}_{\widehat{\Theta}}$)

Output : $\widetilde{\beta}, \widetilde{\Theta}$

To cope with interactions, we instantiate Covariant LEAst-square Refitting (CLEAR) [Deledalle et al., 2017], a framework to simultaneously debias the coefficients along with the algorithm (here, coordinate descent) computing the Elastic Net solution.

Proposition 2.2. *Let us suppose that the coefficients $\widehat{\beta}^k$ and $\widehat{\Theta}^k$ are iteratively updated according to Equation (4) and Equation (5). We define the Jacobian of $\widehat{\beta}^k$ (resp. $\widehat{\Theta}^k$) applied to the residuals as $J_{\widehat{\beta}_j^{k+1}} r^k$ (resp. $J_{\widehat{\Theta}_{jj}^{k+1}} r^k$) and e_j the canonical basis vector :*

$$J_{\widehat{\beta}_j^{k+1}} r^{k+1} = \frac{\left(e_j \|x_j\|_2^2 - X^\top x_j\right)^\top J_{\widehat{\beta}^k} r^k - \left(x_j^\top Z\right)^\top J_{\widehat{\Theta}^k} r^k + x_j^\top r^k}{\|x_j\|^2 + n\alpha_3} \mathbb{1}_{\left\{\left|x_j^\top \left(r^k + \widehat{\beta}_j^k x_j\right)\right| \geqslant n\alpha_1\right\}}$$

$$J_{\widehat{\Theta}_{jj}^{k+1}} r^{k+1} = \frac{\left(e_{jj} \|z_{jj}\|_2^2 - Z^\top z_{jj}\right)^\top J_{\widehat{\Theta}^k} r^k + \left(X^\top z_{jj}\right)^\top J_{\widehat{\beta}^k} r^k + z_{jj}^\top r^k}{\|z_{jj}\|^2 + n\alpha_4} \mathbb{1}_{\left\{\left|z_{jj}^\top \left(r^k + \widehat{\Theta}_{jj}^k z_{jj}\right)\right| \geqslant n\alpha_2\right\}}$$

These updates leads to compute $\rho^{k+1} = \frac{\left\langle [X,Z][J_{\widehat{\beta}^{k+1}} r^{k+1},\ J_{\widehat{\Theta}^{k+1}} r^{k+1}]^\top ; r^{k+1} \right\rangle}{\left\| [X,Z][J_{\widehat{\beta}^{k+1}} r^{k+1},\ J_{\widehat{\Theta}^{k+1}} r^{k+1}]^\top \right\|_2^2}$.
Considering the problem Equation (1)*, the CLEAR approach reads :*

$$\widetilde{\beta}^{k+1} = \widehat{\beta}^{k+1} + \rho^{k+1} J_{\widehat{\beta}^{k+1}} r^{k+1} \ , \tag{7}$$

$$\widetilde{\Theta}^{k+1} = \widehat{\Theta}^{k+1} + \rho^{k+1} J_{\widehat{\Theta}^{k+1}} r^{k+1} \ . \tag{8}$$

This leads to Algorithm 2 where Lines 4, 8 and 10 are evaluated on the fly without Z being built. We call this method CLEAR Least Squares Elastic Net (CLEAR-LSEnet). Setting $\rho = 1$ in Lines 11 and 12 recovers the Ridge estimator associated with the Elastic Net support, instead of a Least Squares version, as in Theorem 2.2.

Algorithm 2: Coordinate Descent Epoch for CLEAR-LSEnet

input : $X \in \mathbb{R}^{n \times p}$, $y \in \mathbb{R}^n$, $\alpha = (\alpha_1, \dots, \alpha_4)^\top$, ...
param. : $\widehat{\beta}(= 0_p)$, $\widehat{\Theta}(= 0_q)$, $J_{\widehat{\beta}} r(= 0_p)$, $J_{\widehat{\Theta}} r(= 0_q)$

1 $jj = 0$; $q = p(p+1)/2$ or $p(p-1)/2$
2 **for** $j_1 = 1, \dots, p$ **do**
3 $\quad \widehat{\beta}_{j_1}^{k+1} = \frac{1}{\|x_{j_1}\|^2 + n\alpha_3} \mathrm{ST}(x_{j_1}^\top (y - r^k + \widehat{\beta}_{j_1}^k x_{j_1}), n\alpha_1)$ // β Elastic Net update
4 $\quad J_{\widehat{\beta}_{j_1}^{k+1}} r^{k+1} = \frac{\left(e_{j_1} \|x_{j_1}\|_2^2 - X^\top x_{j_1}\right)^\top J_{\widehat{\beta}^k} r^k - \left(x_{j_1}^\top Z\right)^\top J_{\widehat{\Theta}^k} r^k + x_{j_1}^\top r^k}{\|x_{j_1}\|^2 + n\alpha_3} \mathbb{1}_{\{|x_{j_1}^\top (r^k + \widehat{\beta}_{j_1}^k x_{j_1})| \geqslant n\alpha_1\}}$
5 $\quad$ **for** $j_2 = 1, \dots, q$ **do**
6 $\quad\quad z_{jj} = x_{j_1} \odot x_{j_2}$ // point-wise multiplication
7 $\quad\quad \widehat{\Theta}_{jj}^{k+1} = \frac{1}{\|z_{jj}\|^2 + n\alpha_4} \mathrm{ST}(z_{jj}^\top (y - r^k + \widehat{\Theta}_{jj}^k z_{jj}), n\alpha_2)$ // Θ Elastic Net update
8 $\quad\quad J_{\widehat{\Theta}_{jj}^{k+1}} r^{k+1} = \frac{\left(e_{jj} \|z_{jj}\|_2^2 - Z^\top z_{jj}\right)^\top J_{\widehat{\Theta}^k} r^k + \left(X^\top z_{jj}\right)^\top J_{\widehat{\beta}^k} r^k + z_{jj}^\top r^k}{\|z_{jj}\|^2 + n\alpha_4} \mathbb{1}_{\{|z_{jj}^\top (r^k + \widehat{\Theta}_{jj}^k z_{jj})| \geqslant n\alpha_2\}}$
9 $\quad\quad jj \mathrel{+}= 1$
10 $\rho^{k+1} = \frac{\langle [X,Z][J_{\widehat{\beta}^{k+1}} r^{k+1},\ J_{\widehat{\Theta}^{k+1}} r^{k+1}]^\top \mid r^{k+1} \rangle}{\|[X,Z][J_{\widehat{\beta}^{k+1}} r^{k+1},\ J_{\widehat{\Theta}^{k+1}} r^{k+1}]^\top\|_2^2}$
11 $\widetilde{\beta}^{k+1} = \widehat{\beta}^{k+1} + \rho^{k+1} J_{\widehat{\beta}^{k+1}} r^{k+1}$ // β CLEAR-LSEnet update
12 $\widetilde{\Theta}^{k+1} = \widehat{\Theta}^{k+1} + \rho^{k+1} J_{\widehat{\Theta}^{k+1}} r^{k+1}$ // Θ CLEAR-LSEnet update

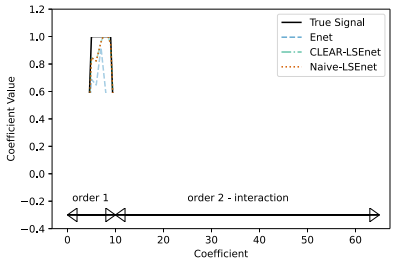
output : $\widehat{\beta}^{k+1}, \widehat{\Theta}^{k+1}\ \widetilde{\beta}^{k+1}, \widetilde{\Theta}^{k+1}$

3 Numerical experiments

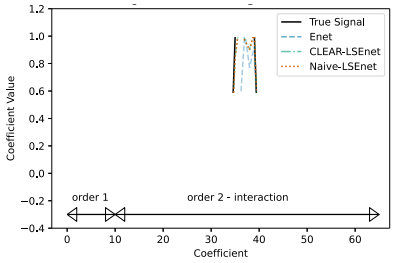
For the Naive-LSEnet, we use the Least Squares solver from `sklearn` [Pedregosa et al., 2011] on the support obtained by Elastic Net. For Figures 1 and 2, we use duality gap as stopping criterion, fixed at 10^{-4} and we set $\alpha_3 = \alpha_4 = 0.001$.

3.1 Artificial datasets

To compare Elastic Net, Naive-LSEnet and CLEAR-LSEnet, we build an artificial dataset, for which X of size $(n, p) = (60, 10)$ is drawn according to a standard Gaus-



(a) CV : $\alpha_1 = \alpha_2 \approx 0.21581$.



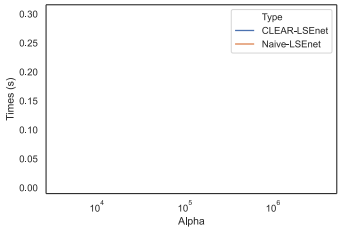
(b) CV (2D grid) : $\alpha_1 \approx 0.159$, $\alpha_2 \approx 0.267$.

Figure 1: Comparison between Elastic Net, CLEAR-LSEnet and Naive-LSEnet, with cross-validation (condensed CV) on the CLEAR-LSEnet result.

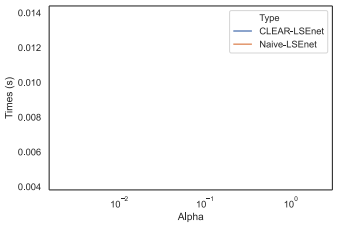
sian distribution, and we set X such that the column x_1, x_2 and x_3 are correlated (we draw x_3 from a Gaussian distribution adding $\frac{1}{2}(x_1 + x_2)$). We include the pure quadratic features leading to 55 interactions features. The true underlying signal has only five non-zero coefficients for the β and five more non-zero coefficients for Θ. Finally, we draw the noise ε from a Gaussian distribution with zero mean and a variance $1/2$. Hence, the response vector $y \in \mathbb{R}^n$ is : $y = X\beta + Z\Theta + \varepsilon$.

In Figure 1, we observe that both CLEAR-LSEnet and Naive-LSEnet estimator recover better coefficients than the Elastic Net. Indeed, both yield coefficients from the true signal than the Elastic Net on the true support. Outside the true support, CLEAR-LSEnet and Naive-LSEnet tends to give a larger coefficients than Elastic Net.

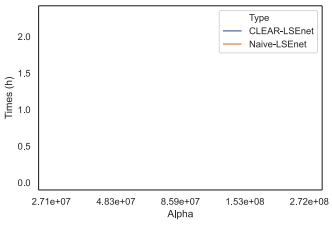
3.2 Real datasets



(a) Boston dataset
size of Z : 368, 4 Kb



(b) Diabetes dataset
size of Z : 194, 5 Kb



(c) Leukemia dataset
size of Z : 14,68 **Gb**

Figure 2: Mean time comparisons : Naive-LSEnet and CLEAR-LSEnet on real-datasets from `sklearn`. Here, α_1 and α_2 are equal.

We see Figure 2, that CLEAR-LSEnet is the same order than Naive-LSEnet. We notice that we can do Naive-LSEnet on Leukemia datasets because for those α_1 and α_2, the support is small, so we can build Z on the support. We must note that Elastic Net

from `sklearn` can handle interactions, but it requires to create and store Z, which is not always feasible. For instance with the Leukemia dataset, Z is almost 14Gb, (and possibly does not fit in memory), whereas our method can handle interactions easily here.

4 Conclusion

We presented a penalized and de-biased regression model able handle quadratic interactions in high-dimension. Future work include sensitivity analysis of the tuning parameters and algorithmic speed up, *e.g.,* following the work by Le Morvan and Vert [2018].

References

C.-A. Deledalle, N. Papadakis, J. Salmon, and S. Vaiter. CLEAR: Covariant LEAst-square Re-fitting with applications to image restoration. *SIAM J. Imaging Sci.*, 10(1): 243–284, 2017.

J. Friedman, T. J. Hastie, H. Höfling, and R. Tibshirani. Pathwise coordinate optimization. *Ann. Appl. Stat.*, 1(2):302–332, 2007.

J. Friedman, T. J. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, 33(1):1–22, 2010.

A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

M. Le Morvan and J.-P. Vert. Whinter: A working set algorithm for high-dimensional sparse second order interaction models. In *ICML*, pages 3632–3641, 2018.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, 12:2825–2830, 2011.

J. Salmon. *On high dimensional regression: computational and statistical perspectives.* Habilitation à diriger des recherches, ENS Paris-Saclay, 2017.

R. Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 58(1):267–288, 1996.

A. N. Tikhonov. On the stability of inverse problems. *Dokl. Akad. Nauk SSSR*, 39: 176–179, 1943.

H. Zou and T. J. Hastie. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(2):301–320, 2005.

Forêt aléatoire interprétable pour des applications industrielles

Clément Bénard [1,*] & Gérard Biau [2] & Sébastien da Veiga [3] & Erwan Scornet [4]

[1,2] *LPSM, Sorbonne Université, 4 place Jussieu, 75005 Paris, France*
[1,3] *Safran Tech, Modelling & Simulation, Rue des Jeunes Bois, Châteaufort, 78114 Magny-Les-Hameaux, France*
[4] *CMAP, École Polytechnique, route de Saclay 91128 Palaiseau, France*
[*] *clement.benard@safrangroup.com*

Résumé. Nous introduisons SIRUS (Ensemble de Règles Stable et Interprétable), un algorithme stable d'apprentissage de règles pour la régression, qui prend la forme d'une liste de règles simple et compacte. Les algorithmes d'apprentissage à l'état de l'art sont souvent qualifiés de "boîtes noires" à cause du grand nombre d'opérations impliquées dans leur mécanisme de prédiction. Malgré leur excellente capacité prédictive, ce manque d'interprétabilité peut être très restrictif lorsque des décisions critiques sont en jeu. D'un autre côté, les algorithmes avec une structure simple, typiquement les arbres de décisions, les modèles de règles, ou les modèles linéaires parcimonieux, sont réputés instables. Ce défaut rend les conclusions de l'analyse statistique non-fiables, ce qui limite fortement leur usage opérationnel. SIRUS est donc conçu pour produire des modèles à la fois simple et stable. Notre algorithme se base sur les forêts aléatoires et conserve leur bonne capacité de prédiction. L'efficacité de SIRUS est prouvée aussi bien empiriquement (à travers des expériences) que théoriquement (en démontrant la stabilité asymptotique). Le package `sirus` fournit une implémentation en `R/C++` disponible sur le `CRAN`.

Mots-clés. interprétabilité, forêts aléatoires, stabilité, règles

Abstract. We introduce SIRUS (**S**table and **I**nterpretable **RU**le **S**et), a stable rule learning algorithm for regression, which takes the form of a short and simple list of rules. State-of-the-art learning algorithms are often referred to as "black boxes" because of the high number of operations involved in their prediction process. Despite their powerful predictivity, this lack of interpretability may be highly restrictive for applications with critical decisions at stake. On the other hand, algorithms with a simple structure—typically decision trees, rule algorithms, or sparse linear models—are well known for their instability. This undesirable feature makes the conclusions of the data analysis unreliable and turns out to be a strong operational limitation. This motivates the design of SIRUS, which combines a simple structure with a remarkable stable behavior. The algorithm is based on random forests, the predictive accuracy of which is preserved. We demonstrate the efficiency of the method both empirically (through experiments) and theoretically (with the proof of its asymptotic stability). A `R/C++` software implementation `sirus` is available on `CRAN`.

Keywords. interpretability, random forests, stability, rules

1 Introduction

Les algorithmes d'apprentissage à l'état de l'art, tels que les forêts aléatoires ou les réseaux de neurones, sont souvent critiqués pour leur aspect «boîte noire». Cette critique provient essentiellement du nombre élevé d'opérations impliquées dans leur mécanisme de prédiction, qui empêche d'appréhender comment les entrées se combinent pour générer les prédictions. Ce manque de transparence est une limitation forte pour de nombreuses applications, en particulier celles impliquant des décisions critiques. L'analyse des processus de production dans l'industrie manufacturière rentre dans cette catégorie. En effet, ces processus impliquent des phénomènes physiques et chimiques complexes qui peuvent en général être modélisés efficacement par des algorithmes d'apprentissage boîte-noire. Cependant, toute modification d'un processus de production a des conséquences lourdes à long terme, et ne peut donc pas simplement reposer sur une modélisation stochastique opaque. C'est pourquoi les algorithmes doivent être interprétables, c'est-à-dire expliciter la relation entre les entrées et la sortie pour guider l'analyse des phénomènes physiques en jeu, et finalement permettre d'améliorer l'efficacité de la production.

Bien qu'il n'y ait pas de consensus sur une définition précise de l'interprétabilité dans la littérature (Murdoch et al., 2019), il est tout de même possible de définir des principes mimimaux : simplicité, stabilité et prédictivité (Bénard et al., 2019). La simplicité d'un algorithme peut être quantifiée par le nombre d'opérations pour calculer une prédiction. Deuxièmement, Yu (2013) soutient que «l'interprétabilité a besoin de stabilité», car les conclusions d'une analyse statistique doivent être robustes aux petites perturbations de données pour avoir du sens. Finalement, une baisse significative de la capacité de prédiction par rapport à un algorithme boîte-noire signifie que le modèle manque certaines tendances dans les données et peut donc induire en erreur (Breiman, 2001b).

Les arbres de décision (Breiman et al., 1984) sont souvent présentés comme interprétables pour leur structure récursive simple. Cependant, ils sont très sensibles aux petites perturbations de données (Breiman, 2001b), ce qui limite fortement leur utilisation pratique. Les algorithmes de règles sont un autre type de méthodes non-linéaires avec une structure simple, définie comme un ensemble de règles élémentaires. Une règle est un ensemble de contraintes sur les variables explicatives, qui forme un hyperrectangle dans l'espace d'entrée, et auquel est associé une prédiction constante. Une multitude d'algorithmes de règles a été développée depuis les annéees 1970, mais la plupart d'entre eux sont limités aux problèmes de classification. RuleFit (Friedman and Popescu, 2008) et Node harvest (Meinshausen, 2010) sont deux exceptions notables. Malgré leur bonne capacité de prédiction, ces méthodes ont tendance à produire de longues et instables listes d'une cinquante de règles, limitant sérieusement leur caractère interprétable.

Notre objectif est d'introduire SIRUS (Ensemble de Règles Stables et Interprétables) pour la régression, et de montrer ainsi que les algorithmes de règles peuvent gérer efficacement les problèmes de régression en produisant des listes de règles compactes et stables. Nous nous appuierons sur Bénard et al. (2019), qui introduit SIRUS pour la classification.

2 SIRUS

Nous considérons un cadre classique pour la régression où un échantillon i.i.d. $\mathscr{D}_n = \{(\mathbf{X}_i, Y_i), i = 1, \ldots, n\}$ est observé, avec chaque $(\mathbf{X}_i, Y_i)$ distribué comme la pair générique $(\mathbf{X}, Y)$ indépendante de $\mathscr{D}_n$. Le p-uplet $\mathbf{X} = (X^{(1)}, \ldots, X^{(p)})$ est un vecteur aléatoire à valeurs dans $\mathbb{R}^p$, et $Y \in \mathbb{R}$ est la réponse. Notre objectif est d'estimer la fonction de régression $m(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ avec un petit ensemble stable de règles.

Génération des règles La première étape de SIRUS est d'apprendre une forêt aléatoire avec un grand nombre d'arbres M à partir de l'échantillon $\mathscr{D}_n$. Deux modifications de l'algorithme original de Breiman (Breiman, 2001a) permettent de stabiliser la structure de la forêt : les coupures sont restreintes aux q-quantiles empiriques des marginales $X^{(1)}, \ldots, X^{(p)}$ (typiquement $q = 10$), et la profondeur des arbres est limitée à 2. Observons maintenant que chaque nœud de chaque arbre définit un hyperrectangle dans l'espace d'entrée $\mathbb{R}^p$. On peut donc définir une règle à partir de chaque nœud comme un estimateur constant en deux morceaux. Sa valeur dépend uniquement de si le nouveau point évalué appartient ou non à l'hyperrectangle associé à la règle. Formellement, un nœud d'un arbre est représenté par un chemin $\mathscr{P}$ qui décrit comment atteindre le nœud considéré depuis la racine de l'arbre. Dans la suite, on note Π la liste finie de tous les chemins possibles. En utilisant ce principe, la première étape de SIRUS consiste à extraire un grand ensemble de chemins d'une forêt aléatoire, de l'ordre de 10^4.

Sélection des règles Dans la deuxième étape, SIRUS sélectionne les chemins pertinents de ce grand ensemble. Malgré la pertubation aléatoire de la construction des arbres, les chemins extraits sont redondants. Une grande fréquence d'apparition d'un chemin dans la forêt est synonyme d'une tendance forte et robuste dans les données, et permet donc d'identifier les bons candidats pour construire un ensemble de règles compact, stable et prédictif. La fréquence est notée $\hat{p}_{M,n}(\mathscr{P})$ pour chaque chemin $\mathscr{P} \in \Pi$. Un seuil $p_0 \in (0, 1)$ (à régler empiriquement) permet de sélectionner l'ensemble des chemins $\hat{\mathscr{P}}_{M,n,p_0} = \{\mathscr{P} \in \Pi : \hat{p}_{M,n}(\mathscr{P}) > p_0\}$. En résumé, SIRUS utilise le principe de bagging aléatoire, mais agrège la structure de la forêt elle-même au lieu des prédictions. Par définition du mécanisme d'extraction des chemins, les règles associées à $\hat{\mathscr{P}}_{M,n,p_0}$ sont linéairement dépendantes. Afin de pouvoir bien définir une agrégation linéaire des règles, $\hat{\mathscr{P}}_{M,n,p_0}$ est filtré de la façon suivante : si une règle associée à un chemin $\mathscr{P} \in \hat{\mathscr{P}}_{M,n,p_0}$ est une combinaison linéaire de règles associées à des fréquences supérieures d'apparition dans la forêt, alors $\mathscr{P}$ est supprimée de $\hat{\mathscr{P}}_{M,n,p_0}$.

Agrégation des règles Les précédentes étapes de SIRUS permettent d'obtenir un petit ensemble de règles. La règle $\hat{g}_{n,\mathscr{P}}$ associée au chemin $\mathscr{P}$ est un estimateur constant en deux morceaux : si un nouveau point $\mathbf{x}$ appartient à l'hyperrectangle $H_{\mathscr{P}} \subset \mathbb{R}^p$ induit par

$\mathscr{P}$, la règle renvoie la moyenne des Y_i pour les points d'apprentissage appartenant à $H_{\mathscr{P}}$. Si $\mathbf{x} \notin H_{\mathscr{P}}$, c'est la moyenne des Y_i pour les points à l'extérieur de $H_{\mathscr{P}}$ qui est renvoyée. Un poids positif est associé à chaque règle afin de les combiner en un estimateur de $m(\mathbf{x})$. Ces poids sont définis comme la solution de la régression avec une pénalisation L_2, où chaque variable d'entrée est une règle de $\hat{\mathscr{P}}_{M,n,p_0}$, et λ est un paramètre qui pondère la pénalisation (réglé empiriquement). Ainsi, l'estimateur final $\hat{m}_{M,n,p_0}(\mathbf{x})$ de $m(\mathbf{x})$ s'écrit

$$\hat{m}_{M,n,p_0}(\mathbf{x}) = \hat{\beta}_0 + \sum_{\mathscr{P} \in \hat{\mathscr{P}}_{M,n,p_0}} \hat{\beta}_{n,\mathscr{P}} \hat{g}_{n,\mathscr{P}}(\mathbf{x}). \tag{2.1}$$

En notant $\hat{\boldsymbol{\beta}}_{n,p_0}$ le vecteur dont les composantes sont les coefficients $\hat{\beta}_{n,\mathscr{P}}$ pour $\mathscr{P} \in \hat{\mathscr{P}}_{M,n,p_0}$, $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, $\boldsymbol{\Gamma}_{n,p_0}$ la matrice dont les lignes sont les valeurs des règles $\hat{g}_{n,\mathscr{P}}(\mathbf{X}_i)$ pour $i \in \{1, \ldots, n\}$, et $\mathbf{1_n} = (1, \ldots, 1)^T$, $(\hat{\boldsymbol{\beta}}_{n,p_0}, \hat{\beta}_0)$ est defini par

$$(\hat{\boldsymbol{\beta}}_{n,p_0}, \hat{\beta}_0) = \underset{\boldsymbol{\beta} \geq 0, \beta_0}{\operatorname{argmin}} \frac{1}{n} ||\mathbf{Y} - \beta_0 \mathbf{1_n} - \boldsymbol{\Gamma}_{n,p_0} \boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_2^2.$$

Interprétabilité Pour approndir la discussion, il convient de quantifier les trois propriétés des modèles interprétables. La simplicité est mesurée par le nombre de règles dans le modèle final. La notion de stabilité est définie à partir de l'intuition suivante : l'algorithme est stable si deux estimations indépendantes basées sur deux échantillons indépendants génèrent la même liste de règles. En notant $\hat{\mathscr{P}}'_{M,n,p_0}$ la liste finale de chemins produite par SIRUS à partir d'un échantillon indépendant $\mathscr{D}'_n$, l'indice de Dice-Sorensen $\hat{S}_{M,n,p_0}$ donne la proportion de règles partagée par $\hat{\mathscr{P}}_{M,n,p_0}$ et $\hat{\mathscr{P}}'_{M,n,p_0}$ et s'écrit

$$\hat{S}_{M,n,p_0} = \frac{2\left|\hat{\mathscr{P}}_{M,n,p_0} \cap \hat{\mathscr{P}}'_{M,n,p_0}\right|}{\left|\hat{\mathscr{P}}_{M,n,p_0}\right| + \left|\hat{\mathscr{P}}'_{M,n,p_0}\right|}.$$

En pratique, un échantillon $\mathscr{D}'_n$ n'est pas accessible et une validation-croisée est utilisée pour simuler la perturbation de données. Enfin, l'erreur du modèle est mesurée par la proportion de variance non-expliquée.

3 Expériences

Des expériences sont réalisées sur une dizaine de jeux de données publics (Dua and Graff, 2017). La capacité de prédiction de SIRUS est comparable à ses principaux concurrents, RuleFit et Node harvest, tout en produisant des modèles plus compacts. La stabilité est considérablement améliorée, comme le montre la Table 1. En outre, la Table 2 illustre SIRUS sur le jeu de données "LA Ozone" d'enregistrements météo à Los Angeles en 1976. La variation de la concentration d'ozone est modélisée en fonction des conditions météo (vent, température, humidité...).

Dataset	RuleFit	Node Harvest	SIRUS-R
Ozone	0.22	0.30	**0.62**
Mpg	0.25	0.43	**0.83**
Prostate	0.32	0.23	**0.48**
Housing	0.19	0.40	**0.80**
Diabetes	0.18	0.39	**0.66**
Machine	0.23	0.29	**0.88**
Galaxy	0.40	0.39	**0.77**
Abalone	0.31	0.38	**0.82**
Bones	0.59	0.52	**0.89**

Table 1: Stabilité pour différents jeux de données publics : proportion moyenne de règles en commun entre deux modèles d'une validation-croisée (10 folds).

4 Analyse Théorique

Les algorithmes de règles présentent une structure simple par construction, et de nombreux exemples dans la littérature illustrent leur excellente capacité prédictive. En revanche, ce type d'algorithme est généralement instable (Murdoch et al., 2019). En conséquence, l'analyse théorique est consacrée à la stabilité asymptotique de SIRUS. Nous établissons que si l'ensemble d'apprentissage est suffisamment grand, SIRUS génère systématiquement la même liste de règles à partir d'échantillons indépendants. Dans cet optique, il est nécessaire d'introduire la probabilité $p^\star(\mathscr{P})$, le pendant théorique de $\hat{p}_{M,n}(\mathscr{P})$ (la fréquence d'apparition d'un chemin $\mathscr{P}$ dans la forêt empirique) : $p^\star(\mathscr{P})$ est définie comme la probabilité que $\mathscr{P}$ appartienne à un arbre théorique CART aléatoire. L'arbre théorique n'est plus construit à partir d'un échantillon $\mathscr{D}_n$, mais uniquement à partir de la loi de probabilité de $(\mathbf{X}, Y)$. Enfin, on note $\mathcal{U}^\star = \{p^\star(\mathscr{P}) : \mathscr{P} \in \Pi\}$. Le résultat se base sur les hypothèses peu restrictives suivantes :

(H1) Le nombre de points ré-échantillonnés a_n satisfait $\lim_{n\to\infty} a_n = \infty$ et $\lim_{n\to\infty} \frac{a_n}{n} = 0$.

(H2) Le nombre d'arbres M_n satisfait $\lim_{n\to\infty} M_n = \infty$.

(H3) La variable aléatoire $\mathbf{X}$ admet une densité strictement positive f par rapport à la mesure de Lebesgue sur $\mathbb{R}^p$. De plus, pour tout $j \in \{1, \ldots, p\}$, la densité marginale $f^{(j)}$ of $X^{(j)}$ est continue, bornée, et strictement positive. Finalement la variable aléatoire Y est bornée.

Theorem 1. *Si les hypothèses (H1)-(H3) sont satisfaites, alors pour tout* $p_0 \in [0,1] \setminus \mathcal{U}^\star$ *et* $\lambda > 0$*, en probabilité,*

$$\lim_{n\to\infty} \hat{S}_{M_n,n,p_0} = 1.$$

Average Ozone = 12			**Intercept** = −7.8				
Frequency				**Rule**			**Weight**
0.29	**if**	temp < 65	**then**	Ozone = 7	**else**	Ozone = 19	0.12
0.17	**if**	ibt < 189	**then**	Ozone = 7	**else**	Ozone = 18	0.07
0.063	**if**	temp ≥ 65 & vis < 150	**then**	Ozone = 20	**else**	Ozone = 7	0.31
0.061	**if**	vh < 5840	**then**	Ozone = 10	**else**	Ozone = 20	0.072
0.060	**if**	ibh < 2110	**then**	Ozone = 16	**else**	Ozone = 7	0.14
0.058	**if**	ibh < 2960	**then**	Ozone = 15	**else**	Ozone = 6	0.10
0.051	**if**	temp ≥ 65 & ibh < 2110	**then**	Ozone = 21	**else**	Ozone = 8	0.16
0.048	**if**	vis < 150	**then**	Ozone = 14	**else**	Ozone = 7	0.18
0.043	**if**	temp < 65 & ibt < 120	**then**	Ozone = 5	**else**	Ozone = 15	0.15
0.040	**if**	temp < 70	**then**	Ozone = 8	**else**	Ozone = 20	0.14
0.039	**if**	ibt < 227	**then**	Ozone = 9	**else**	Ozone = 22	0.21

Table 2: Liste de règles générée par SIRUS pour les données "LA Ozone".

References

C. Bénard, G. Biau, S. Da Veiga, and E. Scornet. SIRUS: Making random forests interpretable. *arXiv:1908.06852*, 2019.

L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001a.

L. Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16:199–231, 2001b.

L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification and Regression Trees*. Chapman & Hall/CRC, Boca Raton, 1984.

D. Dua and C. Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

J.H. Friedman and B.E. Popescu. Predictive learning via rule ensembles. *The Annals of Applied Statistics*, 2:916–954, 2008.

N. Meinshausen. Node harvest. *The Annals of Applied Statistics*, 4:2049–2072, 2010.

W.J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu. Interpretable machine learning: Definitions, methods, and applications. *arXiv:1901.04592*, 2019.

B. Yu. Stability. *Bernoulli*, 19:1484–1500, 2013.

Prédiction de blessure sans contact chez les footballeurs professionnels

Mathieu Berthe [1] & Pierre Druilhet [2] & Stéphanie Léger [3]

[1,2,3] *Université Clermont Auvergne*
Laboratoire de Mathématiques Blaise Pascal UMR 6620 - CNRS
[1] *Mathieu.Berthe@math.univ-bpclermont.fr*
[2] *Pierre.Druilhet@math.univ-bpclermont.fr*
[3] *Stephanie.Leger@math.univ-bpclermont.fr*

Résumé. La blessure musculaire dans le sport professionnel peut avoir un impact important sur les performances de l'athlète. Les mécanismes et les facteurs de risque de la blessure sont mal connus. Par ailleurs, les blessures sont des évènements relativement rares ce qui rend la construction de modèles statistiques de prévision délicate. En effet les jeux de données déséquilibrées dégradent la performance des techniques de modélisation habituelles et rendent les outils de comparaison de modèles inefficaces et biaisés. Pour pallier ces problèmes, de nombreuses méthodes ont été développées, que ce soit des méthodes dites de sampling (oversampling, undersampling, SMOTE...), des améliorations et corrections de biais sur des méthodes existantes (régression logistique pour événements rares) ou encore des méthodes dites ensemblistes (Bagging, boosting). Nous proposons à l'aide de données réelles, basées sur le suivi longitudinal d'une équipe de football professionnel, de comparer les méthodes afin de construire le meilleur modèle prédictif. Ensuite, à l'aide de l'analyse de la sensibilité, nous mettrons en évidence certains facteurs de risque important dans le mécanisme de la blessure.

Mots-clés. Évènements rares ; Analyse de sensibilité ; Boosting ; Bagging ; SVM ; Régression logistique, Arbre de classification ; Morriss ; Sobol ;

Abstract. Muscle injury in professional sports has a huge impact on the athlete's performance. The mechanisms and risk factors for injury are largely unknown. Moreover injuries are most often rare events, which makes its prediction complicated. Indeed imbalanced dataset degrade the performance of common modeling techniques and make the tools used to compare different models inefficient and biased. In order to overcome such issues, many methods have been developed whether it be sampling methods (Oversampling, undersampling, SMOTE), upgrading and correcting bias to adjust existing methods (regression logistic for unusual events), or aggregating methods (Bagging, Boosting). Advantages and drawbacks of each method will be assessed thanks to a literature review. Every method will be compared through simulations in order to highlight the best ones. This will also allow us to highlight, using sensitivity analysis, the important risk factors in the mechanism of injury.

Keywords. Rare event ; Boosting ; Bagging ; SVM ; Neural network ; Logistic regression ; classification tree ; Sensitivity analysis ; Moriss ; Sobol ;

1 Introduction

La blessure musculaire bien que relativement rare (2.48 blessures pour 1000h d'exposition) dans le football professionnel a un impact important pour l'athlète, qui peut perdre en capacité physique. Le mécanisme des blessures musculaires est compliqué et encore mal connu. De nombreux travaux tendent à améliorer cette connaissance. Ils consistent à combiner des facteurs de risque internes, ou biologiques tels que l'âge ou le sexe et des facteurs de risque externes comme l'équipement ou le type de pelouse utilisé.
On appellera évènement rare tout évènement se produisant avec une probabilité p petite (voire très petite) ($p \leq 0.05$.). Les méthodes habituelles, comme la régression logistique, les arbres de classification ou les SVM, ont tendance à ne pas modéliser correctement les jeux de données déséquilibrées [1]. Pour améliorer la prédiction, de nombreuses améliorations ont été apportées à ces modèles [1]. Des méthodes dites de sampling apportent une amélioration de la prédiction, celles-ci consistent à modifier de façon aléatoire ou non le jeu de données, le plus souvent en rajoutant ou en supprimant des individus. Les méthodes dites d'assemblage montrent une amélioration notable de la prédiction des évènements rares. L'assemblage consiste à construire plusieurs "classifieurs" et à les agréger en un seul. Ces groupes de techniques peuvent être utilisées conjointement pour obtenir des résultats très prometteurs, ce qui engendre un nombre de méthodes possibles très important.
Nous nous proposons donc de présenter les principales méthodes et de les confronter entre elles, afin de voir lesquelles sont les plus performantes dans la prédiction des évènements rares et donc dans la prédiction des blessures sans contact. De plus nous tenterons de mettre en avant l'importance des facteurs de risque utilisés à l'aide de l'analyse de la sensibilité et plus particulièrement les méthodes de SOBOL et MORRIS.

2 Modèle prédictif pour les évènements rares

Les méthodes de prédiction habituelles ont tendance à sous estimer la probabilité des évènements rares et donc à ne pas correctement les détecter. Afin de pallier à ce problème nous avons choisi d'utiliser la régression logistique combinée à des méthodes de sampling et d'assemblage pour prédire les évènements rares.

2.1 Echantillonnage des données

Les méthodes d'échantillonnage consistent à créer, à partir d'un jeu de données Z déséquilibrées de taille N_Z, un jeu de données $\tilde{Z}$ modifiées plus équilibrées de taille $N_{\tilde{Z}}$.
Les deux méthodes les plus utilisées pour rééquilibrer les données sont l'undersampling [2], qui consiste à supprimer aléatoirement des individus de la classe majoritaire (on a alors $N_Z > N_{\tilde{Z}}$), et l'oversampling [2] qui consiste à dupliquer de manière aléatoire des individus de la classe minoritaire (on a alors $N_Z < N_{\tilde{Z}}$). De nombreuses autres méthodes plus

complexes sont également disponibles, la principale étant SMOTE [3]. SMOTE est une méthode d'oversampling qui consiste à créer selon certains critères un individu synthétique (qui n'existe pas dans la base de données) de la classe minoritaire, qui se situe entre chaque individu de la classe minoritaire et ses k-plus proches voisins de la classe majoritaire.

2.2 Méthode par assemblage

Les méthodes d'assemblage consistent à construire plusieurs "classifieurs" à partir de modèles choisis arbitrairement (arbre de classification/régression, SVM, régression logistique, etc.), puis à les regrouper par vote ou moyenne, pour obtenir la prédiction ou la classification. Prenons la méthode de Bagging (Bootstrap aggregating)[4] ; elle consiste à construire m modèles à partir de m échantillons obtenus par bootstrap et à les regrouper par vote pour la classification et par moyenne pour la régression. La seconde méthode que l'on peut citer est le boosting [6], et plus particulièrement ADABOOST [5] qui repose sur la construction itérative de "classifieurs". A chaque itération, les exemples mal classés par le "classifieur" courant sont pondérés et leurs importances augmentent pour la construction du "classifieur" suivant. Finalement, les "classifieurs" sont regroupés et pondérés en fonction de leur taux de bonne classification, pour obtenir une prédiction finale. Les méthodes d'assemblage sont très réputées pour augmenter la performance des prédictions lorsque les données sont rares.

3 Outils de mesure

Pour comparer les performances des différents modèles de prédictions, le taux global de bien classés est mal adapté aux données déséquilibrées (prévalence ≤ 0.05). En effet, prenons le cas d'un échantillon déséquilibré de taille $N = 1000$, décomposée en $N_+ = 950$ de la classe minoritaire et $N_- = 50$ la classe minoritaire. Alors, le modèle consistant à classer tous les individus dans la classe majoritaire aura un taux de bien classés$= 0.95$ (considéré comme bon) mais avec une sensibilité de 0. Nous comparerons donc la performance de nos modèles à l'aide du score de Peirce ([6]) qui est égal à $SP = \textit{sensibilité} + \textit{spécifité} - 1$. Il est compris entre -1 et 1. Nous utiliserons également le critère AUC (Area Under Curve). Ces deux critères peuvent s'obtenir à partir de la courbe ROC du classifieur.

4 Analyse de la sensibilité

L'analyse de la sensibilité permet d'étudier l'influence des variables d'entrée (facteurs de risque) sur la réponse (risque de blessure). Les deux principales méthodes sont les indices de Sobol et la méthode de Morris [7]. Nous présentons ici uniquement cette dernière. Elle consiste, à l'aide d'un plan d'expérience, à calculer pour chaque variable j μ_j^* qui mesure la variation engendré par la variable j sur la variable réponse Y et σ qui mesure

l'interaction entre les variables ou leur effet non linéaire. Plus μ^* est important plus la variable a un effet important dans le modèle, plus σ est important, plus l'interaction entre les variables est grand ou leur influence est non linéaire. Pour comparer les variables entre elle il suffit alors de regarder le graphique de Moriss (voir figure 1)

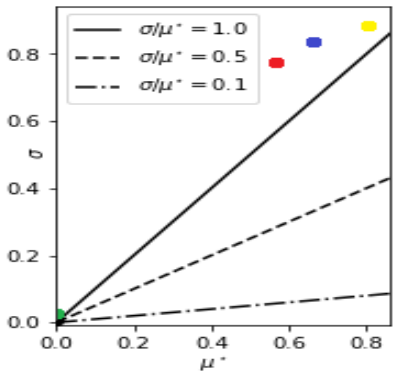


FIGURE 1 – Graphique de Morris

5 Résultats

Nous utilisons les données réelles de footballeurs professionnels évoluant dans un club de première division du championnat français de 2015 à 2020. Les variables d'entrées pour chaque joueur sont : la charge de travail et le temps de jeu cumulé sur 21 jours, le temps de récupération entre deux matchs, le risque de rechute correspondant au rapport entre le pourcentage de jours d'invalidité causé par les blessures du joueur et le pourcentage de jours moyen d'invalidité pour blessures de l'équipe. Enfin pour tenir compte du rythme et des particularités de chaque joueur nous utiliserons l'identifiant du joueur. Ces 5 variables nous permettrons de produire une probabilité qu'un joueur se blesse avant chaque rencontre que nous comparerons à la réalité. Les résultats obtenus en utilisant la régression logistique et les méthodes de bagging, random over/undersampling ainsi que la technique SMOTE sont présentés dans le tableau 2. La régression logistique seule obtient l'un des résultats les plus faibles (AUC : 0.64, Pierce : 0.36). Les résultats sont largement améliorés lorsque l'on ajoute la méthode d'oversampling notamment avec un ratio 1 :1, c'est à dire 1 non évènement pour 1 évènement (AUC : 0.80, Pierce : 0.53). L'effet de l'undersampling est plus compliqué à évaluer puisque utilisé seul les résultats obtenus ne sont pas modifiés (AUC : 0.66, Pierce : 0.38). En revanche lorsqu'il est utilisé en combinaison avec l'oversampling les résultats sont largement améliorés (AUC : 0.80, Pierce : 0.53). Le bagging de manière général stabilise et améliore les résultats de la régression logistique quasi systématiquement, sans over-undersampling (AUC : 0.75, Pierce : 0.52), avec oversampling 1 :1 (AUC : 0.77, Pierce : 0.50) et avec la combinaison oversampling et undersampling (AUC :

Bagging	Undersampling[1]	Oversampling[2]	SMOTE[3]	AUC	Peirce[4]	Sensibilité[5]	Spécificité[5]
Non	Non	Non	Non	0.64	0.36	0.81	0.56
Non	Non	1:1	Non	0.73	0.54	0.87	0.63
Non	0.5	Non	Non	0.66	0.38	0.69	0.70
Non	0.5	1:1	Non	0.80	0.53	0.88	0.67
20	Non	Non	Non	0.75	0.52	0.81	0.71
20	Non	1:1	Non	0.77	0.50	0.88	0.64
20	0.5	Non	Non	0.73	0.49	0.88	0.62
20	0.5	1:1	Non	0.80	0.59	0.81	0.78
Non	Non	Non	3:1	0.54	0.22	0.81	0.42
Non	Non	Non	1:1	0.65	0.38	0.63	0.62
Non	Non	2:1	1:1	0.81	0.60	0.93	0.67
20	Non	2:1	1:1	0.80	0.60	0.81	0.79
20	0.7	10:9	1:1	0.79	0.55	0.81	0.74

(1) : Pourcentage de la classe majoritaire conservé après undersampling aléatoire
(2) : Rapport entre la classe majoritaire et minoritaire obtenue après oversampling aléatoire
(3) : Rapport entre la classe majoritaire et minoritaire obtenue après application de la méthode SMOTE
(4) : Score de Peirce le plus élevé.
(5) : Sensibilité et spécificité pour l'indice de Peirce le plus élevé.

FIGURE 2 – Résultats de la Régression logistique

0.80, Pierce : 0.59). La méthodes SMOTE utilisée seule, n'a pas montré une amélioration des résultats, cependant les meilleurs résultats ont été obtenus en utilisant SMOTE 1 :1 en combinaison avec la méthode de random oversampling 2 :1(AUC : 0.81, Pierce : 0.60) et avec Bagging (AUC : 0.80, Pierce : 0.60) permettant d'obtenir la prédiction de 81% des blessures et 79% des joueurs non blessés.

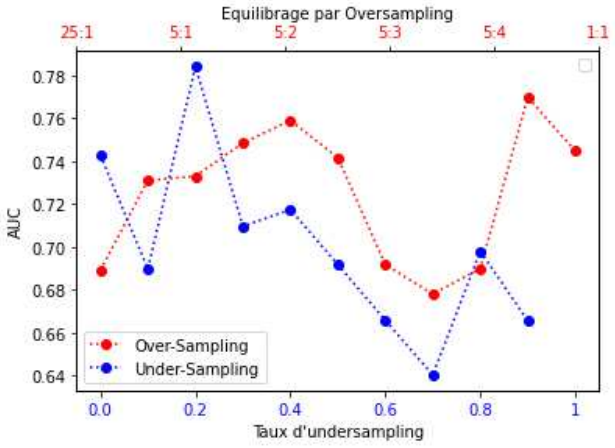


FIGURE 3 – Effet de l'over/undersampling

La figure 3 présente l'effet de différent taux d'over/undersampling, on remarque que plus le taux d'undersampling est important moins l'AUC est important puisque l'on perd de l'information sur la classe majoritaire. Pour l'oversampling, une augmentation de l'AUC se produit lorsque le rééquilibrage n'est pas trop important (5 non événement pour 1

évènement) ou lorsqu'il est vraiment important (1 non événement pour 1 évènement) entre ses deux bornes l'effet est quasi inexistant. Ces résultats sont à confirmer sur d'autre jeux de données.

L'analyse de sensibilité par la méthode de Morris (figure 2) montre que les variables indice de blessure (jaune), charge de travail (bleu) et temps de jeu en match (rouge) impactent fortement le risque de blessure mais de manière non linéaire ou avec interaction (car ils sont au dessus de la première bissectrice). En revanche on peut voir que l'impact de l'effet joueur (vert) et du temps de récupération (noir) est négligeable.

6 Conclusion

Il existe une multitude de méthodes ou combinaisons de méthodes pour modéliser les évènements rares que ce soit par choix du modèle, par méthodes d'échantillonnage, ou encore par méthodes d'assemblage. Il semble intéressant d'évaluer les différentes stratégies afin de choisir la meilleure et de mettre en avant leurs défauts et leurs qualités à l'aide des outils de comparaison adaptés.

Remerciement : L'opération Prévention du risque de blessure musculaire chez le joueur de football professionnel est cofinancée par l'Union européenne dans le cadre du Fonds Européen de Développement Régional (FEDER).

Bibliographie

[1] King, Gary et Zeng, L. (2001). Logistic Regression in Rare Event Data, *Political Analysis*, 9, pp. 137-163.
[2] Drummond, C. et Holte, R.C. , D. J. (2003). Class Imbalance , and Cost Sensitivity : Why Under-Sampling beats OverSampling.
[3] Chawla, N.V. et Bowyer, K.C. et Hall, L.O. et Kegelmeyer, W.P. (2002). SMOTE : Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, pp. 321-357.
[4] Breiman, L. (1996). Bagging predictors. *Machine Learning*,24, pp. 123-140.
[5] Freund, Y. et Schapire, R.E. (1999). A Short Introduction to Boosting. *Journal of Japanese Society for Artificial Intelligence*,14, pp. 771-780.
[6] Peirce, C.S. (1884). The numerical measure of the success of predictions. *Science*,4, pp. 453-454.
[7] Morris Max D. Factorial sampling plans for preliminary computational experiments. Technometrics, May 1991
[8] Sobol I. M. Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. *Mathematics and Computers in Simulation*, 2001.

Récupération du support pour les estimateurs pivotaux

Quentin Bertrand[1], Mathurin Massias[2], Alexandre Gramfort[1] & Joseph Salmon[3]

[1] *INRIA, Université Paris-Saclay*
[2] *University of Genova, Italy*
[3] *IMAG, Univ. Montpellier, CNRS, Montpellier, France*

Résumé. Dans le cadre de la régression parcimonieuse en grande dimension, les estimateurs pivotaux sont des estimateurs pour lesquels le paramètre de régularisation optimal est indépendant du niveau de bruit. L'estimateur pivotal canonique est le square-root Lasso, formulé tout comme ses dérivés comme un problème d'optimisation "non lisse + non lisse". Les techniques pour résoudre ces problèmes incluent le lissage du terme d'attache des données, afin de pouvoir utiliser des algorithmes proximaux rapides et efficaces. Dans ce travail, nous montrons des taux de convergence en norme $\ell_{2,\infty}$ "minimax" pour les estimateurs de type square root Lasso, simple tâche et multitâches, ainsi que non lissés et lissés. Grâce à notre analyse théorique, nous dérivons des lignes directrices sur la façon de régler l'hyperparamètre de lissage, et illustrons sur des données synthétiques leur intérêt. Cet article est une version abrégée de Massias et al. (2020).

Mots-clés. Régression, Lasso, parcimonie, optimisation convexe, lissage

Abstract. In high dimensional sparse regression, pivotal estimators are estimators for which the optimal regularization parameter is independent of the noise level. The canonical pivotal estimator is the square-root Lasso, formulated along with its derivatives as a "non-smooth + non-smooth" optimization problem. Techniques to solve these include smoothing the datafitting term, to benefit from fast efficient proximal algorithms. In this work we show minimax sup-norm convergence rates for non smoothed and smoothed, single task and multitask square-root Lasso-type estimators. Thanks to our theoretical analysis, we provide guidelines on the smoothing hyperparameter setting, and illustrate their interest on synthetic data. This work is a short version of Massias et al. 2020.

Keywords. Lasso, sparsity, convex optimization, smoothing

1 Introduction

Depuis le milieu des années 1990 et le développement du Lasso (Tibshirani, 1996), la regréssion linéaire parcimonieuse en grande dimension a été largement étudiée. L'analyse statistique du Lasso montre qu'il atteint des taux optimaux (à un facteur logarithmique près, Bickel et al. 2009); voir aussi Bühlmann and van de Geer (2011) pour une revue

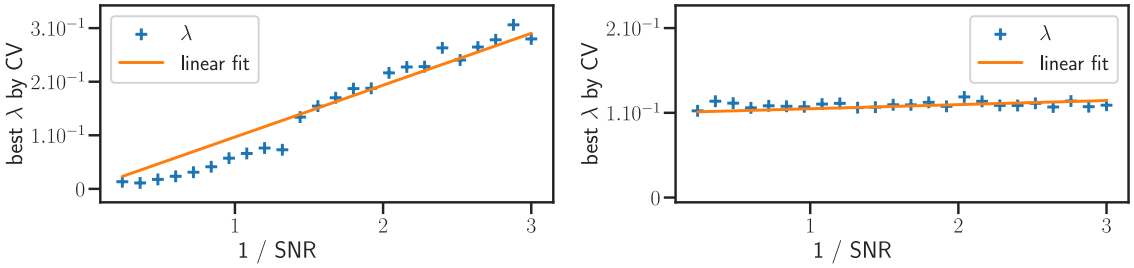


Figure 1: Paramètres de régularisation optimaux λ pour le Lasso (gauche) et le square-root Lasso (droite), déterminés par validation croisée sur l'erreur de prédiction (bleu), en fonction du niveau de bruit, pour des valeurs simulées de y. Comme indiqué par la théorie, le λ optimal du Lasso croît linéairement avec le niveau de bruit, tandis qu'il reste constant pour le square-root Lasso.

détaillée. Pourtant, cet estimateur nécessite un calibrage spécifique pour atteindre un tel taux: le paramètre de régularisation doit être proportionnel au niveau de bruit. Cette quantité est généralement inconnue en pratique, d'où le développement de méthodes adaptatives au niveau de bruit. C'est dans le but de pallier cette faiblesse que Belloni et al. (2011) ont introduit le square-root Lasso, défini pour un vecteur d'observation $y \in \mathbb{R}^n$, une matrice de régresseurs $X \in \mathbb{R}^{n \times p}$ et un paramètre de régularisation λ par

$$\arg\min_{\beta \in \mathbb{R}^p} \frac{1}{\sqrt{n}} \|y - X\beta\|_2 + \lambda \|\beta\|_1 \ . \tag{1}$$

Il a été démontré qu'il est *pivotal* pour le niveau de bruit par Belloni et al. (2011): le paramètre de régularisation optimal de leur analyse ne dépend pas du vrai niveau de bruit. Cette propriété est vérifiée en pratique comme l'illustre la Figure 1. Malgré cet avantage théorique, la résolution du square-root Lasso nécessite de résoudre un problème d'optimisation "non lisse + non lisse". Ce type de problèmes peut être résolu par programmation conique (Belloni et al., 2011) ou à l'aide d'algorithmes primaux-duaux (Chambolle and Pock, 2011), pour lesquels la convergence pratique repose souvent sur des hyper-paramètres difficiles à fixer. Une autre méthode consiste à utiliser des formulations variationnelles de normes, par exemple en réexprimant la norme ℓ_1 à partir de la formulation variationnelle suivante pour la valeur absolue: $|x| = \min_{\sigma > 0} \frac{x^2}{2\sigma} + \frac{\sigma}{2}$ (Bach et al. 2012, Sec. 5.1).

Cela conduit à une *estimation concomitante* (Huber and Dutter, 1974), c'est-à-dire des problèmes d'optimisation par rapport aux coefficients de régression *et une variable supplémentaire*. Dans le cadre de la régression parcimonieuse, l'approche concomitante séminale est le Lasso concomitant (Owen, 2007):

$$\arg\min_{\beta \in \mathbb{R}^p, \sigma > 0} \frac{1}{2n\sigma} \|y - X\beta\|_2^2 + \frac{\sigma}{2} + \lambda \|\beta\|_1 \ , \tag{2}$$

qui donne la même estimation $\hat{\beta}$ que Problème (1) si $y - X\hat{\beta} \neq 0$. Le Problème (2) est cependant plus facile à résoudre: il est convexe et le terme d'attache aux données est différentiable. Néanmoins, le terme d'attache aux données n'est toujours pas lisse, car σ peut s'approcher arbitrairement de 0: la convergence des algorithmes proximaux n'est pas garantie. Une solution consiste à introduire une contrainte $\sigma \geq \underline{\sigma}$ (Ndiaye et al., 2017), qui équivaut à *lisser* (Nesterov, 2005; Beck and Teboulle, 2012) le square-root Lasso, c'est-à-dire remplacer son terme d'attache aux donnée non lisse par une approximation lisse (voir les détails dans Massias et al. 2020).

Il existe un moyen simple de généraliser le square-root Lasso au cas multitâches (observations $Y \in \mathbb{R}^{n \times q}$): le square-root Lasso multitâches,

$$\underset{\mathrm{B} \in \mathbb{R}^{p \times q}}{\arg\min} \frac{1}{\sqrt{nq}} \|Y - X\mathrm{B}\|_F + \lambda \|\mathrm{B}\|_{2,1} \quad , \tag{3}$$

où $\|\mathrm{B}\|_{2,1}$ est la norme ℓ_1 des normes ℓ_2 des lignes. Il est clair que le square-root Lasso multitâches (Problème (3)) souffre des mêmes faiblesses numériques que le square-root Lasso. Une version plus facile à résoudre a été introduite par Bertrand et al. (2019, Prop. 21): le square-root Lasso multitâches *lissé* est obtenu en remplaçant la fonction non-lisse $\|\cdot\|_F$ par une approximation lisse, en fonction d'un paramètre $\underline{\sigma} > 0$:

$$\underset{\mathrm{B} \in \mathbb{R}^{p \times q}}{\arg\min} \left(\|\cdot\|_F \square \left(\tfrac{1}{2\underline{\sigma}} \|\cdot\|^2 + \tfrac{\underline{\sigma}}{2} \right) \right) \left(\tfrac{Y - X\mathrm{B}}{\sqrt{nq}} \right) + \lambda \|\mathrm{B}\|_{2,1} \quad , \tag{4}$$

$\square$ désignant la convolution infimale entre deux fonctions. Une autre extension du square-root Lasso au cas multitâches est le square-root Lasso multivarié [1] (van de Geer, 2016, Sec. 3.8):

$$\underset{\mathrm{B} \in \mathbb{R}^{p \times q}}{\arg\min} \frac{1}{\sqrt{nq(n \wedge q)}} \|Y - X\mathrm{B}\|_* + \lambda \|\mathrm{B}\|_{2,1} \quad . \tag{5}$$

Dans l'analyse du square root Lasso, la non-différentiabilité en 0 peut être évitée en excluant le cas pathologique où les résidus $y - X\hat{\beta}$ s'annulent. Cependant, l'analyse du square root Lasso multivarié à travers sa formulation concomitante présente une faiblesse évidente: elle nécessite d'exclure le cas où les residus ne sont pas de rang plein. Motivé par des applications numériques, Massias et al. (2018) ont introduit une borne inférieure sur la plus petite valeur propre de S ($S \succeq \underline{\sigma}\,\mathrm{Id}_n$) pour contourner ce problème. Comme observé par Bertrand et al. (2019, Sec 3.1), cela revient à lisser la norme nucléaire.

Notre objectif est de prouver des taux de convergence en norme $\ell_{2,\infty}$ et des garanties de récupération de support pour les estimateurs présentés ci-dessus, ainsi que pour leurs homologues lissés.

Travaux connexes Les propriétés statistiques du Lasso ont été étudiées sous différents cadres et hypothèses. Bickel et al. (2009) ont montré qu'avec une forte probabilité, $\|X(\hat{\beta} -$

[1] modifié ici avec une pénalité de groupe sur les lignes au lieu de la pénalité ℓ_1

$\beta^*)\|_2$ tend vers 0 au taux minimax (convergence en prédiction), tandis que Lounici (2008) a prouvé la convergence en norme ℓ_∞ et la récupération de support du Lasso (convergence en estimation), *i.e.,* un contrôle de $\|\hat{\beta} - \beta^*\|_\infty$. Ce dernier résultat a été étendu au cas multitâches par Lounici et al. (2011).

Depuis, d'autres estimateurs de type Lasso ont été proposés et étudiés, comme le square-root Lasso (Belloni et al., 2011) ou le scaled Lasso (Sun and Zhang, 2012). Dans le cas multitâches, van de Geer and Stucky (2016); Molstad (2019) ont étudié le square-root Lasso multivarié. Il a été prouvé que ces estimateurs convergent *en prédiction.* Cependant, mis à part Bunea et al. (2014) pour le square-root Lasso, nous ne connaissons pas d'autres travaux montrant une convergence en norme $\ell_{2,\infty}$ [2] de ces estimateurs.

Notation Les colonnes et les lignes de matrices sont désignées respectivement par $A_{:i}$ et $A_{i:}$. Pour tout $\mathrm{B} \in \mathbb{R}^{p\times q}$ nous définissons $\mathcal{S}(\mathrm{B}) := \{j \in [p] \,:\, ||\mathrm{B}_{j:}||_2 \neq 0\}$ le support en ligne de B. Nous écrivons $\mathcal{S}_*$ pour le support ligne par ligne de la vraie matrice de coefficients $\mathrm{B}^* \in \mathbb{R}^{p\times q}$. Les coefficients de régression estimés sont écrits $\hat{\mathrm{B}}$. La convolution infimale entre deux fonctions f_1 et f_2 de $\mathbb{R}^d$ à $\mathbb{R}$ est indiquée par $f_1 \,\square\, f_2$ et est définie pour chaque x comme $\inf\{f_1(xy) + f_2(y) : y \in \mathbb{R}^d\}$. Les normes Frobenius et nucléaires sont notées respectivement $\|\cdot\|_F$ et $\|\cdot\|_*$. Pour les matrices, $\|\cdot\|_{2,1}$ et $\|\cdot\|_{2,\infty}$ sont les lignes $\ell_{2,1}$ et $\ell_{2,\infty}$ normes, c'est-à-dire respectivement la somme et le maximum des normes de lignes. Pour une matrice positive définie symétrique S, $\|x\|_S = \sqrt{\mathrm{Tr}\, x^\top S x}$.

2 Résultats

Nos hypothèses sont classiques: bruit i.i.d., incohérence mutuelle, et grands coefficients:

Hypothèse 1. $Y = X\mathrm{B}^* + \mathrm{E}$ et les entrées de la matrice de bruit E sont des variables aléatoires i.i.d. $\mathcal{N}(0, \sigma^{*2})$,

Hypothèse 2 (Incohérence mutuelle)**.** La *matrice de Gram* $\Psi := \frac{1}{n} X^\top X$ vérifie

$$\Psi_{jj} = 1 \ , \text{ and } \max_{j' \neq j} |\Psi_{jj'}| \leq \tfrac{1}{7\alpha s}, \forall j \in [p] \ , \tag{6}$$

pour un certain $s \geq 1$ et une certaine constante $\alpha > 1$,

Hypothèse 3 (van de Geer 2016)**.** Il existe $\eta > 0$ vérifiant $\lambda \|\mathrm{B}^*\|_{2,1} \leq \eta\sigma^*$,

Nous montrons alors que le square-root Lasso lissé a les mêmes propriétés statistiques que le square-root Lasso:

[2] d'un intérêt particulier: combinés à une hypothèse de grands coefficients, elle implique l'identification du support

Proposition 4. (Voir Massias et al. (2020) pour la preuve et sa généralisation à Problème (5).) Notons $\hat{\mathrm{B}}$ l'estimateur square-root Lasso (3) ou sa version lissée (4). Sous les Hypothèses 1, 2 et 3, pour $C = \left(1 + \frac{16}{7(\alpha-1)}\right)$, $A > \sqrt{2}$ et $\lambda = \frac{2\sqrt{2}}{\sqrt{nq}}\left(1 + A\sqrt{(\log p)/q}\right)$, si $\underline{\sigma} \leq \frac{\sigma^*}{\sqrt{2}}$ alors avec probabilité au moins $1 - p^{1-A^2/2} - (1+e^2)e^{-nq/24}$,

$$\frac{1}{q}\|\hat{\mathrm{B}} - \mathrm{B}^*\|_{2,\infty} \leq C(3+\eta)\lambda\sigma^* \ . \tag{7}$$

De plus, si

$$\min_{j \in \mathcal{S}^*} \frac{1}{q}\|\mathrm{B}^*_{j:}\|_2 > 2C(3+\eta)\lambda\sigma^* \ , \tag{8}$$

alors avec la même probabilité, le support estimé

$$\hat{\mathcal{S}} := \{j \in [p] : \frac{1}{q}\|\hat{\mathrm{B}}_{j:}\|_2 > C(3+\eta)\lambda\sigma^*\} \tag{9}$$

retrouve le vrai support: $\hat{\mathcal{S}} = \mathcal{S}^*$.

References

F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Convex optimization with sparsity-inducing norms. *Foundations and Trends in Machine Learning*, 4(1):1–106, 2012.

A. Beck and M. Teboulle. Smoothing and first order methods: A unified framework. *SIAM J. Optim.*, 22(2):557–580, 2012.

A. Belloni, V. Chernozhukov, and L. Wang. Square-root Lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

Q. Bertrand, M. Massias, A. Gramfort, and J. Salmon. Handling correlated and repeated measurements with the smoothed multivariate square-root lasso. *NeurIPS*, 2019.

P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 2009.

P. Bühlmann and S. van de Geer. *Statistics for high-dimensional data*. Springer Series in Statistics. Springer, Heidelberg, 2011. Methods, theory and applications.

F. Bunea, J. Lederer, and Y. She. The group square-root Lasso: Theoretical properties and fast algorithms. *IEEE Trans. Inf. Theory*, 60(2):1313–1325, 2014.

A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vis.*, 40(1):120–145, 2011.

P. J. Huber and R. Dutter. Numerical solution of robust regression problems. In *Compstat 1974 (Proc. Sympos. Computational Statist., Univ. Vienna, Vienna, 1974)*, pages 165–172. Physica Verlag, Vienna, 1974.

K. Lounici. Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electron. J. Stat.*, 2:90–102, 2008.

K. Lounici, M. Pontil, S. van de Geer, and A. B. Tsybakov. Oracle inequalities and optimal inference under group sparsity. *Ann. Statist.*, 39(4):2164–2204, 2011.

M. Massias, O. Fercoq, A. Gramfort, and J. Salmon. Generalized concomitant multi-task lasso for sparse multimodal regression. In *AISTATS*, volume 84, pages 998–1007, 2018.

M. Massias, Q. Bertrand, A. Gramfort, and J. Salmon. Support recovery and sup-norm convergence rates for sparse pivotal estimation. *AISTATS*, 2020.

A. J. Molstad. Insights and algorithms for the multivariate square-root lasso. *arXiv preprint arXiv:1909.05041*, 2019.

E. Ndiaye, O. Fercoq, A. Gramfort, V. Leclère, and J. Salmon. Efficient smoothed concomitant lasso estimation for high dimensional regression. *Journal of Physics: Conference Series*, 904(1):012006, 2017.

Y. Nesterov. Smooth minimization of non-smooth functions. *Math. Program.*, 103(1): 127–152, 2005.

A. B. Owen. A robust hybrid of lasso and ridge regression. *Contemporary Mathematics*, 443:59–72, 2007.

T. Sun and C.-H. Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012.

R. Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 58(1):267–288, 1996.

S. van de Geer. *Estimation and testing under sparsity*, volume 2159 of *Lecture Notes in Mathematics*. Springer, 2016. Lecture notes from the 45th Probability Summer School held in Saint-Four, 2015, École d'Été de Probabilités de Saint-Flour.

S. van de Geer and B. Stucky. χ 2-confidence sets in high-dimensional regression. In *Statistical analysis for high-dimensional data*, pages 279–306. Springer, 2016.

Processus d'Ornstein-Uhlenbeck sur un arbre pour la détection de bactéries différentiellement abondantes

Antoine Bichat[1], Mahendra Mariadassou[2] & Christophe Ambroise[3]

[1] *LaMME, UEVE, Enterome – antoine.bichat@univ-evry.fr*
[2] *MaIAGE, INRAE – mahendra.mariadassou@inrae.fr*
[3] *LaMME, UEVE – christophe.ambroise@univ-evry.fr*

Résumé. En métagénomique, il est courant de réaliser des études dites d'abondance différentielle pour identifier les bactéries dont l'abondance est associée à une variable donnée, comme par exemple l'état d'un patient.

La plupart des méthodes actuelles effectuent un test indépendant par espèce. Cependant, une telle pratique ne prend en compte ni la problématique de tests multiples, ni des liens de parenté entre ces espèces et donc d'éventuelles corrélations entre leurs abondances qui pourraient entrainer une dépendance entre les tests. Ces liens de parentés sont généralement capturés par un arbre phylogénétique.

Ce travail propose une méthode de détection d'espèces différentiellement abondantes fondée sur la prise en compte de cette phylogénie. Les p-valeurs sont modélisées comme une fonction d'un processus d'Ornstein-Uhlenbeck avec sauts qui évolue le long de l'arbre phylogénétique. La méthode proposée estime la position et l'intensité des sauts, qui sont ensuite répercutés jusqu'aux feuilles – *i.e.* les espèces – pour identifier les espèces différentiellement abondantes.

Mots-clefs. Biostatistique, Métagénomique, Analyse d'abondance différentielle, Processus stochastiques, Arbres.

Abstract. In metagenomics, differential abundances studies are commonly used to identify species whose abundances are associated to a phenotype of interest, *e.g.* healthy versus diseased.

Most state of the art methods carry out independent tests to know whether each species is differentially abundant or not. However, such procedures do not take into account evolutionary relationships between species which may result in correlated abundances and then correlated p-values. These relationships are captured by the species phylogenetic tree.

This work proposes a method to detect differentially abundant species based on the phylogeny. p-values are seen as a function of an Ornstein-Uhlenbeck process with shifts on the phylogenetic tree. Our method estimates the positions and intensities of those shifts. Shifts are then propagated back to the leaves – i.e. species – in order to identify the differentially abundant ones.

Keywords. Biostatistics, Metagenomics, Differential Abundance Study, Stochastic Processes, Trees.

1 Introduction

Le microbiote peut se définir comme l'ensemble des micro-organismes présents dans un environnement donné. Depuis plusieurs années, la recherche scientifique est très active sur ce sujet et découvre fréquemment de nouvelles corrélations entre microbiote et maladies [10, 15, 17] ou entre microbiote et comportements spécifiques [4, 16].

Les données métagénomiques apportent des informations sur les espèces et les échantillons considérés dans l'étude : une table d'abondance représentant le comptage (ou la proportion) de chaque espèce dans chaque échantillon, les informations relatives aux échantillons (maladie, environnement, prise d'antibiotiques, etc) et une phylogénie.

Ici, nous nous intéressons à des problématiques de type « abondance différentielle », dans lesquelles les échantillons ou patients sont divisés en plusieurs groupes. L'objectif est d'identifier les espèces dont l'abondance varie entre les groupes.

Différentes méthodes d'abondance différentielle sont couramment utilisées sur les données métagénomiques : analyse de la variance, modèle linéaires généralisés [7, 9, 11] ou test des rangs de Wilcoxon par exemple, généralement suivi d'une procédure de correction pour la multiplicité des tests [2, 3].

Cependant, ces méthodes ne prennent pas en compte les relations de parenté qui existent entre les espèces. En particulier, on s'attend à ce que des espèces apparentées soient plus susceptibles d'être différentiablement abondantes simultanément que des espèces éloignées dans l'arbre. Nous nous proposons donc de prendre en compte cette dépendance en la modélisant par un processus d'Ornstein-Uhlenbeck.

Note. Dans ce résumé, nous utilisons le mot espèce de façon générique pour désigner les différentes composantes du microbiote qui sont testées. Cela peut être une unité taxonomique bien définie comme l'espèce, le genre, etc, ou avec une définition opérationnelle comme l'OTU (*Operational Taxonomic Unit* [6]), l'ASV (*Amplicon Sequence Variant* [5]) ou encore la MSP (*Metagenomic Species Pangenome* [13]).

2 Processus d'Ornstein-Uhlenbeck sur un arbre

Un processus d'Ornstein-Uhlenbeck est un processus gaussien qui satisfait l'équation différentielle stochastique suivante :

$$\mathrm{d}W_t = -\alpha_{\mathrm{ou}}(W_t - \beta_{\mathrm{ou}})\mathrm{d}t + \sigma_{\mathrm{ou}}\mathrm{d}B_t,$$

où α_{ou} est le taux de convergence et β_{ou} la valeur optimale du processus. De plus,

$$\mathbb{E}\left[W_t \mid W_0\right] = W_0 e^{-\alpha_{\mathrm{ou}} t} + \beta_{\mathrm{ou}}\left(1 - e^{-\alpha_{\mathrm{ou}} t}\right) \quad \text{et} \quad \mathbb{C}\mathrm{ov}\left[W_t, W_s \mid W_0\right] = \frac{\sigma_{\mathrm{ou}}^2}{2\alpha_{\mathrm{ou}}}\left(e^{-\alpha_{\mathrm{ou}}|t-s|} - e^{-\alpha_{\mathrm{ou}}(t+s)}\right).$$

Il est également possible de faire évoluer un processus d'Ornstein-Uhlenbeck sur un arbre [1]. Le long d'une branche, les paramètres du processus sont fixes. À chaque nœud,

une branche se divise (en deux dans le cas d'un arbre binaire) et le processus donne naissance à deux copies indépendantes ayant la même valeur initiale au point de branchement. Cela induit notamment une dépendance statistique entre toutes les descendants issus d'un même ancêtre. Cette dépendance est d'autant plus forte que l'ancêtre est récent. Sur la Figure 1, le processus vert partant de N_0 jusqu'à N_1 donne naissance à deux processus T_4 et T_5, jaune et bleu, lorsqu'il arrive au nœud N_1.

De plus, à chaque branchement, un changement dans les paramètres du processus est susceptible d'advenir. Dans ce cas, le processus garde la même valeur au nœud mais continue sa trajectoire avec les nouveaux paramètres. C'est le cas en N_2 dans la Figure 1 où le processus orange a subi un changement dans sa valeur optimale β_{ou} par rapport au processus rouge : la trajectoire est continue et le processus dérive vers la nouvelle valeur optimale.

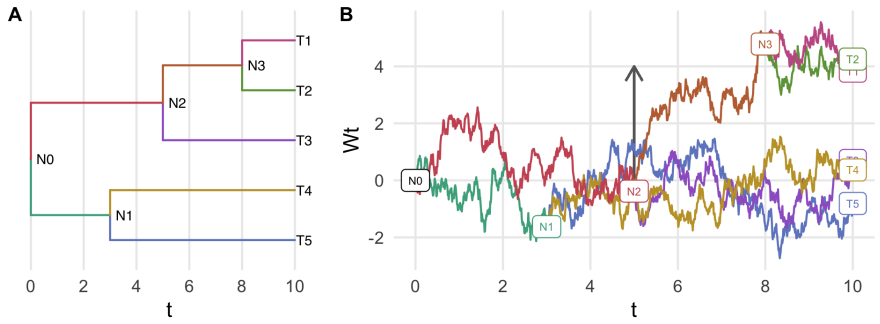


FIGURE 1 – Exemple d'un processus d'Ornstein-Uhlenbeck sur un arbre à 5 feuilles. À chaque branchement, le processus se scinde en deux processus indépendants ayant la même valeur initiale. Les paramètres sont conservés sauf lors d'un saut dans la valeur optimale, comme en N_2.

En notant
— $\text{pa}(i)$ le nœud parent de i,
— t_i la distance entre la racine et le nœud i,
— $t_{i,j}$ la distance entre la racine et le premier ancêtre commun à i et j,
— $d_{i,j} = t_i + t_j - 2t_{i,j}$ la distance entre les nœuds i et j,
— $\ell_i = t_i - t_{\text{pa}(i)}$ la distance entre le nœud i et son parent,
on a

$$X_i | X_{\text{pa}(i)} \sim \mathcal{N}\left(X_{\text{pa}(i)} e^{-\alpha_{\text{ou}}\ell_i} + \beta_{\text{ou},i}\left(1 - e^{-\alpha_{\text{ou}}\ell_i}\right), \frac{\sigma^2_{\text{ou}}}{2\alpha_{\text{ou}}}\left(1 - e^{-2\alpha_{\text{ou}}\ell_i}\right)\right)$$

et

$$\mathbb{C}\text{ov}\left[X_i, X_j\right] = \frac{\sigma^2_{\text{ou}}}{2\alpha_{\text{ou}}}\left(1 - e^{-2\alpha_{\text{ou}}t_{i,j}}\right) \times e^{-\alpha_{\text{ou}}d_{i,j}}.$$

3 Détection d'espèces différentiellement abondantes

Nous disposons des abondances de m espèces disposées sur une phylogénie $\mathcal{T}$ à n branches. Notons $T \in \{0,1\}^{m\times n}$ sa matrice d'incidence définie par $T_{i,j} = 1$ si et seulement si la branche j est sur le chemin menant de la racine à la feuille i.

Notre procédure commence classiquement par effectuer un test d'analyse différentielle par espèce (Wilcoxon, Kruskal-Wallis...) pour obtenir le vecteur $\mathfrak{p} \in [0,1]^m$ des p-valeurs.

Nous transformons ensuite nos p-valeurs en z-scores via la fonction de répartition inverse de la loi normale : $\mathfrak{z}_i = \Phi^{-1}(\mathfrak{p}_i)$.

Sous H_0, $\mathfrak{p}_i \sim \mathcal{U}([0,1])$, ainsi $\mathfrak{z}_i \sim \mathcal{N}(0,1)$. Sous H_1, $\mathfrak{p}_i \preccurlyeq \mathcal{U}$, ce que nous modélisons par $\mathfrak{z}_i \sim \mathcal{N}(\mu_i, 1)$ avec $\mu_i < 0$.

Considérons maintenant que $\mathfrak{z}$ est un vecteur gaussien issue de la réalisation d'un processus d'Ornstein-Uhlenbeck avec sauts dans la valeur optimale sur $\mathcal{T}$:

$$\mathfrak{z} \sim \mathcal{N}_m(\mu, \Sigma),$$

où $\Sigma_{i,j} = \frac{\sigma_{\text{ou}}^2}{2\alpha_{\text{ou}}}\left(1 - e^{-2\alpha_{\text{ou}} t_{i,j}}\right) e^{-\alpha d_{i,j}}$ avec σ_{ou} tel que $\Sigma_{i,i} = 1$, pour assurer $\mathfrak{z}_i \sim \mathcal{N}(0,1)$ sous H_0.

Nous commençons par déterminer la localisation et l'intensité des sauts $\beta \in \mathbb{R}^n$ sur l'arbre afin d'estimer les valeurs moyennes aux feuilles via la relation $\mu = T\beta$.

Dans notre modélisation, la contrainte $\mu_i < 0$ se traduit par la contrainte convexe suivante :

$$\beta \in \mathcal{C}_n = \left\{X \in \mathbb{R}^n \text{ tel que } TX \in \mathbb{R}_-^m\right\}.$$

En imposant une contrainte supplémentaire de parcimonie ℓ_1 sur les sauts pour remédier au fait que T n'est pas de plein rang, le problème se ramène au problème d'optimisation convexe suivant :

$$\hat{\beta} = \text{argmin}_{\beta\in\mathcal{C}_n} \left(\mathfrak{z} - T\beta\right)^T (\Sigma)^{-1} \left(\mathfrak{z} - T\beta\right) + \lambda\|\beta\|_1.$$

Nous résolvons le lasso contraint grâce à un algorithme du *shooting* [8, 12], dont chaque itération est en $\mathcal{O}(nm)$. Une fois ceci fait, nous débiaisons l'estimateur [19] afin d'obtenir des intervalles de confiance sur les sauts, qui sont ensuite propagés par la matrice d'incidence et permettent d'obtenir un vecteur de p-valeurs lissées pour les feuilles.

La procédure complète a été implémentée dans le package `R` [14] *zazou*, qui prend en entrée le vecteur des p-valeurs et renvoie le vecteur des p-valeurs lissées. Appliqué à des données simulées, *zazou* atteint des performances meilleures que des tests indépendants ou que d'autres méthodes hiérarchiques [18].

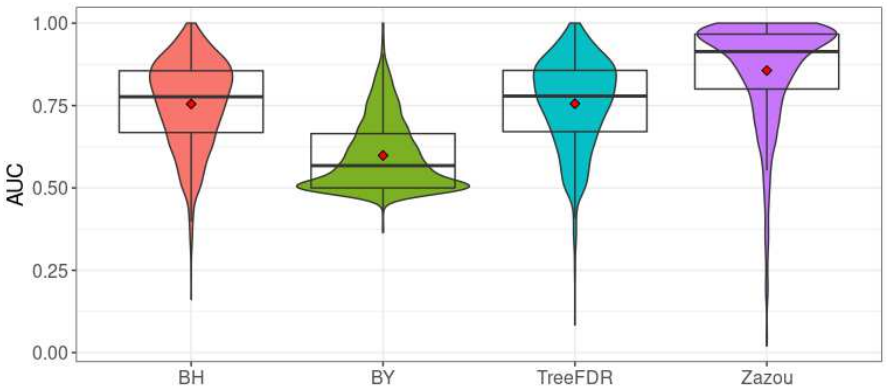


FIGURE 2 – Comparaison des AUC obtenus sans correction phylogénétique (Benjamini-Hochberg, Benjamini-Yekutieli) et avec (TreeFDR, Zazou).

Références

[1] Paul Bastide, Mahendra Mariadassou, and Stéphane Robin. Detection of adaptive shifts on phylogenies by using shifted stochastic processes on a tree. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 79(4) :1067–1093, 2017.

[2] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate : a practical and powerful approach to multiple testing. *Journal of the Royal statistical society : series B (Methodological)*, 57(1) :289–300, 1995.

[3] Yoav Benjamini, Daniel Yekutieli, et al. The control of the false discovery rate in multiple testing under dependency. *The annals of statistics*, 29(4) :1165–1188, 2001.

[4] Nicholas A Bokulich, Jennifer Chung, Thomas Battaglia, Nora Henderson, Melanie Jay, Huilin Li, Arnon D Lieber, Fen Wu, Guillermo I Perez-Perez, Yu Chen, et al. Antibiotics, birth mode, and diet shape microbiome maturation during early life. *Science translational medicine*, 8(343) :343ra82–343ra82, 2016.

[5] Benjamin J Callahan, Paul J McMurdie, Michael J Rosen, Andrew W Han, Amy Jo A Johnson, and Susan P Holmes. Dada2 : high-resolution sample inference from illumina amplicon data. *Nature methods*, 13(7) :581, 2016.

[6] J Gregory Caporaso, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D Bushman, Elizabeth K Costello, Noah Fierer, Antonio Gonzalez Pena, Julia K Goodrich, Jeffrey I Gordon, et al. Qiime allows analysis of high-throughput community sequencing data. *Nature methods*, 7(5) :335, 2010.

[7] Jun Chen, Emily King, Rebecca Deek, Zhi Wei, Yue Yu, Diane Grill, and Karla Ballman. An omnibus test for differential distribution analysis of microbiome sequencing data. *Bioinformatics*, 34(4) :643–651, 2018.

[8] Wenjiang J Fu. Penalized regressions : the bridge versus the lasso. *Journal of computational and graphical statistics*, 7(3) :397–416, 1998.

[9] Michael Love, Simon Anders, and Wolfgang Huber. Differential analysis of count data–the deseq2 package. *Genome Biol*, 15(550) :10–1186, 2014.

[10] Xochitl C Morgan, Timothy L Tickle, Harry Sokol, Dirk Gevers, Kathryn L Devaney, Doyle V Ward, Joshua A Reyes, Samir A Shah, Neal LeLeiko, Scott B Snapper, et al. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome biology*, 13(9) :R79, 2012.

[11] Joseph N Paulson, O Colin Stine, Héctor Corrada Bravo, and Mihai Pop. Differential abundance analysis for microbial marker-gene surveys. *Nature methods*, 10(12) :1200, 2013.

[12] Gautam V Pendse. A tutorial on the lasso and the shooting algorithm. *Harvard Medical School*, 13, 2011.

[13] Florian Plaza Oñate, Emmanuelle Le Chatelier, Mathieu Almeida, Alessandra CL Cervino, Franck Gauthier, Frédéric Magoulès, S Dusko Ehrlich, Matthieu Pichaud, and Jonathan Wren. Mspminer : abundance-based reconstitution of microbial pangenomes from shotgun metagenomic data. *Bioinformatics*, 2018.

[14] R Core Team. *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020.

[15] Jacques Ravel, Pawel Gajer, Zaid Abdo, G Maria Schneider, Sara SK Koenig, Stacey L McCulle, Shara Karlebach, Reshma Gorle, Jennifer Russell, Carol O Tacket, et al. Vaginal microbiome of reproductive-age women. *Proceedings of the National Academy of Sciences*, 108(Supplement 1) :4680–4687, 2011.

[16] Gil Sharon, Daniel Segal, John M Ringo, Abraham Hefetz, Ilana Zilber-Rosenberg, and Eugene Rosenberg. Commensal bacteria play a role in mating preference of drosophila melanogaster. *Proceedings of the National Academy of Sciences*, page 201009906, 2010.

[17] Peter J Turnbaugh, Micah Hamady, Tanya Yatsunenko, Brandi L Cantarel, Alexis Duncan, Ruth E Ley, Mitchell L Sogin, William J Jones, Bruce A Roe, Jason P Affourtit, et al. A core gut microbiome in obese and lean twins. *nature*, 457(7228) :480, 2009.

[18] Jian Xiao, Hongyuan Cao, and Jun Chen. False discovery rate control incorporating phylogenetic tree increases detection power in microbiome-wide multiple testing. *Bioinformatics*, 33(18) :2873–2881, 2017.

[19] Cun-Hui Zhang and Stephanie S Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 76(1) :217–242, 2014.

Sensor selection on graphs via data-driven node sub-sampling in network time series

Jérémie Bigot [1], Yiye Jiang [1,2] & Sofian Maabout [2]

[1] *Institut de Mathématiques de Bordeaux. jeremie.bigot@math.u-bordeaux.fr yiye.jiang@math.u-bordeaux.fr*
[2] *Laboratoire Bordelais de Recherche en Informatique. maabout@u-bordeaux.fr*

Résumé. Cet article aborde la sélection d'un ensemble d'échantillonnage optimal de capteurs sur un réseau de séries temporelles dans le but de récupérer le signal au niveau de capteurs non observés avec une erreur de reconstruction minimale. La stratégie proposée est motivée par des applications dans lesquelles des signaux définis sur un graphe et dépendant du temps sont collectés sur des réseaux redondants. Dans ce cadre, on peut souhaiter utiliser uniquement un sous-ensemble de capteurs pour prédire les flux de données sur l'ensemble de la collection de noeuds dans le graphe sous-jacent. Une application typique est la possibilité de réduire la consommation d'énergie dans un réseau de capteurs dont les batteries peuvent être limitées. Nous proposons et comparons différentes stratégies basées sur les données pour sélectionner un ensemble de noeuds d'échantillonnage ou de manière équivalente pour désactiver un nombre fixe de capteurs. Nous relions également notre approche à la littérature existante sur la sélection des capteurs à partir de données multivariées avec une structure graphique sous-jacente (éventuellement). Notre méthodologie combine des outils d'analyse de séries temporelles multivariées, de traitement du signal sur des graphes et d'apprentissage statistique en grande dimension. Pour illustrer les performances de notre approche, nous proposons une analyse de données réelles issues de réseaux de vélos en libre-service dans différentes villes.

Mots-clés. Sélection des capteurs, séries temporelles du réseau, traitement du signal sur des graphes, ensemble d'échantillonnage, statistiques de grande dimension, Laplacian et noyau de graphe, transformation de Fourier sur des graphes, réseau neuronal convolutif sur des graphes, réseaux de vélos en libre-service.

Abstract. This paper is concerned by the problem of selecting an optimal sampling set of sensors over a network of time series for the purpose of signal recovery at non-observed sensors with a minimal reconstruction error. The proposed strategy is motivated by applications where time-dependent graph signals are collected over redundant networks. In this setting, one may wish to only use a subset of sensors to predict data streams over the whole collection of nodes in the underlying graph. A typical application is the possibility to reduce the power consumption in a network of sensors that may have limited battery supplies. We propose and compare various data-driven strategies to select a sampling set of nodes or equivalently to turn off a fixed number of sensors. We also relate our approach to the existing literature on sensor selection from multivariate data with a (possibly) underlying graph structure. Our methodology combines tools from multivariate

time series analysis, graph signal processing and statistical learning in high-dimension. To illustrate the performances of our approach, we report numerical experiments on the analysis of real data from bike sharing networks in different cities.

Keywords. Sensor selection, network time series, signal processing on graphs, sampling set, high-dimension statistics, Laplacian and graph kernel, graph Fourier transform, graph convolutional neural networks, bike sharing networks.

1 Notations and methodologies of sensor selection

First of all, we represent a network of sensors as a graph $\mathcal{G} = \{\mathcal{N}, \mathcal{E}, A\}$ consisting of a finite set of nodes (or vertices) $\mathcal{N} = (V_i)_i$ with $|\mathcal{N}| = N$, a set of edges $\mathcal{E} = (i,j)_{i,j}$, and an adjacency matrix $A = (a_{ij})_{i,j}$.

A key hypothesis of our approach is to assume that observations of time-dependent signals $\mathbf{x}_t$ on the graph $\mathcal{G}$ are available for a sufficiently large number of time points $1 \leq t \leq T_0$. At each time t, a *graph signal* on $\mathcal{G}$ is defined as a mapping $x_t : \mathcal{N} \to \mathbb{R}$ with $x_t(V_i)$ representing the observation at time t and node V_i. The collection of data $(\mathbf{x}_t)_{1 \leq t \leq T_0}$ is a multivariate time series that we shall also refer to as a *network time series*. Equivalently the signal $\mathbf{x}_t$ may be represented as a vector $\mathbf{x}_t = (x_{1t}, x_{2t}, ..., x_{Nt}) \in \mathbb{R}^N$ with $x_{it} = x_t(V_i)$. Then, after time T_0, we wish to turn off a subset $\{V_i, i \in I\}$ of sensors with given cardinality $|I| = p > 0$, and to reconstruct as accurately as possible the signal over all nodes of the graph at time $T_0 < t \leq T_1$ using observations from the remaining sensors. We call the values of sensors until T_0 the historical data and those after T_0 the current data. We denote $\mathbf{x}_{I,t} = (x_{it})_{i \in I}$ and $\mathbf{x}_{I^c,t} = (x_{it})_{i \in I^c}$ where $I^c = \mathcal{N} \backslash I$.

Then, the first step in our procedure of sensor selection is to choose a *methodology of reconstruction* that is defined as a parametric class of functions

$$f_\Theta^{\mathcal{G}} : \mathbb{R}^q \to \mathbb{R}^p, \quad \text{where} \quad q = (N-p)(H+1),$$

indexed by a set of real parameters Θ with dimension d. The values of $f_\Theta^{\mathcal{G}}$ may depend on the graph $\mathcal{G}$ and data from the past up to time lag $H \geq 0$, with $H \ll T_0$. The parameter H represents the amount of past information to be used for signal recovery at unobserved nodes, and we denote by $\mathbf{x}_{I^c,t}^H = [\mathbf{x}_{I^c,t}, \mathbf{x}_{I^c,t-1}, ..., \mathbf{x}_{I^c,t-H}]$ the $\mathbb{R}^q$ vector containing the observed signals from time $t-H$ to t.

In a second step, given a subset I of fixed cardinality $|I| = p$, one trains $f_\Theta^{\mathcal{G}}$ by minimizing the reconstruction error on all historical data $\mathbf{x}_t, t \leq T_0$ as follows

$$\mathbf{F}(I, \hat{\Theta}(I)) = \min_{\Theta \in \mathbb{R}^d} \mathbf{F}(I, \Theta), \quad \text{where} \quad \mathbf{F}(I, \Theta) := \frac{1}{T_0} \sum_{t=H+1}^{T_0} \|\mathbf{x}_{I,t} - f_\Theta^{\mathcal{G}}(\mathbf{x}_{I^c,t}^H)\|_{\ell_2}^2, \qquad (1)$$

where $\|\cdot\|_{\ell_2}$ denotes the usual Euclidean norm. The third step is to select a subset $\hat{I}$ of cardinality p minimizing the reconstruction error of the historical data that is

$$\hat{I} = \underset{I\subset\mathcal{N}\,:\,|I|=p}{\arg\min}\ \mathbf{F}(I,\hat{\Theta}(I)). \tag{2}$$

In this paper, we have considered three classes of reconstruction methods, which are *linear estimators* (without graph prior), *graph kernel approaches*, and *graph convolutional neural networks* (with graph prior). For the first two classes, minimizing (2) is a delicate combinatorial problem as some of its simplest instance using a linear reconstruction approach is known to be NP-complete, for example see [3]. Hence, we shall introduce greedy strategies to approximate a solution to (2) which selects firstly the best node $i_{(1)}$ from $\mathcal{N}$, then to find the best node $i_{(2)}$ from the reduced network consisting of the node set $\mathcal{N}\setminus\{i_{(1)}\}$ along with the edges among them, and so on. The algorithm stops after the number of selected sensors reaches the pre-given hyperparameter p. Whereas for the deep learning class, we have designed the special architectures as well as training schemes to circumvent this problem. All of our methods provide an order of removing for selected sensor.

1.1 Linear estimators and minimization of partial variance

We now assume $(\mathbf{x}_t)_t$ is a multivariate stationary process with zero mean[1], and consider the reconstruction function $f^{\mathcal{G}}_{\Theta}(\mathbf{x}^H_{I^c,t}) := \Theta\mathbf{x}^H_{I^c,t}$. Then the theoretical reconstruction error writes as

$$\mathcal{F}(I,\Theta) = \mathbb{E}\|\mathbf{x}_{I,t} - \Theta\mathbf{x}^H_{I^c,t}\|^2_{l_2},$$

where $\Theta\in\mathbb{R}^{p\times q}$. From the population point of view, the best set I is given by

$$\begin{aligned} I^* &= \underset{I\subset\mathcal{N}\,:\,|I|=p}{\arg\min}\ \mathcal{F}(I,\hat{\Theta}(I)) = \underset{I\subset\mathcal{N}\,:\,|I|=p}{\arg\min}\ \min_{\Theta\in\mathbb{R}^{p\times q}}\mathcal{F}(I,\Theta) \\ &= \underset{I\subset\mathcal{N}\,:\,|I|=p}{\arg\min}\ tr(\mathbf{\Sigma}_I - [\beta^H_{II^c}][\alpha^H_{I^c}]^{-1}[\beta^H_{II^c}]^t) = \underset{I\subset\mathcal{N}\,:\,|I|=p}{\arg\min}\sum_{i\in I}\sigma_i^2 - [\beta^H_{iI^c}][\alpha^H_{I^c}]^{-1}[\beta^H_{iI^c}]^t, \end{aligned} \tag{3}$$

where $\mathbf{\Sigma}_I = \mathrm{Cov}(\mathbf{x}_{I,t})$, $\alpha^H_{I^c} = \mathrm{Cov}(\mathbf{x}^H_{I^c,t})$, and $\beta^H_{II^c} = \mathrm{Cov}(\mathbf{x}_{I,t}\,,\,\mathbf{x}^H_{I^c,t})$. Note that quantity $\sigma_i^2 - [\beta^H_{iI^c}][\alpha^H_{I^c}]^{-1}[\beta^H_{iI^c}]^t$ is the *partial variance* of variable x_{it} given $\mathbf{x}^H_{I^c,t}$. Thus the formula aims to find the set of most predictable variables, measured by their partial variances.

Furthermore, it can be proven that if we replace all the population covariances with their corresponding sample covariances in formula (3), we can obtain the minimal empirical reconstruction error so that the selection result $\hat{I}$ readily, which reads as

$$\hat{I} = \underset{I\subset\mathcal{N}\,:\,|I|=p}{\arg\min}\sum_{i\in I}\hat{\sigma}_i^2 - [\hat{\beta}^H_{iI^c}][\hat{\alpha}^H_{I^c}]^{-1}[\hat{\beta}^H_{iI^c}]^t. \tag{4}$$

[1]To facilitate the analysis, in section 1.1, 1.2, we assume the graph signal is centered, in the sense that $\bar{\mathbf{x}} = \frac{1}{T_0}\sum_{t=1}^{T_0}\mathbf{x}_t = \mathbf{0}$.

In practice, we derived the greedy algorithm according to criterion (4), meaning that each time we remove one sensor from the current graph which has the minimal partial variance. When $H = 0$, the resulting greedy algorithm coincides with the one in [4, 5], which is based on Gaussian process assumption and entropy measurement. The only difference is that, in our setting, we did not add the Gaussian assumption, allowing the algorithm applied over various population distributions, only if their second moment exists. While the trade-off is, for some distributions linear function is not the best reconstruction function.

1.2 Graph kernel approaches

We still assume the process $(\mathbf{x}_t)_t$ is stationary and zero mean. In the case where the graph prior is provided, we can consider graph kernel ridge regression as reconstruction method[2]. We propose to define kernel as $k : \mathcal{X} \times \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X} = \mathcal{N} \times \mathbb{Z}$ and $\mathcal{Y} = \mathbb{R}$. Because the process $(\mathbf{x}_t)_t$ is stationary, we require the kernel value to only depend on the time difference rather than specific time stamps. In other words, this means that $k\left[(i,t),(j,t-l)\right] = k\left[(i,t'),(j,t'-l)\right]$. Thus, we can denote the kernel value $k\left[(i,t),(j,t-l)\right]$ by $k(i,j,l)$. We first set up the following regression problem to acquire the reconstruction function $f^{\mathcal{G}}_{\Theta}$, by computing

$$\alpha^*_t = \underset{\alpha\in\mathbb{R}^q}{\arg\min} \|\mathbf{x}^H_{I^c,t} - K^H_{I^c}\alpha\|^2_{\ell^2} + \lambda\alpha^t K^H_{I^c}\alpha\,,$$

where $K^H_{I^c} \in \mathbb{R}^{q\times q}$ is the Gram matrix of input $I^c \times [t, t-1, ..., t-H]$. Then $f^{\mathcal{G}}_{\Theta}(\mathbf{x}^H_{I^c,t}) = K^H_{II^c}\alpha^*_t = K^H_{II^c}(K^H_{I^c} + \lambda\mathrm{Id})^{-1}\mathbf{x}^H_{I^c,t}$, where $K^H_{II^c} = [K(I, I^c, 0), K(I, I^c, 1), ..., K(I, I^c, H)] \in \mathbb{R}^{p\times q}$, with $K(I, I^c, l) \in \mathbb{R}^{p\times(N-p)}$ whose (i,j)th entity equal to $k(i,j,l)$ with $i \in I, j \in I^c$. Thus the minimal reconstruction error and the best turned-off set $\hat{I}$ is given by

$$\begin{aligned}\hat{I} = \underset{I\subset\mathcal{N}\,:\,|I|=p}{\arg\min}\ \mathbf{F}(I,\hat{\Theta}(I)) &= \underset{I\subset\mathcal{N}\,:\,|I|=p}{\arg\min}\ \frac{1}{T_0}\sum_{t=H+1}^{T_0} \|\mathbf{x}_{I,t} - K^H_{II^c}(K^H_{I^c} + \lambda\mathrm{Id})^{-1}\mathbf{x}^H_{I^c,t}\|^2_{\ell_2} \\ &= \underset{I\subset\mathcal{N}\,:\,|I|=p}{\arg\min}\ \sum_{i\in I} \hat{\sigma}^2_i - 2\hat{\beta}^H_{iI^c}\hat{\Theta}(i)^t + \hat{\Theta}(i)\hat{\alpha}^H_{I^c}\hat{\Theta}(i)^t\,,\end{aligned} \tag{5}$$

where $\hat{\Theta}(i)$ denotes the quantity $K^H_{iI^c}(K^H_{I^c} + \lambda\mathrm{Id})^{-1}$. In practice, we derive the greedy algorithm from criterion (5) to select sensors.

Note that, this section is the nonlinear generalization of previous section. Indeed, if we take the specific kernel as sample autocovariance, then $K^H_{I^c} = \hat{\alpha}^H_{I^c}$, $K^H_{II^c} = \hat{\beta}^H_{II^c}$, and the kernel ridge regression becomes linear ridge regression. Furthermore, if λ is set to be 0, then criterion (5) becomes (4). In our application, we use the following kernel design.

$$k\left[\,(i,t),(j,t')\,\right] = k_{\mathcal{G}}(i,j)\cdot k_{rbf}(t,t'),$$

where $k_{\mathcal{G}}$ is the graph Laplacian kernel, and k_{rbf} is the Gaussian kernel.

[2]For the background of reproducing kernel Hilbert space and its induced graph kernel regression on independent and identically distributed data, see [2].

1.3 Graph convolutional neural networks (GCN)

Our basic network architecture is from [1]. On the top of it, as a reconstruction method of missing set I, we first complete the graph signals $\mathbf{x}^H_{I^c,t}$ to $\mathbf{x}^H_t$ by inserting zeros on the missing places $\mathbf{x}^H_{I,t}$, then input the whole signal to GCN. We require GCN to output the recovered missing values $\hat{\mathbf{x}}_{I,t}$ by setting the loss function as

$$J = \frac{1}{|Batch|} \sum_{t \in \text{Batch}} \|\mathbf{x}_{I,t} - \hat{\mathbf{x}}_{I,t}\|^2_{\ell_2}.$$

We then study two ways of transforming a prediction network into a selection network which adapts to its own prediction capability. We refer to the transformed selection networks as masked GCN and dropout GCN respectively. For both selection networks, we set their input and output dimensions as N, without changing the prediction architecture in between. Masked GCN uses ℓ_1 norm regularization in the sensor selection, where the number of selected sensors is controlled by the regularization parameter. Dropout GCN leverages the dropout technique [6], to block the inputs from p random sensors at each optimization step, i.e. the dropout rate is set as p/N, meanwhile only collects their outputs to calculate the prediction error. This method finally provides a scoring for each sensor which takes into account all these missing situations the prediction network has seen. The scoring furthermore gives us the selection result.

2 Sensor selection on a city bike-sharing network

Our data comes from the bike-sharing network in Toulouse city, which consists of 273 nodes and 4300 hours[3]. Every bike station has been installed a sensor to record the number of empty docks. Time series variable x_{it} represents the ratio of empty docks (used bikes) at sensor i, time t. In this case, the graph is constructed based on the geographical distance between stations. We applied the above-mentioned methodologies of sensor selection over this network. Figure 1, 2 show their selection results. Figure 3 shows some methods' reconstruction performance of their first selected sensors.

These results reveal that, there are two types of signals in the network, which are both sensible to be selected. One is from the remote sensor (e.g. top signal in Figure 3). It is noise of small magnitude, therefore naturally leading to lower reconstruction error. The other comes from the city center (e.g. bottom signal in Figure 3), which contains meaningful patterns. We use scaled data over the kernel approach in order to guide the model to detect and learn the patterns, at the cost of not selecting the sensors of absolute minimal reconstruction error, which is also the case for dropout GCN with R2 score.

[3]We gratefully acknowledge Max Halford for providing this dataset available at `https://maxhalford.github.io/blog/a-short-introduction-and-conclusion-to-the-openbikes-2016-challenge/`

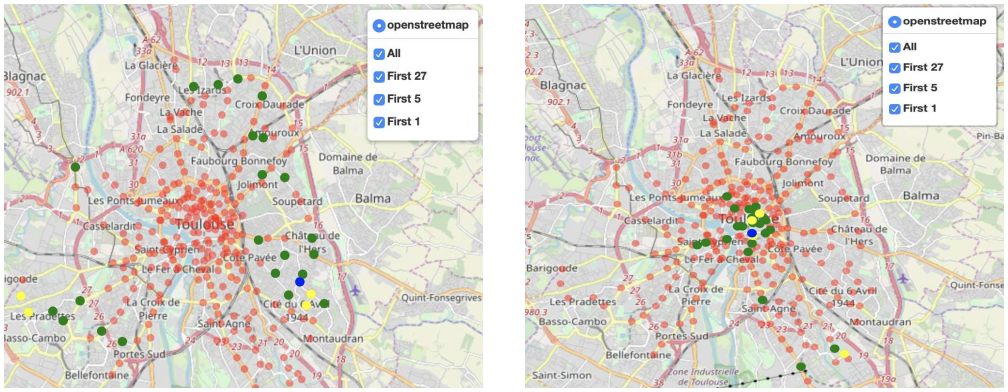


Figure 1: *Sensor selection results* ($p = 0.1N$). Left: linear ridge regression ($H = 15$, $\mathbf{F}(\hat{I}, \hat{\Theta}(\hat{I}))$ on test set: 1.5717). Right: graph kernel ridge regression ($H = 1$, scaled data, $\mathbf{F}(\hat{I}, \hat{\Theta}(\hat{I}))$ on test set: 1.4015).

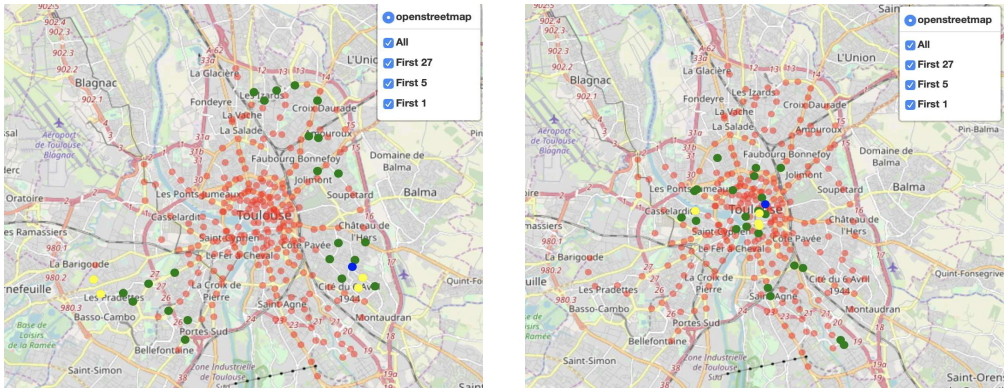


Figure 2: *Sensor selection results* ($p = 0.1N$). Left: masked GCN ($H = 0$, $\mathbf{F}(\hat{I}, \hat{\Theta}(\hat{I}))$ on test set: 1.1768). Right: dropout GCN ($H = 0$, R2 score, $\mathbf{F}(\hat{I}, \hat{\Theta}(\hat{I}))$ on test set: 1.3995).

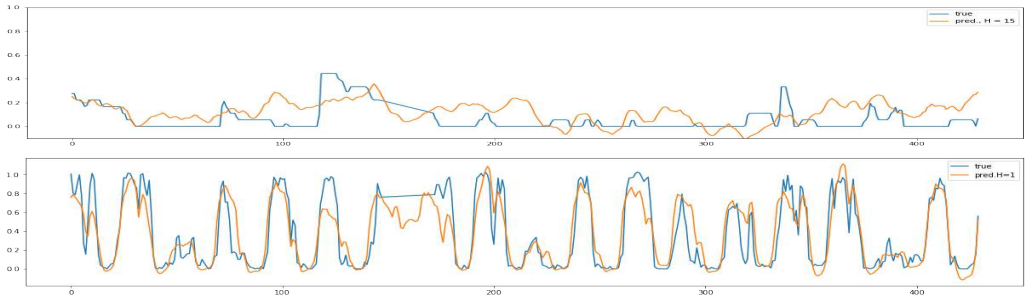


Figure 3: *The reconstruction of* 1*st removed sensor.* Top: linear ridge regression ($H = 15$). Bottom: graph kernel ridge regression ($H = 1$, scaled data).

References

[1] Defferrard, M., Bresson, X., and Vandergheynst, P. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in neural information processing systems* (2016), pp. 3844–3852.

[2] Kolaczyk, E. D. *Statistical Analysis of Network Data: Methods and Models*, 1st ed. Springer Publishing Company, Incorporated, 2009.

[3] Krause, A., Singh, A., and Guestrin, C. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *J. Mach. Learn. Res. 9* (June 2008), 235–284.

[4] Sakiyama, A., Tanaka, Y., Tanaka, T., and Ortega, A. Eigendecomposition-free sampling set selection for graph signals. *IEEE Transactions on Signal Processing 67*, 10 (2019), 2679–2692.

[5] Shewry, M. C., and Wynn, H. P. Maximum entropy sampling. *Journal of Applied Statistics 14*, 2 (1987), 165–170.

[6] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research 15*, 1 (2014), 1929–1958.

Etude de co-localisation en génomique avec des processus de Hawkes

Anna Bonnet [1] & François Gindraud [2] & Franck Picard[3] & Vincent Rivoirard[4]

[1] *Sorbonne Université - LPSM, 4, Place Jussieu 75005 Paris anna.bonnet@upmc.fr*
[2]*Laboratoire de Biométrie et Biologie Evolutive UMR 5558 CNRS Univ. Lyon 1 - Bât. Grégor Mendel 43 bd du 11 novembre 1918 69622 Villeurbanne francois.gindraud@univ-lyon1.fr*
[3]*Laboratoire de Biométrie et Biologie Evolutive UMR 5558 CNRS Univ. Lyon 1 - Bât. Grégor Mendel 43 bd du 11 novembre 1918 69622 Villeurbanne franck.picard@univ-lyon1.fr*
[4] *CEREMADE UMR CNRS 7534 Groupe Probabilités et Statistique Université Paris Dauphine Place du Maréchal De Lattre De Tassigny 75775 PARIS Cedex 16 vincent.rivoirard@dauphine.fr*

Résumé. Les protocoles de séquençage ChIP-Seq, très utilisés en génomique, permettent de localiser des interactions entre ADN et protéines à l'échelle du génome. Nous cherchons à déterminer les informations biologiques pertinentes parmi ces données spatiales, en étudiant notamment la co-localisation de plusieurs processus le long du génome. Une des particularités de ces données est qu'elles sont bruitées à cause des incertitudes dans les méthodes de détection. Nous proposons d'utiliser un modèle de processus ponctuel multivarié, le modèle de Hawkes, et nous proposons une extension des méthodes d'estimation non paramétrique de Reynaud-Bouret et al (2014) dans le cas où les données sont observées avec un bruit uniforme. Nous appliquons cette méthode à plusieurs jeux de données issus de la génomique et nous mettons en évidence la nécessité d'utiliser une méthode d'analyse multivariée ainsi que l'intérêt de prendre en compte l'incertitude sur la position des données.

Mots-clés. Processus de Hawkes multivarié, Convolution, Applications en génomique

Abstract. The spatial localization of many DNA-protein interactions is now available thanks to the development of ChIP-Seq protocoles, and their investigation calls for adapted statistical methods. Our goal is to study co-localization of biological features spatialized along the genome in order to extract relevant biological information. One specificity of the data is that it is are detected with uncertainties regarding the precise position. We propose to use a multivariate Hawkes model with an extension of the inference procedure developed by Reynaud-Bouret et al (2014) when the positions are directly observed. We apply this method to several genomic datasets.

Keywords. Multivariate Hawkes process, Convolution, Genomic applications

1 Introduction

1.1 Contexte biologique et technologique

Les nouvelles technologies de séquençage permettent désormais d'étudier le fonctionnement des génomes à une résolution jamais atteinte. La disponibilité des génomes complets a permis la localisation d'éléments régulateurs du génome comme les marques chromatiniennes qui caractérisent l'état de compaction de l'ADN et les enhancers qui contrôlent l'expression des gènes. La prise en compte de cette information spatiale s'avère fondamentale, notamment pour étudier des phénomènes intrinsèquement spatiaux comme la réplication du génome. Ces données sont généralement disponibles sous forme d'intervalles génomiques, ce qui est dû à certaines imprécisions dans la détection des observations. Un des principaux défis réside désormais dans le traitement de ces données en masse, afin d'en extraire les connaissances permettant de mieux comprendre le fonctionnement global de la régulation des génomes. L'objectif est de développer un cadre statistique permettant de modéliser et d'estimer les interactions spatiales entre éléments localisés sur le génome.

1.2 Etat de l'art

De nombreuses méthodes ont été développées pour étudier les co-occurrences de plusieurs éléments génomiques and testant notamment des associations deux à deux. Ces méthodes très générales fournissent des schémas de co-occurrences pour toutes les données spatiales décrites par des points ou des intervalles, en particulier les données génomiques. Favorov et al. (2012) proposent par exemple quatre procédures standard implémentées dans le package R `GenometriCorr` pour évaluer la proximité spatiale de deux listes de points ou d'intervalles. D'autres méthodes ont été également proposées pour tenir compte notamment de l'hétérogénéité du génome (par exemple la méthode ChromHMM proposée par Wai et al. (2016)) ou encore Chikina et al. (2012) qui proposent une nouvelle métrique pour évaluer la proximité de listes d'intervalles.
La limite de ces différentes méthodes est qu'elles sont adaptées à des comparaisons deux à deux et ne permettent pas de détecter les corrélations artéfactuelles dues à une corrélation commune avec un autre processus. Carstensen et al. (2010) ont proposé de modéliser les interactions spatiales grâce à des processus de Hawkes, qui permettent de caractériser précisément les interactions spatiales (phénomènes d'attraction/répulsion, intensité, distances typiques d'interactions) et également de proposer un cadre unifié permettant d'étudier de nombreux processus simultanément. Toutefois, ce modèle présente l'inconvénient de revenir à une modélisation ponctuelle et de ne pas tenir compte de l'incertitude sur la positions des données.

Nous proposons ici de combiner la modélisation avec des processus de Hawkes utilisée par Carstensen et al. (2010) en intégrant dans notre modèle l'imprécision des positions

observées. Nous développons ensuite une méthode d'inférence en nous appuyant sur les travaux de Reynaud-Bouret et al. (2014), qui ont proposé une méthode d'estimation non paramétrique des fonctions d'interaction du modèle de Hawkes dans le cas où l'on observe les positions exactes.

2 Modèle et inférence statistique

Nous présentons dans un premier temps le modèle de Hawkes multivarié où l'on observe précisément les positions des occurrences et ensuite le modèle de Hawkes bruité.

2.1 Modèle de Hawkes multivarié

On observe un ensemble de positions que l'on note $N^m = \{X_1^m, \dots, X_{n_m}^m\}$, de sorte que X_i^m est la position du i-ème pic du processus m. Dans la suite, $N = (N^1, \dots, N^M)$ correspond à l'ensemble des positions des M processus. On suppose que ces positions sont aléatoires et qu'elles peuvent être modélisées par un processus ponctuel. On introduit la représentation infinitésimale telle que l'événement $\{dN_x^m = 1\}$ représente l'augmentation infinitésimale du nombre de points de N^m à la position x. Chaque processus m est caractérisé par la fonction d'intensité λ^m qui modélise la probabilité d'observer une nouvelle occurrence de N^m conditionnellement à toutes les occurrences précédemment observées. En notant $\mathcal{F}_{x^-}^1, \dots, \mathcal{F}_{x^-}^M$ les occurrences de N observées avant x, on considère :

$$\mathbb{P}\left(dN_x^m = 1 \mid \mathcal{F}_{x^-}^1, \dots, \mathcal{F}_{x^-}^M\right) = \lambda^m(x).$$

La modèle de Hawkes consiste à écrire l'intensité conditionnelle sous la forme

$$\forall m \in [1, M],\ \lambda^m(x) \ = \ \nu^m + \sum_{\ell=1}^{M} \sum_{X \in N^\ell} h_\ell^m(x - X),$$

de sorte que la probabilité d'observer une nouvelle occurrence de N^m dépende de toutes les occurrences de N observées avant x au travers de fonctions d'interaction $h^m = (h_1^m, \dots, h_M^m)$. Plus précisément, la fonction (inconnue) h_ℓ^m caractérise l'influence des points de N^ℓ sur la probabilité d'observer des points de N^m et cette influence ne dépend que de la distance entre les occurrences des processus ℓ et m. Un aspect essentiel de la modélisation avec des processus de Hawkes repose dans la définition de l'intensité conditionnelle, qui garantit de s'affranchir des interactions artéfactuelles puisque les interactions entre les processus N^m et N^ℓ sont corrigées par toutes les autres interactions entre les processus de N. Remarquons que h_m^m décrit l' auto-interaction du processus m, ce qui permet de modéliser la dépendance entre les occurrences d'un même processus. Enfin, ν^m est appelé taux spontané et peut être considéré comme la part de l'intensité qui ne peut être expliquée par la présence des occurrences de $(N^1, \dots, N^M)$.

2.2 Modélisation de l'incertitude sur les positions des occurrences

Les vraies positions des occurrences sont en fait inconnues à cause des imprécisions dues aux méthodes de détection. En effet, les observations sont fournies sous forme d'intervalles sur lesquels se trouvent les occurrences sans que l'on connaisse leur position exacte. On considère alors un processus observé $\{X_1^m, \dots, X_{n_m}^m\}$ qui est une version bruitée des vraies positions $\{X_1^{*m}, \dots, X_{n_m}^{*m}\}$, de sorte que $\forall m \in \{1, \dots, M\}$, $\forall i \in \{1, \dots, n_m\}$

$$X_i^{*m} = X_i^m + E_i^m,$$

où $E_i^m \sim \mathcal{U}[0, \eta^m]$ est indépendant de X_i^m. En pratique, on observe un intervalle de longueur η^m : on considérera que l'observation est donnée par le début de l'intervalle X_i^m et que l'occurrence appartient en réalité à $[X_i^m, X_i^m + \eta^m]$. La longueur maximale de l'erreur de position η_i^m est connue et dépend de chaque occurrence. On note

$$W^m(x) = \frac{1}{\eta^m} 1_{x \in [0, \eta^m]}$$

la fonction de densité de la distribution uniforme sur $[0, \eta^m]$. En notant $\mathcal{E}^m$ l'ensemble des erreurs de position sur N^{*m}, l'intensité conditionnelle de N^{*m} conditionnellement aux positions observées et aux erreurs s'écrit :

$$\mathbb{P}\left(dN_x^{*m} = 1 \mid \mathcal{F}_{x^-}^1, \dots, \mathcal{F}_{x^-}^M, \mathcal{E}^1, \dots, \mathcal{E}^M\right) = \nu^m + \sum_{\ell=1}^{M} \sum_{i=1}^{n_\ell} h_\ell^m\big(x - [X_i^\ell + E_i^\ell]\big),$$

Finalement, on peut exprimer la probabilité d'observer une occurrence de N^{*m} au point x en fonction des observations $N^1, \dots, N^M$ et des noyau uniformes $W^1, \dots, W^M$:

$$\begin{aligned}
\mathbb{P}\left(dN_x^{*m} = 1 \mid \mathcal{F}_{x^-}^1, \dots, \mathcal{F}_{x^-}^M\right) &= \int \cdots \int \prod_{\ell=1}^{M} \prod_{i=1}^{n_\ell} \nu^m W^\ell(e_i^\ell) de_i^\ell \\
&+ \int \cdots \int \prod_{\ell=1}^{M} \prod_{i=1}^{n_\ell} \left(\sum_{\ell=1}^{M} \sum_{i=1}^{n_\ell} h_\ell^m\big(x - [X_i^\ell + e_i^\ell]\big) \right) W^\ell(e_i^\ell) de_i^\ell \\
\lambda^{*m}(x) &= \nu^m + \sum_{\ell=1}^{M} \sum_{X \in N^\ell} W^\ell * h_\ell^m(x - X),
\end{aligned}$$

où $*$ correspond au produit de convolution.

2.3 Inférence non paramétrique

En s'inspirant de la méthode d'estimation non paramétrique proposée par Reynaud-Bouret et al. (2014) dans le cas d'un modèle de Hawkes où l'on observe les positions, nous proposons une méthode d'estimation des fonctions d'interaction $(h_\ell^m)_{1\leq \ell,m\leq M}$ dans le cas du modèle bruité.

3 Application

3.1 Application 1 : Etude du processus de réplication

Le programme de réplication du génome humain est un programme spatio-temporel complexe qui permet une réplication fidèle du matériel génétique. Picard et al. (2014) ont proposé des méthodes d'identification des points de départ du processus de réplication, appelées origines de réplication ; néanmoins, les mécanismes associés à l'initiation de la réplication sont encore peu connus. Plusieurs études ont proposé d'étudier la corrélation entre la présence des origines de réplications et différents facteurs génétiques et épigénétiques, en particulier les marques de chromatine qui caractérisent l'état de compaction de l'ADN. Des co-localisations ont été détectées grâce à des études d'association standard, et on s'intéresse à la caractérisation précise de ces interactions spatiales. Nous montrons notamment que certaines marques de chromatine augmentent considérablement la présence d'origines de réplication.

3.2 Application 2: Etude du phénomène de recombinaison

La recombinaison de l'ADN correspond à l'échange de matériel génétique entre plusieurs chromosome ou entre plusieurs régions d'un même chromosome. La conséquence de cette recombinaison est la production de descendants avec des gènes hérités de tous ses grand-parents, ce qui favorise la diversité génétique. Bien que courante, la recombinaison génétique est un processus complexe. En particulier, certaines régions sont appelées des points chauds de recombinaison car elles contiennent un taux de recombinaison particulièrement élevé. Nous proposons d'étudier le contexte génétique des points chauds de recombinaison.

Bibliographie

Carstensen, L., Sandelin A., Winther O. and Hansen, N.R. (2010), Multivariate hawkes process models of the occurrence of regulatory elements, *BMC Bioinformatics*,11(1):456.

Chikina, M.D. and Troyanskaya O.G. (2012) An effective statistical evaluation of chipseq dataset similarity, *Bioinformatics*, 28(5):607–613.

Favorov A.V., Mularoni L., Cope L.M. Medvedeva Y.A., Mironov A.A. Makeev V.J., Wheelan S.J. (2012),Exploring massive, genome scale datasets with the genometricorr package, *PLoS Computational Biology*, 8(5)

Picard F., Cadoret J-C., Audit B., Arneodo A., Alberti A.; Battail C., Duret L. and Prioleau M-N. (2014) The spatiotem- poral program of dna replication is associated with specific combinations of chromatin marks in human cells, *PLoS Genetics*, 10(5):e1004282, 2014.

Reynaud-Bouret P., Rivoirard V., Grammont F. and Tuleau-Malot C. (2014), Goodness-of-fit tests and nonparametric adaptive estimation for spike train analysis, *Journal of Mathematical Neuroscience*, page 4:3

Wei Y. and Wu H. (2016) Measuring the spatial correlations of protein binding sites. *Bioinformatics*,32(12):1766–1772

Vitesse de convergence dans les théorèmes centraux limite pour des statistiques résumées de processus ponctuels spatiaux

Christophe A.N. Biscio [1] & Florent Bonneu [2]

[1] *Department of Mathematical Sciences, Aalborg University, Skjernvej 4A, DK-9220 Aalborg, Denmark.*
christophe@math.aau.dk
[2] *LMA EA2151, Avignon Université, F-84000 Avignon, France.*
florent.bonneu@univ-avignon.fr

Résumé. En théorie des processus ponctuels spatiaux, des théorèmes centraux limite ont été établis ces dernières années dans des contextes différents (modèles, hypothèses, cadre asymptotique...) pour des statistiques résumées spécifiques, alors que leurs vitesses de convergence n'ont jamais été étudiées. La vitesse de convergence est notamment utile pour étudier les propriétés asymptotiques des estimateurs sur le plan théorique et pour contrôler l'erreur d'approximation faite en utilisant ces théorèmes en pratique. Pour combler ce manque, nous étudions la vitesse de convergence dans un théorème central limite de statistiques résumées pour des processus ponctuels spatiaux alpha-mélangeants. Nous démontrons tout d'abord une inégalité de type Berry-Esseen pour des champs aléatoires alpha-mélangeants par une méthode de Stein-Tikhomirov revisitée et utilisons ce résultat pour obtenir la vitesse de convergence dans le cadre de processus ponctuels spatiaux.

Mots-clés. Processus ponctuels spatiaux, statistique asymptotique, α-mélange, statistiques résumées, champs aléatoires.

Abstract. In spatial point process theory, central limit theorems were established in the past years under several contexts (models, assumptions, asymptotic framework...) on specific summary statistics, whereas their rates of convergence remained unstudied. The rate of convergence is particularly useful for studying asymptotic properties of estimators in theory and for controlling the approximation error made by using these theorems in practice. To fill this gap, we study the rate of convergence in a central limit theorem for summary statistics of alpha-mixing spatial point processes. We firstly prove a Berry-Esseen inequality for alpha-mixing random fields by a revisited Stein-Tikhomirov's method and use this result to obtain the rate of convergence in the spatial point process framework.

Keywords. Spatial point processes, asymptotic statistics, α-mixing, summary statistics, random fields.

1 Introduction

Soit $\mathbf{X}$ un processus ponctuel spatial sur $\mathbb{R}^d$ observé dans $W \subset \mathbb{R}^d$. En statistique des processus ponctuels spatiaux on s'intéresse souvent à des statistiques résumées $T_W(\mathbf{X})$ de la forme

$$T_W(\mathbf{X}) = \sum_{\xi_1,\ldots,\xi_p \in \mathbf{X} \cap W}^{\neq} h(\xi_1, \ldots, \xi_p) \tag{1}$$

où $h : \mathbb{R}^{dp} \to \mathbb{R}^q$, $p, q \geq 1$, et le signe $\neq$ signifie que la sommation s'effectue sur des paires de points distincts. Des exemples de statistiques résumées de cette forme incluent l'estimateur de la fonction K de Ripley et des équations d'estimation. Le comportement asymptotique de ces statistiques résumées quand l'aire de W augmente a été étudié ces dernières années sous diverses propriétés de $\mathbf{X}$ et hypothèses sur h.

Quand $\mathbf{X}$ verifie certaines propriétés d'alpha-mélange, divers résultats ont été établis dans des articles comme Guan et Sherman (2007), Prokešová et Jensen (2013), et Xu et al. (2018) en utilisant une approche basée sur la technique par bloc de Bernstein. Une autre approche pour prouver les théorèmes centraux limites sous des conditions d'alpha-mélange est basée sur Bolthausen (1982) qui considère des champs aléatoires stationnaires et dont les preuves sont généralisées dans Guyon (1995) et Karácsony (2006) pour des champs aléatoires non-stationnaires. Cette dernière approche a été utilisée dans plusieurs articles comme Waagepetersen et Guan (2009), Biscio et Coeurjolly (2016), et Coeurjolly (2017) avant d'être généralisée dans Biscio et Waagepetersen (2019).

Cependant, dans toutes les références sur les processus ponctuels mentionnées ci-dessus la vitesse de convergence des théorèmes centraux limite reste non étudiée bien qu'elle fasse l'objet de nombreuses recherches dans d'autres domaines comme par exemple en statistique non-paramétrique ou en statistique spatiale pour des champs aléatoires (Prokhorov et Statulevičius, 2000). En particulier, les vitesses de convergence pour des théorèmes centraux limites de statistiques résumées pour des champs aléatoires ont été étudiées dans Guyon et Richardson (1984), Guyon (1995) et Sunklodas (1986). Leurs techniques de preuve sont basées sur la méthode de Stein-Tikhomirov.

2 Vitesse de convergence

Nous considérons une suite croissante de fenêtres d'observation compactes $\{W_n\}_{n\in\mathbb{N}}$ vérifiant $W_1 \subset W_2 \subset \ldots$ and $|\bigcup_{l=1}^{\infty} W_l| = \infty$, ainsi que les hypercubes $C_n(\mathbf{l})$ centrés sur $\mathbf{l} \in \mathbb{L}$ un lattice (par la suite $\mathbb{L} := s_n\mathbb{Z}^d$) qui forment une partition disjointe de $\mathbb{R}^d$.

La statistique $T_{W_n}(\mathbf{X})$ définie par la forme (1) peut être décomposée comme

$$T_{W_n}(\mathbf{X}) = \sum_{\mathbf{l}\in\mathcal{D}_n} f_{n,\mathbf{l},W_n}(\mathbf{X}) \tag{2}$$

où pour $n \in \mathbb{N}$, $\mathcal{D}_n = \{\mathbf{l} \in \mathbb{L} : C_n(\mathbf{l}) \cap W_n \neq \emptyset\}$, et $f_{n,\mathbf{l},W_n}$ est la fonction définie par

$$f_{n,\mathbf{l},W_n}(\mathbf{X}) = \sum_{\xi_1 \in X \cap C_n(l) \cap W_n} \left(\sum_{\xi_2,\ldots,\xi_p \in (\mathbf{X} \cap W_n) \setminus \{\xi_1\}}^{\neq} h(\xi_1, \ldots, \xi_p) \right) \tag{3}$$

Pour $n \in \mathbb{N}$ et $\mathbf{l} \in \mathcal{D}_n$, $f_{n,\mathbf{l},W_n}(\mathbf{X})$ mesure la contribution locale à la statistique pour chaque $C_n(\mathbf{l})$. Ainsi, nous définissons le champ aléatoire centré $\{Z_n(\mathbf{l})\}_{\mathbf{l} \in \mathcal{D}_n}$ défini par $Z_n(\mathbf{l}) = f_{n,\mathbf{l},W_n}(\mathbf{X}) - \mathbb{E}(f_{n,\mathbf{l},W_n}(\mathbf{X}))$. On pose $S_n = \frac{1}{\sigma_n}(T_{W_n}(\mathbf{X}) - \mathbb{E}(T_{W_n}(\mathbf{X}))) = \sum_{\mathbf{l} \in \mathcal{D}_n} \frac{Z_n(\mathbf{l})}{\sigma_n}$, où σ_n^2 correspond à $\mathrm{Var}(T_{W_n}(\mathbf{X}))$; et nous nous intéresserons par la suite à son comportement asymptotique.

Afin d'étudier la vitesse de convergence de S_n nous démontrons le théorème suivant proposé pour tout champ aléatoire centré α-mélangeant $\{Z_\mathbf{j}\}_{\mathbf{j} \in \mathbb{Z}^d}$ vérifiant certaines hypothèses et observé sur un ensemble de fenêtres d'observation croissante $W_n \in \mathbb{R}^d$ et dont le coefficient de mélange est défini par

$$\alpha^Z_{c_1,c_2}(m) = \sup \{\alpha\left(\sigma((Z(\mathbf{l}) : \mathbf{l} \in I_1)), \sigma((Z(\mathbf{k}) : \mathbf{k} \in I_2))\right) : \\ I_1 \subset \mathbb{L}, I_2 \subset \mathbb{L}, |I_1| \leq c_1, |I_2| \leq c_2, \mathrm{d}(I_1, I_2) \geq m\}$$

Théorème 1. *Sous différentes hypothèses que nous développerons, nous obtenons les résultats suivants :*

1. *Dans le cas exponentiel*
   S'il existe $\lambda > 0$ tel que $\sup_{n \in \mathbb{N}} \alpha^Z_{\infty,\infty}(m) = O(e^{-\lambda m})$ alors

$$\sup_{t \in \mathbb{R}} \left| \mathbb{P}\Big(\sum_{\mathbf{j} \in W_n} \frac{Z_\mathbf{j}}{\sigma_n} \leq t \Big) - \Phi(t) \right| \leq C \frac{\ln(|W_n|)^{d \min(1+\tau,2)}}{|W_n|^{\frac{\min(1,\tau)}{2}}}.$$

2. *Dans le cas polynomial*
   Si pour $\mu > \frac{8d}{\min(1,\tau)}$, $\sup_{n \in \mathbb{N}} \alpha^Z_{\infty,\infty}(s) = O(m^{-\mu})$ alors

$$\sup_{t \in \mathbb{R}} \left| \mathbb{P}\Big(\sum_{\mathbf{j} \in W_n} \frac{Z_\mathbf{j}}{\sigma_n} \leq t \Big) - \Phi(t) \right| \leq C |W_n|^{-\frac{2+\min(1,\tau)}{2} \frac{\min(1+\tau,2)\mu}{2+\min(1+\tau,2)\mu}}$$

où nous connaissons les constantes C explicitement par rapport à τ, α, d, and $\sup_{\mathbf{l} \in \mathbb{L}} |Z_\mathbf{l}|_{2+\tau}$.

Nous présentons ensuite les travaux sur la vitesse de convergence du théorème central limite dans le cadre des processus ponctuels spatiaux.

Bibliographie

Biscio, C.A.N. et Waagepetersen, R. (2019). A general central limit theorem and subsampling variance estimator for α-mixing point processes, *Scandinavian Journal of Statistics*, 46, 1168–1190.

Biscio, C.A.N. et Coeurjolly, J.F. (2016). Standard and robust intensity parameter estimation for stationary determinantal point processes. *Spatial Statistics*, 18, 24-39.

Bolthausen, E. (1982). On the central limit theorem for stationary mixing random fields. *The Annals of Probability*, 10(4), 1047–1050.

Coeurjolly, J.F. (2017). Median-based estimation of the intensity of a spatial point process. *Annals of the Institute of Statistical Mathematics*, 69, 303-331.

Guan, Y. et Sherman, M. (2007). On least squares fitting for stationary spatial point processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(1), 31–49.

Guyon, X. et Richardson, S. (1984). Vitesse de convergence du théorème de la limite centrale pour des champs faiblement dépendants. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 66(2), 297–314.

Guyon, X. (1985). *Random fields on a network: modeling, statistics, and applications*. New York: Springer-Verlag.

Karacsony, Z. (2006). A central limit theorem for mixing random fields. *Miskolc Mathematical Notes. A Publication of the University of Miskolc*, 7(2), 147-160.

Prokhorov, Yu. V. et Statulevičius, V. (2000). *Limit theorems of Probability Theory*. Springer-Verlag Berlin Heidelberg.

Prokesova, M. et Jensen, E. (2013). Asymptotic Palm likelihood theory for stationary point processes. *Annals of the Institute of Statistical Mathematics*, 65, 387–412.

Sunklodas J. (1986). Estimate the rate of convergence in the central limit theorem for weakly dependent random fields. *Lithuanian Mathematical Journal*, 26, 272–287.

Waagepetersen, R. et Yongtao, G. (2009). Two-step estimation for inhomogeneous spatial point processes. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 71(3), 685–702.

Xu, G., Waagepetersen, R. et Guan, Y. (2018). Stochastic Quasi-likelihood for Case-Control Point Pattern Data. *Journal of the American Statistical Association*.

Regression modelling of interval censored data based on the adaptive ridge procedure

Olivier Bouaziz [1] & Eva Lauridsen [2] & Grégory Nuel [3]

[1] *Laboratoire MAP5 (UMR 8145), Université de Paris, France - olivier.bouaziz@u-paris.fr*
[2] *Ressource Center for Rare Oral Diseases, Copenhagen University Hospital, Rigshospitalet, Denmark*
[3] *LPSM (CNRS 8001), Sorbonne Université, France*

Résumé. Dans un contexte de données censurées à gauche, par intervalle et à droite, nous introduisons un modèle de Cox avec risque de base constant par morceaux. L'estimation se fait à partir de l'algorithme EM en considérant les vrais temps comme des données non observées. Cette procédure d'estimation produit une matrice Hessienne diagonale sur le bloc des paramètres correspondants au risque de base. Grâce à cette propriété, nous utilisons une méthode de vraisemblance pénalisée L_0 pour déterminer automatiquement le nombre et les localisations des ruptures du risque de base. La méthode s'étend directement à l'inclusion de données exactes et au modèle de guérison. Des résultats théoriques sont obtenus qui nous permettent de construire des intervalles de confiance et des tests sur les paramètres du modèle. La méthode proposée a été utilisée pour analyser des données dentaires sur le risque d'ankylose de dents qui ont été replantées et, à l'aide de données simulées, nous montrons que la méthode produit de bon résultats en pratique.

Mots-clés. Adaptive ridge, algorithme EM, censure par intervalle, modèle de guérison, risque instantané constant par morceaux, vraisemblance pénalisée.

Abstract. In a context of left-censored, interval-censored and right-censored data, a Cox model with piecewise constant baseline hazard is introduced. Estimation is carried out with the EM algorithm by treating the true event times as unobserved variables. This estimation procedure is shown to produce a block diagonal Hessian matrix of the baseline parameters. Taking advantage of this interesting feature of the estimation method a L_0 penalised likelihood method is implemented in order to automatically determine the number and locations of the cuts of the baseline hazard. The method can be directly extended to the inclusion of exact observations and to a cure fraction. Theoretical results are obtained which allow to derive statistical inference of the model parameters from asymptotic likelihood theory. A dental dataset on replanted teeth was analysed using the proposed method to study the time to ankylosis complication and simulation studies were conducted which show that the penalisation technique provides a good fit of the baseline hazard and precise estimations of the resulting regression parameters.

Keywords. Adaptive Ridge procedure, Cure model, EM algorithm, Interval censoring, Penalised likelihood, Piecewise constant hazard.

1 Introduction

Interval censored data arise in situations where the event of interest is only known to have occurred between two observation times. These types of data are commonly encountered when the patients are intermittently followed up at medical examinations. This is the case for instance in AIDS studies, when HIV infection onset is determined by periodic testing, or in oncology where the time-to-tumour progression is assessed by measuring the tumour size at periodic testing. Dental data are another examples which are usually interval-censored because the teeth status of the patients are only examined at visits to the dentist. While interval-censored data are ubiquitous in medical applications it is still a common practice to replace the observation times with their midpoints or endpoints and to consider these data as exact. This allows to analyse the data using standard survival approach but may result in a large bias of the estimators.

In the present talk we study the Cox model with piecewise constant baseline hazard. Treating the unobserved true event times as missing variables we use the EM algorithm to perform estimation. As a result, the Hessian of the log-likelihood to be maximised is seen to be diagonal. This is a remarkable feature of the method that easily allows to perform estimation with the piecewise constant baseline using arbitrarily large set of cuts. In contrast, this model had been already introduced in [2] but maximisation of the model parameters was achieved using the observed likelihood which resulted in a full rank Hessian matrix. As a consequence, the authors warn against computational issues which may force the user to reduce the number of cuts by combining adjacent intervals. Taking advantage of the sparse structure of the Hessian matrix, our method can be combined with a L_0 penalty designed to detect the location and number of cuts. This is performed through the adaptive ridge procedure, a regularisation method that was introduced in [3] and then applied in a survival context (without covariates) in [1]. This penalisation technique results in a flexible method where the cuts and locations of the piecewise constant baseline are automatically chosen from the data, thus providing a good compromise between purely non-parametric and parametric baseline functions.

Another advantage of using the EM algorithm is to provide direct extensions of the Cox model. In this work we also consider the inclusion of exact data and the inclusion of a fraction of non-susceptible patients in the estimation method. This last situation is modelled using the cure model of [4], with a logit link for the probability of being cured. With our method, those two extensions are straightforward. In particular, the E-step in the cure model results in a weighted log-likelihood with the weights corresponding to the probability of being cured.

2 A piecewise constant hazard model for interval censored data

Let T denote the time to occurrence of the event of interest. We consider a situation where all individuals are subject to interval censoring defined by the random variables (L, R) such that L and R are observed and $\mathbb{P}(T \in [L, R]) = 1$. The situation $L = 0$ and $R < \infty$ corresponds to left-censoring, $0 < L < R < \infty$ corresponds to strictly interval censoring and $L < R = \infty$ to right censoring. The special case $L = R$ is also allowed which corresponds to exact observations of the time of interest. We introduce a column covariate vector Z of dimension d_Z and for convenience we also introduce δ which equals 0 if an individual is right censored and 1 if he/she is left, interval censored or exactly observed. The variable T is considered continuous and we assume independent censoring in the following way: $\mathbb{P}(T \leq t \mid L = l, R = r, Z) = \mathbb{P}(T \leq t \mid l \leq T \leq r, Z)$. This supposes that the variables (L, R) do not convey additional information on the law of T apart from assuming T to be bracketed by L and R. Finally, we assume non-informative censoring in the sense that the distribution of L and R does not depend on the model parameters involved in the distribution of T.

We consider the following Cox proportional hazard model for the time variable T:

$$\lambda(t \mid Z) = \lambda_0(t) \exp(\beta Z),$$

where β is an unknown row parameter vector of dimension d_Z. We model the baseline function λ_0 through a piecewise constant hazard. Let $c_0, c_1, \ldots, c_K$ represent $K + 1$ cuts, with the convention that $c_0 = 0$ and $c_K = +\infty$. Let $I_k(t) = I(c_{k-1} < t \leq c_k)$, with $I(\cdot)$ denoting the indicator function. We suppose that $\lambda_0(t) = \sum_{k=1}^{K} I_k(t) \exp(a_k)$. Under this model, note that the survival and density functions are respectively equal to:

$$S(t \mid Z) = \exp\Big(- \sum_{k=1}^{K} e^{a_k + \beta Z} (t \wedge c_k - c_{k-1}) I(c_{k-1} \leq t) \Big),$$

$$f(t \mid Z) = \sum_{k=1}^{K} I_k(t) \exp\Big(a_k + \beta Z - \sum_{j=1}^{k} e^{a_j + \beta Z} (t \wedge c_j - c_{j-1}) \Big).$$

We set $\boldsymbol{\theta} = (a_1, \ldots, a_K, \beta)$ the model parameter we aim to estimate. In the following, we will also study the so-called nonparametric situation, when no covariates are available, which is encompassed in our modelling approach as the special case where $Z = 0$. In this context the hazard function is simply equal to λ_0 which is assumed to be piecewise constant and the model parameter is $\boldsymbol{\theta} = (a_1, \ldots, a_K)$. The observed data consist of data $= \{\text{data}_i, i = 1, \ldots, n\}$ with $\text{data}_i = (L_i, R_i, \delta_i, Z_i)$ while T_i is considered as incompletely observed. We introduce the notation $a_{i,k} = a_k + \beta Z_i$.

3 The EM algorithm for left, right and interval censored observations

In the following, we only develop the method for left, right and interval censored observations. The EM algorithm is based on the complete likelihood, defined by: $\mathrm{L}_n(\boldsymbol{\theta}) = \prod_{i=1}^n f(T_i \mid Z_i, \boldsymbol{\theta})$. Denote by $\boldsymbol{\theta}_{\text{old}}$ the current parameter value. The E-step takes the expectation of the complete log-likelihood with respect to the T_i's, given the L_i's, R_i's, δ_i's, Z_i's and $\boldsymbol{\theta}_{\text{old}}$. Write

$$Q_i(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{old}}) := \mathbb{E}[\log(f(T_i \mid Z_i, \boldsymbol{\theta})) \mid \text{data}_i, \boldsymbol{\theta}_{\text{old}}] = \int f(t \mid \text{data}_i, \boldsymbol{\theta}_{\text{old}}) \log f(t \mid Z_i, \boldsymbol{\theta}) dt,$$

where $f(t \mid \text{data}_i, \boldsymbol{\theta}_{\text{old}})$ represents the conditional density of T_i given data_i and $\boldsymbol{\theta}_{\text{old}}$, evaluated at t. Under the independent censoring assumption,

$$f(t \mid \text{data}_i, \boldsymbol{\theta}_{\text{old}}) = \frac{f(t \mid Z_i, \boldsymbol{\theta}_{\text{old}}) I(L_i < t < R_i)}{S(L_i \mid Z_i, \boldsymbol{\theta}_{\text{old}}) - S(R_i \mid Z_i, \boldsymbol{\theta}_{\text{old}})}.$$

The E-step consists of computing the quantity $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{old}}) = \sum_i Q_i(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{old}})$. We have:

$$\begin{aligned}
Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{old}}) =& \sum_{i=1}^n \frac{\int_{L_i}^{R_i} f(t \mid Z_i, \boldsymbol{\theta}_{\text{old}}) \log f(t \mid Z_i; \boldsymbol{\theta}) dt}{S(L_i \mid Z_i, \boldsymbol{\theta}_{\text{old}}) - S(R_i \mid Z_i, \boldsymbol{\theta}_{\text{old}})} \\
Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{old}}) =& \sum_{i=1}^n \left\{ \frac{1}{S(L_i \mid Z_i, \boldsymbol{\theta}_{\text{old}}) - S(R_i \mid Z_i, \boldsymbol{\theta}_{\text{old}})} \right. \\
& \left. \times \sum_{k=1}^K J_{k,i} \int_{c_{k-1} \vee L_i}^{c_k \wedge R_i} \exp\left(a_{i,k}^{\text{old}} - \sum_{j=1}^k e^{a_{i,j}^{\text{old}}} (t \wedge c_j - c_{j-1})\right) \left(a_{i,k} - \sum_{j=1}^k e^{a_{j,k}} (t \wedge c_j - c_{j-1})\right) dt \right\},
\end{aligned}$$

where $J_{k,i}$ is the indicator $I\{(L_i, R_i) \cap (c_{k-1}, c_k) \neq \emptyset\}$ and $b_1 \wedge b_2$, $b_1 \vee b_2$ respectively denote $\min(b_1, b_2)$, $\max(b_1, b_2)$. Finally, the M-step corresponds of maximising, with respect to $\boldsymbol{\theta}$, the quantity

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{old}}) = \sum_{i=1}^n \sum_{k=1}^K \left\{ \left(a_{i,k} - \sum_{j=1}^{k-1} (c_j - c_{j-1}) e^{a_{i,j}}\right) A_{k,i}^{\text{old}} - e^{a_{i,k}} B_{k,i}^{\text{old}} \right\},$$

where exact expressions of the statistics $A_{k,i}^{\text{old}}$ and $B_{k,i}^{\text{old}}$ are not given here.

Maximisation of $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{old}})$ is then performed using the Newton-Raphson algorithm. It should be noted that the Hessian is diagonal and can thus be computed in $\mathcal{O}(K)$ computation time.

4 A penalised EM algorithm

If the number of cuts is not known in advance, we choose a large grid of cuts (i.e K large) and we penalise the log-likelihood in the manner of [3] and [1]. This penalisation is designed to enforce consecutive values of the a_ks that are close to each other to be equal. It is defined in the following way:

$$\ell^{\text{pen}}(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{old}}) = Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{old}}) - \frac{\text{pen}}{2} \sum_{k=1}^{K-1} \hat{w}_k (a_{k+1} - a_k)^2, \tag{1}$$

where $\hat{\boldsymbol{w}} = (\hat{w}_1, \ldots, \hat{w}_{K-1})$ are non-negative weights that will be iteratively updated in order for the weighted ridge penalty term to approximate the L_0 penalty. The pen term is a tuning parameter that describes the degree of penalisation. Note that the two extreme situations pen $= 0$ and pen $= \infty$ respectively correspond to the unpenalised log-likelihood model of the previous section and to the Cox model with exponential baseline.

It can easily be shown that the block matrix corresponding to the second order derivatives with respect to the a_ks is tridiagonal. Taking advantage of this, inversion of the Hessian can be performed in $\mathcal{O}(K)$ computation time. Therefore, for a given value of pen and of the weight vector $\hat{\boldsymbol{w}}$, maximisation can be easily performed using the Newton-Raphson algorithm. Once the Newton-Raphson algorithm has reached convergence, the weights are updated at the lth step from the equation

$$\hat{w}_k^{(l)} = \left((\hat{a}_{k+1}^{(l)} - \hat{a}_k^{(l)})^2 + \varepsilon^2 \right)^{-1},$$

for $k = 1, \ldots, K-1$ with $\varepsilon = 10^{-5}$ (recommended value from [3]) and where the $\hat{a}_k^{(l)}$'s represent the estimates of the a_k's obtained through the Newton-Raphson algorithm. This form of weights is motivated by the fact that $w_k(a_{k+1} - a_k)^2$ is close to 0 when $\mid a_{k+1} - a_k \mid < \varepsilon$ and close to 1 when $\mid a_{k+1} - a_k \mid > \varepsilon$. Hence the penalty term tends to approximate the L_0 norm. The weights are initialized by $\hat{w}_k^{(0)} = 1$, which gives the standard ridge estimate of $\boldsymbol{a}$.

Finally, for a given value of pen, once the adaptive ridge algorithm has reached convergence, a set of cuts is found for the $\hat{a}_k$'s verifying $\hat{w}_k(\hat{a}_{k+1} - \hat{a}_k)^2 > 0.99$. This hard thresholding allows to provide a sparse collection of cuts. The non-penalised log-likelihood Q is then maximised using this set of cuts and the final maximum likelihood estimate is derived using the results of the previous section. It is important to stress that the penalised likelihood is used only to select a set of cuts. Reimplementing the non-penalised log-likelihood Q in the final step enables to reduce the bias classically induced by penalised maximisation techniques.

5 Choice of the penalty term

A Bayesian Information Criterion (BIC) is introduced in order to choose the penalty term. As explained in the previous section, for each penalty value the penalised EM likelihood (1) selects a set of cuts. For a selected set of cuts we denote by m the total number of parameters to be estimated and by $\hat{\boldsymbol{\theta}}_m$ the corresponding non-penalised estimated model parameter obtained by maximisation of the Q function. The BIC is then defined as: $\text{BIC}(m) = -2\log(\text{L}_n^{\text{obs}}(\hat{\boldsymbol{\theta}}_m)) + m\log(n)$, where L_n^{obs} represents the observed likelihood.

Note that the BIC is expressed here in terms of selected models. Since different penalty values can yield the same selection of cuts, the BIC needs only to be computed for all different selected models (and not for all different penalties). The final set of cuts along with its estimator $\hat{\boldsymbol{\theta}}_{\hat{m}}$ is chosen such that $\text{BIC}(\hat{m})$ is minimal.

6 Conclusion

The method has been further extended to the inclusion of exact data and a fraction of non-susceptible individuals. Those two extensions are straightforward thanks to the EM algorithm. From a theoretical point of view, we also studied the asymptotic distribution of the estimated parameters obtained from the adaptive ridge algorithm. We proved that inference after selection of the cuts is still valid if we consider the cuts as fixed. Finally, extensive simulation studies have been conducted which show the good performance of the proposed method. The method has also been applied to a dental dataset of replanted teeth which allowed to display prognostic covariates for the risk of ankylosis and also to detect specific areas of time were patients are at greater risk of developing the complication and should therefore be closely monitored.

References

[1] O. Bouaziz and G. Nuel. L0 regularization for the estimation of piecewise constant hazard rates in survival analysis. *Applied Mathematics*, 8(3), 2017.

[2] B. Carstensen. Regression models for interval censored survival data: application to hiv infection in danish homosexual men. *Statistics in Medicine*, 15(20):2177–2189, 1996.

[3] F. Frommlet and G. Nuel. An adaptive ridge procedure for l_0 regularization. *PLoS ONE*, 11(2), 2016.

[4] J. P. Sy and J. M. Taylor. Estimation in a cox proportional hazards cure model. *Biometrics*, 56(1):227–236, 2000.

Improved estimation of the precision matrix of a mixture of Wishart distributions in high dimensions

Djamila Boukehil [1,2] & Dominique Fourdrinier [1] & Fatiha Mezoued [2]& William E. Strawderman [3]

[1]*LITIS-Université de Rouen, FRANCE* [2]*LAMOPS-ENSSEA, ALGÉRIE* [3]*Rutgers University, USA*

djamila.boukehil@etu.univ-rouen.fr dominique.fourdrinier@univ-rouen.fr
famezoued@yahoo.fr straw@stat.rutgers.edu

Résumé. Dans ce travail, nous nous intéressons à l'estimation de la matrice de précision Σ^{-1}, de dimension $p \times p$, pour des distributions mélange de lois de Wishart sous les coûts de type Efron-Morris $\mathrm{tr}[\{\hat{\Sigma}^{-1} - \Sigma^{-1}\}^2 S^k]$, pour tout $k = 0, 1, 2 \ldots$. Nous proposons des estimateurs orthogonalement invariants qui améliorent les estimateurs usuels de la forme $a\, S^+$, où a est une constante positive et S^+ est l'inverse de Moore Penrose de la matrice de covariance empirique S.

Mots-clés. Matrice de précision, distributions mélanges de lois de Wishart, identité de type Haff-Stein.

Abstract. This paper deals with the problem of estimating the precision matrix for a scale mixture of Wishart distributions under Efron-Morris type losses $\mathrm{tr}[\{\hat{\Sigma}^{-1} - \Sigma^{-1}\}^2 S^k]$, for $k = 0, 1, 2 \ldots$, where S is the sample covariance matrix. We derive orthogonally invariant estimators that dominate estimators of the form $a\, S^+$, where a is a positive constant and S^+ is the Moore Penrose inverse of S.

Keywords. Precision matrix, scale mixtures of Wisharts, Stein-Haff type identity.

1 Introduction

Many data analysis techniques necessitate a good estimate of the precision matrix Σ^{-1} of the underlying distribution of the data, especially when the dimension p of the variable is large compared to the sample size n. In such a situation, the standard estimator based on the inverse of the sample covariance matrix S cannot be defined. Kubokawa and Srivastava (2008) considered estimators based on the Moore Penrose inverse S^+ of S which are of the form

$$\hat{\Sigma}_a^{-1} = a\, S^+, \tag{1.1}$$

where a is a positive constant. These authors proposed alternative estimators $\hat{\Sigma}^{-1}$ of Σ^{-1}, improving on the usual estimators $\hat{\Sigma}_a^{-1}$ in the Gaussian setting. In the same distributional

context, Kubokawa et Inoue (2014) provided general types of ridge estimators for Σ^{-1} and they showed, through simulations, that for a specific ridge parameter their estimators are better than Σ_a^{-1} under $\mathrm{tr}[\{\hat{\Sigma}^{-1}\Sigma - I\}^2]$ loss. Fourdrinier, Mezoued and Wells (2016) provided estimators of Σ_a^{-1}, in a general elliptical setting, and proved that these estimators improved over Σ_a^{-1} under the quadratic loss $\mathrm{tr}[\{\hat{\Sigma}^{-1} - \Sigma^{-1}\}^2]$.

In this work, we deal with an observed $p \times p$ matrix S from the mixture model

$$\begin{cases} S \mid V \sim \mathcal{W}_p(n, V\,\Sigma) \\ V \sim \mathcal{H}(\cdot) \end{cases}, \tag{1.2}$$

where $\mathcal{W}_p(n, V\,\Sigma)$ denotes the Wishart distribution with n degrees of freedom and covariance matrix $V\,\Sigma$ and where $\mathcal{H}(\cdot)$ is a distribution on $\mathbb{R}_+$. Note that we may view the distribution of S as that of $V\,\tilde{S}$, where $\tilde{S}$ has the Wishart distribution $\mathcal{W}_p(n, \Sigma)$ and is independent of V.

This paper is structured as follows. In Section 2, we consider alternative estimators $\hat{\Sigma}^{-1}$ of Σ^{-1} which we evaluate under the risk associated with Efron-Morris type loss functions

$$R_k(\Sigma^{-1}, \hat{\Sigma}^{-1}) = \mathrm{E}_\Sigma[\mathrm{tr}[\{\hat{\Sigma}^{-1} - \Sigma^{-1}\}^2 S^k]], \tag{1.3}$$

for $k = 0, 1, 2...$, where E_Σ denotes the expectation with respect to the model in (1.2) and $\mathrm{tr}\{A\}$ denotes the trace of A. We provide an unbiased estimator of the risk difference between $\hat{\Sigma}^{-1}$ and $\hat{\Sigma}_a^{-1}$ thanks to a new Stein-Haff type identity. Then we derive a sufficient domination condition of $\hat{\Sigma}^{-1}$ over $\hat{\Sigma}_a^{-1}$ in (1.1). In Section 3, we show that there exists an optimal constant a^* for the estimators $a\,S^+$, when $k = 1, 2, 3$. We suggest choices of a for $k = 0$ and $k \geq 4$. We illustrate these dominance results with Haff type estimators. We propose in Section 4 a generalization of these alternative estimators. Conclusions are stated in Section 5.

2 Competitive estimators

We consider estimators of the form

$$\hat{\Sigma}_{a,c}^{-1} = a\,S^+ + c\,S\,G(S), \tag{2.1}$$

where c is a constant and the matrix function $G(S)$ is homogeneous of order α.

The following lemma gives the so-called Stein-Haff identity for the singular Wishart distribution $\mathcal{W}_p(n, V\,\Sigma)$ derived by Fourdrinier, Haddouche and Mezoued (2019). This identity will be used to develop the risk difference between any alternative estimator in (2.1) and the usual estimator in (1.1).

Lemma 2.1. *Let $\tilde{S}$ be a $p \times p$ matrix having a Wishart distribution $\mathcal{W}_p(n, \Sigma)$. For any $p \times p$ matrix function $G(\tilde{S})$ which is weakly differentiable with respect to $\tilde{S}$ and such that $\mathrm{E}_\Sigma\left[|\mathrm{tr}\{\Sigma^{-1}\,\tilde{S}\,G(\tilde{S})|\}\right] < \infty$, we have*

$$\mathrm{E}_\Sigma\left[\mathrm{tr}\{\Sigma^{-1}\,\tilde{S}\,G(\tilde{S})\}\right] = \mathrm{E}_\Sigma\left[\mathrm{tr}\{-\tilde{S}^+\,\tilde{S}\,G(\tilde{S}) + 2\,\tilde{S}^+\,\tilde{S}\,D_{\tilde{S}}\,\{G^\top(\tilde{S})\,\tilde{S}\}\}\right],$$

where the differential operator $D_{\tilde{S}}$ for a matrix $\tilde{S}$ is defined by

$$D_{\tilde{S}} = \left(\frac{1}{2}(1+\delta_{ij})\,\frac{\partial}{\partial \tilde{S}_{ij}}\right)_{1\le i,j\le p},$$

with $\delta_{ij} = 1$ when $i = j$ and $\delta_{ij} = 0$ when $i \neq j$.

Thanks to Lemma 2.1, we obtain the following expression of the risk difference between estimators $\hat{\Sigma}_{a,c}$ and $\hat{\Sigma}_a$.

Theorem 2.1. *Assume that $R_k(\Sigma^{-1}, \hat{\Sigma}_a^{-1}) < \infty$ and that $R_k(\Sigma^{-1}, \hat{\Sigma}_{a,c}^{-1}) < \infty$. Assume also that $S^k\,G(S)$ satisfies the conditions of Lemma 2.1. Then the risk difference between $\hat{\Sigma}_{a,c}^{-1}$ in (2.1) and $\hat{\Sigma}_a^{-1} = a\,S^+$ is expressed through $\tilde{S}$ as*

$$\Delta(G) = \mathrm{E}_\Sigma\left[\delta(G)\right]$$

where $\delta(G)$ is an unbiased estimator of the risk difference which equals

$$\begin{aligned}\delta(G) =& c^2\,\mu_{k+2\alpha+2}\,\mathrm{tr}\{\tilde{S}^k\,[\tilde{S}\,G(\tilde{S})]^2\}\\ &+ 2\,c\left[a\,\mu_{k+\alpha} + \mu_{k+\alpha+1}\right]\mathrm{tr}\{\tilde{S}^+\,\tilde{S}^{k+1}\,G(\tilde{S})\}\\ &- 4\,c\,\mu_{k+\alpha+1}\mathrm{tr}\{\tilde{S}^+\,\tilde{S}\,D_{\tilde{S}}\,\{G^\top(\tilde{S})\,\tilde{S}^{k+1}\}\},\end{aligned} \tag{2.2}$$

where, for $\beta \in \mathbb{R}$, $\mu_\beta = E_{\mathcal{H}}[V^\beta]$ is the moment of order β of V. As a consequence, a sufficient condition for improvement of $\hat{\Sigma}_{a,c}^{-1}$ over $\hat{\Sigma}_a^{-1}$ is that $\delta(G)$ is non positive.

3 Haff type estimators

Let $G(S) = S^+/\mathrm{tr}\{S\}$, with homogeneity order $\alpha = -2$. We consider Haff type estimators of the form

$$\hat{\Sigma}_{HF}^{-1} = a\,S^+ + c\,S\,S^+/\mathrm{tr}\{S\}. \tag{3.1}$$

By applying Theorem 2.1 with $G(S) = S^+/\mathrm{tr}\{S\}$, $\delta(G)$ in (2.2) is non positive for c between 0 and

$$C_{ak} = 2\left[\frac{\mu_{k-1}}{\mu_{k-2}}\left(p - n - 1 + 2\,k\right) - a\right]\frac{\mathrm{tr}\{L^{k-1}\}\,\mathrm{tr}\{L\}}{\mathrm{tr}\{L^k\}} - 4\,\frac{\mu_{k-1}}{\mu_{k-2}} + 2\,\frac{\mu_{k-1}}{\mu_{k-2}}\,B_{k+1}\,\frac{\mathrm{tr}\{L\}}{\mathrm{tr}\{L^k\}}, \tag{3.2}$$

where $L = (\mathrm{diag}(l_i))_{1\le i\le n}$ with $l_1 > \cdots > l_n > 0$ in the singular value decomposition $S = H_1 L H_1^\top$, with the diagonalizing matrix H_1 satisfying $H_1^\top H_1 = I_n$, and where

$$B_{k+1} = \sum_{i=1}^{n}\sum_{j\neq i}^{n} \frac{l_i^k - l_j^k}{l_i - l_j}.$$

In order to illustrate domination of Haff type estimators in (3.1), we need to figure out possible optimal estimators within the class $\{a\, S^+/a > 0\}$, that is, estimators which minimize the risk $R_k(\Sigma^{-1}, a\, S^+)$ for some $a > 0$.

By using Lemma 2.1, the risk of the estimator $a\, S^+$ under model (1.2) is given by

$$\begin{aligned} R_k(\Sigma^{-1}, a\, S^+) =& \mathrm{E}_\Sigma[a^2\, \mu_{k-2}\, \mathrm{tr}\{\tilde{S}^k\, (\tilde{S}^+)^2\} + 2\, a\, \mu_{k-1}\, \mathrm{E}_\Sigma[\mathrm{tr}\{\tilde{S}^k\, (\tilde{S}^+)^2\}] \\ &+ 4\, a\, \mu_{k-1}\, \mathrm{E}_\Sigma[\mathrm{tr}\{\tilde{S}^+\, \tilde{S}\, D_{\tilde{S}}\, \{\tilde{S}^k\, \tilde{S^+}\}\}] + \mu_k\, \mathrm{E}_\Sigma[\mathrm{tr}\{\Sigma^{-2}\, S^k\}]. \end{aligned} \tag{3.3}$$

This risk in (3.3) is minimized at

$$a^* = \frac{\mu_{k-1}}{\mu_{k-2}} \left(2\, \frac{\mathrm{E}_\Sigma[\mathrm{tr}\{\tilde{S}^+\, \tilde{S}\, D_{\tilde{S}}\, \{\tilde{S}^k\, \tilde{S^+}\}\}]}{\mathrm{E}_\Sigma[\mathrm{tr}\{\tilde{S}^k\, (\tilde{S}^+)^2\}]} - 1 \right). \tag{3.4}$$

For $k = 1, 2$ and 3, the optimal value a^* in (3.4) is a constant and equals

- $\dfrac{1}{\mu_{-1}}\,[p - n - 1]$ if $k = 1$;
- $\mu_1\, p$ if $k = 2$;
- $\dfrac{\mu_2}{\mu_1}\,[p + n + 1]$ if $k = 3$.

For $k = 0$ and $k \ge 4$, a^* in (3.4) depends on Σ. However, a lower bound a_0^- and an upper bound a_0^+ exist for a^* so that, in each case, for $a < a_0^-$, $a_0^-\, S^+$ improves over $a\, S^+$ and, for $a > a_0^+$, $a_0^+\, S^+$ improves over $a\, S^+$. Specifically,

- $a_0^- = \dfrac{\mu_{-1}}{\mu_{-2}}\,\big[p - 2\,(n+1)\big]$ and $a_0^+ = \dfrac{\mu_{-1}}{\mu_{-2}}\,\big[p - n - 3\big]$ if $k = 0$;
- $a_0^- = \dfrac{\mu_{k-1}}{\mu_{k-2}}\,\big[p + n + k - 2\big]$ and $a_0^+ = \dfrac{\mu_{k-1}}{\mu_{k-2}}\,\big[p + (k-2)\,n + k - 2\big]$ if $k \ge 4$.

From (3.2) and thanks to the above optimal values a^* of a associated to the different values of k, we give a sufficient condition for domination of $\hat{\Sigma}_{HF}^{-1}$ over $\hat{\Sigma}_a^{-1}$ in the following proposition.

Proposition 3.1. *For $a = a^*$ given above, any Haff type estimator $\hat{\Sigma}_{HF}^{-1} = a^*\, S^+ + c\, S\, S^+/\mathrm{tr}\{S\}$ improves on the optimal estimator $a^*\, S^+$, and hence, improves on any estimator $a\, S^+$ with $a > 0$, when $k = 1$ and 2, as soon as*

- $0 < c < 2\,\dfrac{1}{\mu_{-1}}\,\big[n(n+1)-2\big]$ *if* $k = 1$ *and* $n \geq 2$;
- $0 < c < 2\,\mu_1\,[n+1] - 4\,\mu_1$ *if* $k = 2$ *and* $n \geq 2$.

When $k = 3$, *although* $C_{a^*\,3} > 0$, *that quantity is not constant and cannot be bounded from below by a positive constant. However, for* $a < a^*$, *as soon as*

- $0 < c < 2\,\dfrac{\mu_2}{\mu_1}\,(a^* - a) = 2\,\dfrac{\mu_2}{\mu_1}\,(n + p + 1 - a)$,

the estimator $\hat{\Sigma}^{-1}_{HF}$ *improves on* $a\,S^+$.

For $a = a_0^-$ *given above, any Haff type estimator* $\hat{\Sigma}^{-1}_{HF} = a_0^-\,S^+ + c\,S\,S^+/\mathrm{tr}\{S\}$ *improves on* $a_0^-\,S^+$, *and hence, improves on any estimator* $a\,S^+$ *with* $a < a_0^-$, *as soon as*

- $0 < c < 2\,\dfrac{\mu_{-1}}{\mu_{-2}}\,(n-1)$ *if* $k = 0$ *and* $n \geq 2$;
- $0 < c < 2\,\dfrac{\mu_{k-1}}{\mu_{k-2}}\,(k-3)$ *if* $k \geq 4$.

4 A larger class of estimators

Now, we propose a class of estimators which extends the estimators in (2.1), provided $G(S)$ is orthogonally invariant. Let

$$\hat{\Sigma}^{-1}_{a,c,r} = a\,S^+ + c\,r\big(\mathrm{tr}\{S\}\big)\,S\,G(S)\,, \tag{4.1}$$

where c is a constant, the matrix function $G(S)$ is homogeneous of order -2 and r is a real valued function. We assume that G is orthogonally invariant, that is, of the form

$$G = H_1\,\Psi(L)\,H_1^\top\,, \tag{4.2}$$

where $\Psi(L)$ is a $n \times n$ diagonal matrix, $\mathrm{diag}\big(\psi_1(L), \ldots, \psi_n(L)\big)$.

Theorem 4.1. *Let* $k = 0, 1, 2...$ *be fixed. Under model (1.2) and risk (1.3), suppose that the function* $G(S)$ *is orthogonally invariant as in (4.2), homogeneous of order* -2 *and weakly differentiable, and that all moments in Theorem 2.1 exist. Suppose also that, for any* $i = 1, \ldots, n$, $\psi_i(L) > 0$.

Suppose that, for some $c > 0$, $\delta(G)$ *in (2.2) is non positive so that the estimator* $\hat{\Sigma}^{-1}_{a,c} = a\,S^+ + c\,S\,G(S)$ *in (2.1) improves over the estimator* $\hat{\Sigma}^{-1}_{a} = a\,S^+$ *in (1.1). Then the estimator* $\hat{\Sigma}^{-1}_{a,c,r} = a\,S^+ + c\,r\big(\mathrm{tr}\{S\}\big)\,S\,G(S)$ *in (4.1) improves over* $\hat{\Sigma}^{-1}_{a} = a\,S^+$ *provided that the function* r *is differentiable and that, for any* $t \geq 0$, $0 \leq r(t) \leq 1$ *and* $r'(t) \geq 0$.

Note that, as they are orthogonally invariant, Theorem (4.1) can be illustrated with the improved Haff type estimator in Proposition (3.1).

5 Concluding remarks

In this paper, we have considered estimation of the inverse scatter matrix Σ^{-1} of a scale mixture of Wishart distribution under Efron-Morris type losses. We have shown that the standard estimators of the form $a\,S^{+}$ can be improved by alternative estimators through an unbiased estimator of risk difference, thanks to a new Stein-Haff type identity developed by Fourdrinier, Haddouche and Mezoued (2019). Our results extend several classical domination results for the Wishart case to the scale mixture of Wishart case, and also extend results for specific Efron and Morris type losses to the entire class.

Bibliographie

D. Fourdrinier, A.M. Haddouche, F. Mezoued. (2019). Covariance matrix estimation of an elliptically symetric distribution. Technical report, Université de Rouen Normandie and ENSSEA de Tipaza.

D. Fourdrinier, F. Mezoued, M.T. Wells. (2016). Estimation of the inverse scatter matrix of an elliptically symmetric distribution, Journal of Multivariate Analysis, 143, 32–55.

T. Kubokawa, M.S. Srivastava. (2008). Estimation of the precision matrix of a singular Wishart distribution and its application in high-dimensional data, Journal of Multivariate Analysis, 99, 1906–1928.

Kubokawa, T. and Inoue, A. (2014). Estimation of covariance and precision matrices under scale-invariant quadratic loss in high dimension, Electronic Journal of Statistics, 8, 130–158.

Extreme Partial Least-Squares regression

Meryem Bousebata[1,2], Geoffroy Enjolras[2] & Stéphane Girard[1]

[1] *Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France.*
[2] *Univ. Grenoble Alpes, CERAG, 38000 Grenoble, France.*
meryem.bousebata@inria.fr, geoffroy.enjolras@grenoble-iae.fr, stephane.girard@inria.fr

Résumé. L'objectif de cette communication est de proposer une nouvelle approche, appelée Extreme-PLS, pour la réduction de dimension en régression qui soit adaptée aux queues de distributions. Nous nous intéressons aux combinaisons linéaires des prédicteurs qui expliquent au mieux les valeurs extrêmes de la variable réponse dans un contexte de régression inverse non linéaire. Les performances de la méthode sont évaluées par simulations numériques. Une analyse statistique de données du revenu agricole français, considérant des valeurs extrêmes des rendements céréaliers, est fournie à titre d'illustration.

Mots-clés. Valeurs extrêmes, Réduction de dimension, Régression inverse non linéaire, Partial Least Squares.

Abstract. The objective of this communication is to propose a new approach, called Extreme-PLS, for dimension reduction in regression and adapted to distribution tails. We are interested in the linear combinations of predictors that best explain the extreme values of the response variable in a non-linear inverse regression context. The performance of the method is evaluated on numerical simulations. A statistical analysis of French farm income data, considering extreme values of cereal yields, is provided as an illustration.

Keywords. Extreme value, Dimension reduction, Non-linear inverse regression, Partial Least Squares.

1 Introduction

Context. Regression analysis is widely used to study the relationship between a response variable Y and an explanatory p-dimensional vector X. When p grows, a dimension reduction becomes necessary to show only the most relevant directions of high-dimensional data. There exist a number of statistical models for dimension reduction in regression problems. One of the most popular is Partial Least Squares (PLS) regression that combines the characteristics of Principal Component Analysis (PCA) and multiple regression (Tenenhaus, 1999). Its purpose is to find linear combinations of the X coordinates highly correlated with Y. Sliced Inverse Regression (SIR) is an alternative method for dimension reduction in regression which explores the simplicity of the inverse regression view of X against Y (Li, 1991). It aims at replacing X by its projection onto a subspace of smaller

dimension without loss of information. At the same time, there is a growing interest for the modelling of conditional extremes, *i.e.* extremes depending on a covariate. One can mention for instance the estimation of conditional extreme quantiles or more generally, the tail of conditional distributions (Gardes & Girard, 2010). In this communication, we aim to deal with these two lines of works (dimension reduction in regression and conditional extremes) by looking for a linear combination $\beta^t X$ of the covariates that best explains the extreme values of Y in a context of non linear inverse regression. More precisely, we propose an approach, called Extreme-PLS, to estimate β by maximization of the covariance between $\beta^t X$ and Y given Y exceeds a high threshold y. Such an adaptation of the PLS estimator to the extreme-value framework thanks to the non-linear inverse regression model is of great interest. First, it allows for a visual interpretation of conditional extremes by providing the linear projection of X that best explains the large values of Y. Second, the dimension reduction should yield improved rates of convergence for most estimators dealing with conditional extreme values.

An inverse model. Let us consider the following non linear inverse regression model:

(M) $X = g(Y)\beta + \varepsilon$, where X is a p-dimensional random vector, Y is a real random variable, $g : \mathbb{R} \to \mathbb{R}$ is the link function and ε is p-dimensional random vector of error. The parameter $\beta \in \mathbb{R}^p$ is an unknown unit vector, ε may depend on Y and g is an unknown function.

Similar inverse regression models were used to establish the theoretical properties of SIR (Bernard-Michel, Gardes & Girard, 2008). Under model (M), we aim to estimate β by maximizing the covariance between $\beta^t X$ and Y for large values of Y. Indeed, roughly speaking, when Y is large, ε becomes negligible, one thus has $X \simeq g(Y)\beta$ and thus $Y \simeq g^{-1}(\beta^t X)$ without (in)dependence assumption.
The paper is organized as follows. In Section 2, the Extreme-PLS approach is introduced and its theoretical properties are stated. The associated estimator is exhibited in Section 3. The performances of the estimator are investigated through a simulation study in Section 4. Our estimation procedure is applied to study the influence of various parameters on cereal yield observed on French farms in Section 5.

2 Extreme-PLS approach

Let us denote by $w(y)$ the unit vector maximizing the covariance between $w^t X$ and Y given that Y exceeds a large threshold y:

$$w(y) = \arg \max_{\|w\|=1} \operatorname{cov}(w^t X, Y | Y \geq y). \tag{1}$$

This optimization problem benefits from a closed-form solution given in the next proposition and obtained by solving the constrained optimization problem using Lagrange multipliers. For all $y \in \mathbb{R}$, let us denote by $\bar{F}(y) = \mathbb{P}(Y \geq y)$ the survival function of Y and the

tail-moments, whenever they exist, $m_Y(y) = \mathbb{E}(Y\mathbb{1}_{\{Y\geq y\}}) \in \mathbb{R}$, $m_X(y) = \mathbb{E}(X\mathbb{1}_{\{Y\geq y\}}) \in \mathbb{R}^p$, $m_{XY}(y) = \mathbb{E}(XY\mathbb{1}_{\{Y\geq y\}}) \in \mathbb{R}^p$.

Proposition 1. *Suppose that* $\mathbb{E}\|X\| < \infty$, $\mathbb{E}|Y| < \infty$ *and* $\mathbb{E}\|XY\| < \infty$. *Then, the solution of the optimization problem (1) is:*

$$w(y) = v(y)/\|v(y)\| \text{ where } v(y) = \bar{F}(y)m_{XY}(y) - m_X(y)m_Y(y). \tag{2}$$

Let us note that the solution (2) is invariant with respect to the scaling and location of X. In order to establish the convergence of $w(y)$ to β as $y \to \infty$, we use the proximity criterion which measures the similarity between two vectors of an inner product space. It is measured by the cosine of the angle between two vectors and determines whether two vectors are pointing in roughly the same direction. Here, this criterion corresponds to the squared cosine of the angle between the unit vectors $w(y)$ and β and defined as : $\cos^2(w(y), \beta) = (w(y)^t\beta)^2$. A value close to 0 implies a weak proximity ($w(y)$ is almost orthogonal to β) while a value close to 1 means a high colinearity. Note that, when Y and the error ε are independent, we recover the classical PLS framework and it is easily shown that there is a perfect colinearity between $w(y)$ and β for all $y \in \mathbb{R}$. In the following, no assumption made on the (in)dependence between Y and ε, but additional assumptions on the link function g and the distribution tail of Y are considered:

(A1) Y is a positive random variable with density function f regularly varying at infinity with index $-\frac{1}{\gamma} - 1$, $\gamma \in (0, 1)$ *i.e.* for all $t > 0$,

$$\lim_{y\to\infty} \frac{f(ty)}{f(y)} = t^{-\frac{1}{\gamma}-1}.$$

This property is denoted for short by $f \in RV_{-1/\gamma-1}$.

(A2) $g \in RV_c$ with $c > 0$.

(A3) There exists $q > 1/(\gamma c)$ such that $\mathbb{E}(\|\varepsilon\|^q) < \infty$.

Let us note that (A1) implies that $\bar{F} \in RV_{-1/\gamma}$ which is equivalent to assuming that the distribution of Y is in the Fréchet maximum domain of attraction, with extreme-value index $\gamma > 0$, see de Haan & Ferreira (2007). In other words, (A1) entails that Y has a right heavy-tail. The restriction to $\gamma < 1$ ensures that $\mathbb{E}|Y|$ exists. Assumption (A2) means that the link function asymptotically behaves like a power function. Finally, (A3) is a technical assumption which is satisfied for instance by Gaussian distributions.

Proposition 2. *Assume (M), (A1), (A2) and (A3) hold with* $\gamma < 1/(c + 1)$ *and* $q > 1/(1 - \gamma)$. *Then,* $\cos^2(w(y), \beta) \to 1$ *as* $y \to \infty$.

3 Extreme-PLS: Population version

Let (X_i, Y_i), $1 \leq i \leq n$ be an iid sample from model (M) and let $y_n \to \infty$ as the sample size n tends to infinity. The solution (2) is estimated by its empirical counterpart:

$$\hat{w}(y_n) = \hat{v}(y_n)/\|\hat{v}(y_n)\| \text{ where } \hat{v}(y_n) = \hat{\bar{F}}(y_n)\hat{m}_{XY}(y_n) - \hat{m}_X(y_n)\hat{m}_Y(y_n), \tag{3}$$

with $\hat{\bar{F}}$ the empirical survival function and

$$\hat{m}_{XY}(y_n) = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i \mathbb{1}_{\{Y_i \geq y_n\}}, \hat{m}_Y(y_n) = \frac{1}{n}\sum_{i=1}^{n} Y_i \mathbb{1}_{\{Y_i \geq y_n\}}, \hat{m}_X(y_n) = \frac{1}{n}\sum_{i=1}^{n} X_i \mathbb{1}_{\{Y_i \geq y_n\}}.$$

Establishing the asymptotic behaviour of the random variable $\hat{w}(y_n) - w(y_n)$ as $n \to \infty$ is the goal of our current work.

4 Simulation results

Let us consider a sample of size $n = 1000$ and dimension $p = 3$ from model (M) with a link function $g(t) = t^c$, $t > 0$, $c \in \{1/4, 1/2, 1, 3/2, 2\}$. The behavior of the estimator $\hat{w}(y_n)$ of $\beta = (\sqrt{2}/2, \sqrt{2}/2, 0)$ is illustrated on this regression model where Y is distributed from a Pareto$(a = 2, \alpha = 5)$ distribution with survival function $\bar{F}(y) = (y/a)^{-\alpha}$, $y > a$. The error ε follows the Gaussian distribution $\mathcal{N}(0, \sigma^2 I_3)$, where I_3 is the 3×3 identity matrix. We simulate each component $\varepsilon^{(j)}$ of ε dependent of Y using Frank copula C_θ, where $\theta \geq 0$ is a dependence parameter, see for instance Section 4.2 of (Nelsen, 2007). Let us recall that $\theta = 0$ yields independent margins while large values of θ imply high dependence. The standard deviation σ is chosen such that the Signal Noise Ratio (SNR) is equal to 10: SNR$= \sigma^{-1} g(an^{1/\alpha})$. Note that $g(an^{1/\alpha})$ represents the approximate maximum value of $g(\cdot)$ on a n-sample from the Pareto(a, α) distribution. Finally, we compute the mean proximity criterion between $\hat{w}(y_n)$ and β, on $N = 100$ replications, as follows:

$$\mathrm{PC}(y_n) = \frac{1}{N}\sum_{r=1}^{N} \cos^2(\hat{w}(y_n)^{(r)}, \beta), \tag{4}$$

where $\hat{w}(y_n)^{(r)}$ denotes the estimator (3) computed on the r^{th} replication. For each threshold $y_n = \bar{F}^{-1}(1-\tau)$, where $\tau \in \{0.05, 0.10, 0.95\}$, the $\mathrm{PC}(y_n)$ quality measure is plotted in Figure 1 as a function of the number of exceedances $n\bar{F}(y_n) = n(1-\tau)$, for several values $\theta \in \{0, 10, 20, 30\}$ of the dependence parameter. The closer the measurement is to 1, the better the estimate. It appears that, even when the dependence between Y and ε is strong, the estimator $\hat{w}$ is very efficient for small number of exceedances corresponding to large values of the threshold y_n. As expected, the closer Frank copula parameter is to 0 (near independence situation), the better the performance of the estimator for all number of exceedances.

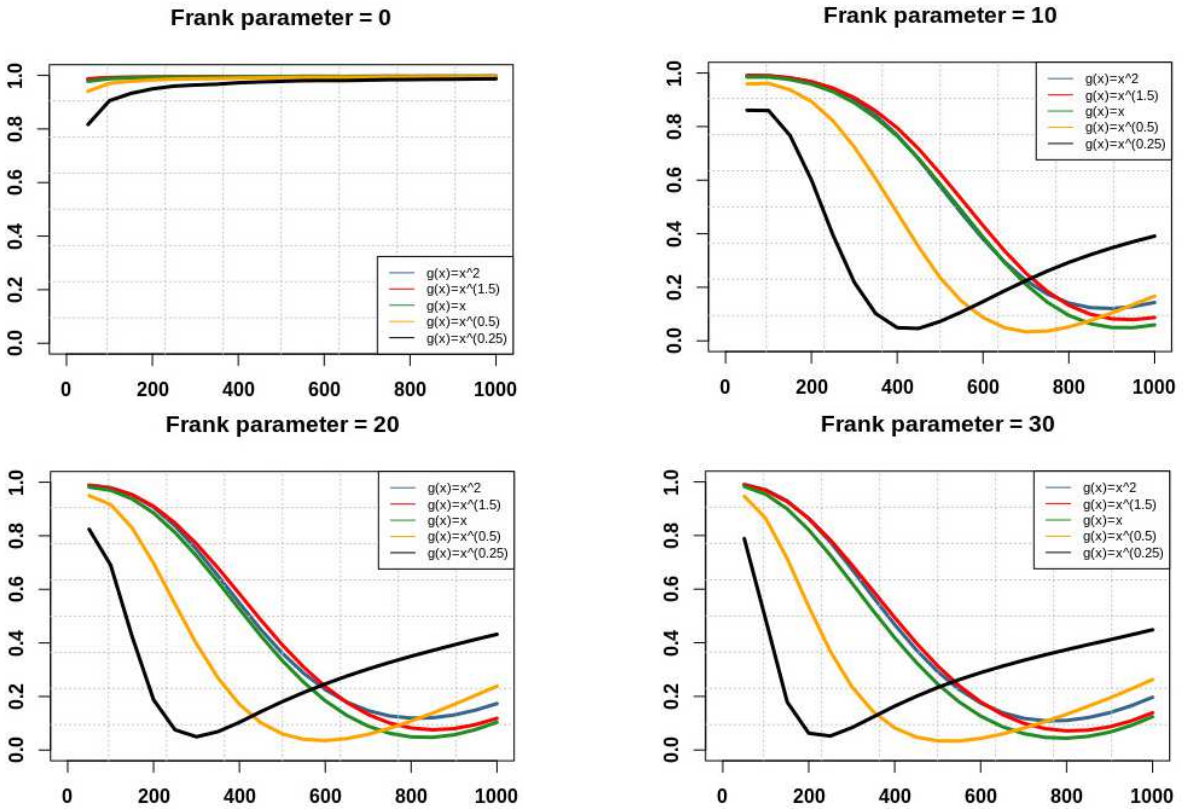


Figure 1: Numerical behaviour of the estimator $\hat{w}(y_n)$ on simulated data, $PC(y_n)$ quality measure as a function of the number of exceedances for several Frank copula parameters.

5 Real data results

In the context of farm income modelling, the impact of various factors on farm yields (expressed in quintals per hectare) is analysed. To this end, our approach is applied to data extracted from the Farm Accountancy Data Network (FADN), an annual database of commercial-sized farm holdings. This dataset of n =9,589 observations contains significant accounting and financial information about French professional farm incomes from 2006 to 2016. The response variable Y is wheat yield and the covariate X includes 7 quantitative variables: year, pesticide, fertilizer, selling prices, region, crop insurance purchased, and insurance claims. The estimator $\hat{w}(y_n)$ is computed for each $y_n = Y_{[n(1-\tau)],n}$ where $\tau \in \{0.01, 0.02, \ldots, 0.99\}$. For the sake of interpretation, we define the conditional correlation between the projected covariate $\hat{w}(y_n)^t X$ and each coordinate $X^{(j)}$ of the covariate as:

$$\rho(\hat{w}(y_n)^t X, X^{(j)} | Y \geq y_n) = \frac{cov(\hat{w}(y_n)^t X, X^{(j)} | Y \geq y_n)}{\sigma(\hat{w}(y_n)^t X | Y \geq y_n)\sigma(X^{(j)} | Y \geq y_n)}.$$

The results are depicted on Figure 2 for the 7 considered covariates. We note that the 1000 largest yields ($Y > y_n = 92$ quintals/hectare) are mainly explained by the purchase of an insurance policy and the claims size. One can also see that the 500 largest yields ($Y > y_n = 98$ quintals/hectare) are still explained by insurance contracts. In contrast, the claims size effect is reduced and a slight pesticide effect appears.

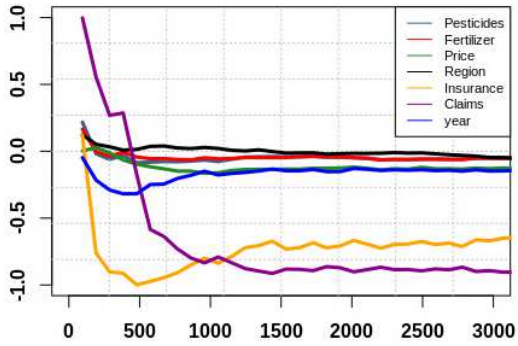


Figure 2: Conditional correlation $\rho(\hat{w}(y_n)^t X, X^{(j)}|Y \geq y_n)$ as a function of the number of exceedances associated with each y_n.

Acknowledgements. This work is supported by the French National Research Agency (ANR) in the framework of the Investissements d'Avenir Program (ANR-15-IDEX-02).

References

Bernard-Michel, C., Gardes, L., & Girard, S. (2008). A note on sliced inverse regression with regularizations. *Biometrics*, 64(3), 982–984.

Bingham, N. H., Goldie, C. M., & Teugels, J. L. (1989). Regular variation (Vol. 27). *Cambridge university press*.

de Haan, L., & Ferreira, A. (2007). Extreme value theory: An introduction. *Springer Science & Business Media*.

Gardes, L., & Girard, S. (2010). Conditional extremes from heavy-tailed distributions: An application to the estimation of extreme rainfall return levels. *Extremes*, 13(2), 177–204.

Li, K. C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414), 316–327.

Nelsen, R. B. (2007). An introduction to copulas. *Springer Science & Business Media*.

Tenenhaus, M. (1999). L'approche PLS. *Revue de Statistique Appliquée*, 47(2), 5–40.

The dynamic latent block model for the co-clustering of evolving binary matrices

Giulia Marchello[1,2], Marco Corneli[2,3] & Charles Bouveyron[2]

[1] *Università degli Studi di Perugia*
[2] *Université Côte d'Azur, Inria, CNRS, Laboratoire J.A. Dieudonné, Maasai team*
[3] *Université Côte d'Azur, Maison de la Simulation et des Interactions (MSI)*
Email: giulia.marchello@inria.fr

Résumé. Nous considérons le problème du co-clustering des matrices binaires qui peuvent évoluer dans le temps et nous introduisons un modèle génératif pour le gérer. Le modèle proposé, appelé dynamic latent block model, étend le modèle des blocs latents binaire classique au cas dynamique. La modélisation de la dynamique en temps continu repose sur un processus de Poisson non homogène, avec une partition latente des intervalles de temps. Nous proposons d'utiliser l'algorithme SEM-Gibbs pour l'inférence du modèle.

Mots-clés. Co-clustering, matrices binaires dynamiques, modèle à blocs latents, algorithme SEM-Gibbs

Abstract. We consider the problem of co-clustering binary matrices that may evolved along the time and we introduce a generative model to handle it. The proposed model, named dynamic latent block model, extend the classical binary latent block model to the dynamic case. The modeling of the dynamic in a continuous time relies on a non-homogeneous Poisson process, with a latent partition of time intervals. We proposed to use the SEM-Gibbs algorithm for model inference.

Keywords. Co-clustering, dynamic binary matrices, latent block model, SEM-Gibbs algorithm

1 Introduction

In many applications, it is now frequent to have to summarize large binary matrices that may evolve along the time. For instance, e-commerce systems record in continuous time all purchases of products made by customers. It is of interest for those companies to cluster both customers and products to better understand the purchasing behaviors. The simultaneous clustering of rows and columns of a matrices is known as a co-clustering problem.

The Dynamic Latent Block Model introduced in the present work is a new and original way to consider, in a dynamic framework, the Latent Block Model [see e.g. [2, 7]] which is a popular generative model for co-clustering. We assume that the number of individuals

(rows) $i = 1, ..., n$ and the number of objects (columns) $j = 1, ..., p$ is fixed during the whole time period taken into account $[0, T]$. Moreover, we consider that the interaction occurring between the individual i and the object j at time $u \in [0, T]$, is represented by the triplet (i, j, u) with $u \leq T$, and it is denoted by $x_{ij}(u)$. Thus, we assume that the number of interactions between individuals and objects follows a non-homogenous Poisson process (NHPP) where the intensity function $\lambda(t)$ only depends on the clusters they belong to [4].

In order to model the continuous time we consider the approach that has been used in Corneli et al. (2018) [3]. In that paper the continuous time is handled by a time partition over $[0, T]$ where the interactions are aggregated on the time intervals of such partition obtaining a sequence of static matrices that allows us to identify the time clusters. Therefore, the basic idea is to partition the time in small windows and to cluster every time period, in this way it is possible to assign every time interval to a time cluster. Intuitively, one time cluster can occur more than once in the temporal line when its peculiar features are repeated after some time. This behaviour, for instance, is typical when the phenomenon of seasonality is observed.

2 The dynamic latent block model

We consider a random matrix $\mathbf{X}(t)$ where the entry $X_{ij}(t)$, with $i = 1, ..., n$, $j = 1, ..., p$ and $t \in [0, T]$, counts the number of interactions between i and j up to time t.

The latent structure of rows and columns of the matrix $\mathbf{x}(t)$ is identified by:

- $\mathbf{z} = (z_{ik}; i = 1, ..., n; k = 1, ..., K)$: represents the clustering of rows into K groups: $\mathcal{A}_1, ..., \mathcal{A}_K$. Row i belongs to cluster $\mathcal{A}_k$ iff $z_{ik} = 1$ and $\mathbf{z_i} = (z_{ik})_k \in (0, 1)^K$ is the group idicator of row i;
- $\mathbf{w} = (w_{j\ell}; j = 1, ..., p; \ell = 1, ..., L)$: represents the clustering of columns into L groups: $\mathcal{B}_1, ..., \mathcal{B}_L$. Coloumn j belongs to cluster B_ℓ iff $w_{j\ell} = 1$ and $\mathbf{w_j} = (w_{j\ell})_\ell \in (0, 1)^L$ is the group idicator of coloumn j;

Moreover, $\mathbf{z}$ and $\mathbf{w}$ are indipendent and they are distributed as follows:

$$p(\mathbf{z}|\gamma) = \prod_{k=1}^{K} \gamma_k^{|\mathcal{A}_k|} \tag{1}$$

where:
$\gamma_k = \mathbb{P}\{z_i = k\}$; $\sum_{k=1}^{K} \gamma_k = 1$ and $|\mathcal{A}_k|$ represents the number of rows in the cluster $\mathcal{A}_k$.

$$p(\mathbf{w}|\rho) = \prod_{\ell=1}^{L} \rho_\ell^{|\mathcal{B}_\ell|} \tag{2}$$

where:

$\rho_\ell = \mathbb{P}\{w_j = \ell\}$; $\sum_{\ell=1}^{L} \rho_\ell = 1$ and $|\mathcal{B}_\ell|$ represent the number of coloums in the cluster $\mathcal{B}_\ell$.

As mentioned previously, a non-homogeneous Poisson process (NHPP) is used to count the interactions between the row i and the column j up to time $t \in [0, T]$, denoted by $X_{ij}(t)$:

$$X_{ij}(t)|z_{ik}w_{j\ell} = 1 \sim \mathcal{P}\left(\int_0^t \lambda_{k\ell}(u)du\right) \tag{3}$$

where $\lambda_{k\ell}(t)$ represents the intensity function that only depends on the considered row cluster k and on the column cluster ℓ. Moreover, $\lambda_{k\ell}(t)$ has to be positive and integrable on the time interval $[0, T]$.

As in Corneli (2016) [4], we discretize the continous time interval [0,T] in U subintervals, where $I_u = [t_{u-1} - t_u[$. , with:

$$0 = t_0 < t_1 < \cdots < t_U = T.$$

The number of interactions between i and j on the considered time partition I_u is summarized by X_{iju} and is defined as:

$$X_{iju} := X_{ij}(t_u) - X_{ij}(t_{u-1}), \forall (i, j, u).$$

We introduce a tensor $X = \{X_{iju}\}_{iju}$ with dimensionality $N \times P \times U$. As previously described in a more theoretical way, each time interval $I_1, ..., I_U$ is assigned to a hidden time cluster $\mathcal{C}_1, .., \mathcal{C}_c$. To model the membership to time clusters, a new latent variable $\mathbf{s}$ has to be introduced, in particular $\mathbf{s}_u = c$ if and only if the time interval I_u belongs to the time cluster $\mathcal{C}_c$. Furthermore, we assume that $\mathbf{s}$ follows a multinomial distribution:

$$p(\mathbf{s}|\delta) = \prod_{c=1}^{C} \delta_c^{|\mathcal{C}_c|}, \tag{4}$$

where:

$\delta_c = \mathbb{P}\{s_u = c\}$; $\sum_{c=1}^{C} \delta_c = 1$ and $|\mathcal{C}_c|$ represents the number of time intervals in the cluster $\mathcal{C}_c$. The originality of this work lies in the fact that a dependence is assumed between the two variables: X_{iju} and S_u.

Once these additional assumpions have been made, we can rewrite Eq; (3) considering that the intensity functions are stepwise constant on each time cluster $\mathcal{C}_c$. Hence, now we are considering a non- homogeneus Poisson process (NHPP) with parmeter $\lambda_{k\ell c}$ and Δ_u :

$$X_{iju}|z_{ik}w_{j\ell}S_{uc} = 1 \sim \mathcal{P}(\lambda_{k\ell c}\Delta_u) \tag{5}$$

where Δ_u indicates the length of the interval I_u that is usually constant, $\Delta_u = \Delta$. Moreover, as pointed out in Corneli et al. (2018) [3], one can set $\Delta_u = 1$ without loss of generality, indeed if, for instance, we have a dataset divided in week, the time interval Δ_u could be set to one week. It is finally possible to introduce a tensor $\Lambda = \{\lambda_{k\ell c}\}_{k\ell c}$ of dimension $K \times L \times C$.

3 The likelihood

Since the increments of a Poisson process are independent it holds that:

$$p(X_{iju}|z_{ik}w_{jl}s_{uc} = 1, \lambda_{k\ell}) = \prod_{u=1}^{U} \left(\frac{(\lambda_{klc})^{X_{iju}}}{X_{iju}!} \exp\left(-\lambda_{klc}\right) \right) \tag{6}$$

Therefore, we can introduce the $K \times L \times C$ tensor of order 3, $\boldsymbol{\lambda}$, identified by the triplet (i, j, u) and whose elements are denoted as $\pi_{k\ell c}$. At this point it is possible to write the complete data likelihood of the model, identified by the following equation:

$$p(X, z, w, s|\gamma, \rho, \delta, \lambda) = p(z|\gamma) \cdot p(w|\rho) \cdot p(s|\delta) \cdot p(X|z, w, s, \lambda) \tag{7}$$

Looking at the right hand side of the Eq. (7) we notice that $p(z|\gamma)$, $p(w|\rho)$ and $p(s|\delta)$ have been defined in the previous section respectively by the Eqs. (1), (2) and (4). While the joint distribution of $\mathbf{X}$, given $\boldsymbol{z}$, $\boldsymbol{w}$, and $\boldsymbol{s}$, can be easily obtained as a generalization of the Eq. (6):

$$p(X|z, w, s, \lambda) = \prod_{k=1}^{K}\prod_{\ell=1}^{L}\prod_{c=1}^{C} \left(\frac{(\lambda_{k\ell c})^{R_{k\ell c}}}{P_{k\ell c}} exp\left(-|\mathcal{A}_k||\mathcal{B}_\ell||\mathcal{C}_c|\lambda_{k\ell c}\right) \right) \tag{8}$$

where:

$$R_{k\ell c} = \sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{u=1}^{U} z_{ik}w_{j\ell}s_{uc}X_{iju}$$

$$P_{k\ell c} = \prod_{i=1}^{n}\prod_{j=1}^{p}\prod_{u=1}^{U} (z_{ik}w_{j\ell}s_{uc}X_{iju})!$$

Denoted as θ the set of the model parameters: $\theta = (\gamma, \rho, \delta, \lambda)$. The complete data log-likelihood can be defined as

$$\ell(\theta|z, w, S, X) = \sum_{z}\sum_{w}\sum_{c} \log p(X, z, w, S|\theta) \tag{9}$$

4 Model inference

As usual, we look for a way to maximize the log-likelihood in order to obtain the estimation of $\boldsymbol{\theta}$. In the co-clustering case, the EM algorithm is computationally infeasible [6], the idea is to use a stochastic version of it, known as SEM-Gibbs, proposed by Keribin C. (2010) [9] and exploited, for instance, by Bouveyron et al. (2017) [1] in the Functional Latent Block Model . Thanks to the Gibbs Sampling, in the SE step a partition for $\boldsymbol{z}$, $\boldsymbol{w}$ and $\boldsymbol{s}$ is generated without computing the joint distribution. The algorithm starts with initial values for the parameter set $\boldsymbol{\theta}^{(0)}$, the column clusters $\boldsymbol{w}^{(0)}$ and the time clusters $s^{(0)}$. Regarding the burn-in period, after a certain number of iterations of the algorithm, we can obtain the final parameters estimation by computing the mean of the sampled distribution. The optimal values for $\boldsymbol{z}$, $\boldsymbol{w}$ and $\boldsymbol{s}$ are estimated by the mode of their sample distributions.

5 Conclusion and further work

We introduced in this short paper a generative model for co-clustering dynamic binary matrices. We proposed to use the SEM-Gibbs algorithm for model inference. Regarding further work, we are currently working on an implementation of this algorithm and we plan to apply it on real-world data from Amazon recording fine food purchases over 8 years.

References

[1] Charles Bouveyron, Laurent Bozzi, Julien Jacques, and François-Xavier Jollois. The functional latent block model for the co-clustering of electricity consumption curves. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67(4):897–915, 2018.

[2] Vincent Brault and Mahendra Mariadassou. Co-clustering through latent bloc model: A review. 2015.

[3] Marco Corneli, Charles Bouveyron, Pierre Latouche, and Fabrice Rossi. The dynamic stochastic topic block model for dynamic networks with textual edges. *Statistics and Computing*, 2018.

[4] Marco Corneli, Pierre Latouche, and Fabrice Rossi. Block modelling in dynamic networks with non-homogeneous poisson processes and exact icl. *Social Network Analysis and Mining*, 6(1):55, 2016.

[5] Gérard Govaert and Mohamed Nadif. Block clustering with bernoulli mixture models: Comparison of different approaches. *Computational Statistics & Data Analysis*, 52(6):3233–3245, 2008.

[6] Gérard Govaert and Mohamed Nadif. *Co-clustering: models, algorithms and applications*. John Wiley & Sons, 2013.

[7] Christine Keribin, Vincent Brault, Gilles Celeux, Gérard Govaert, et al. Model selection for the binary latent block model. In *Proceedings of COMPSTAT*, volume 2012, 2012.

[8] Christine Keribin, Gilles Celeux, and Valérie Robert. The Latent Block Model: a useful model for high dimensional data. In *ISI 2017 - 61st world statistics congress*, pages 1–6, Marrakech, Morocco, July 2017.

[9] Christine Keribin, Gérard Govaert, and Gilles Celeux. Estimation d'un modèle à blocs latents par l'algorithme sem. 2010.

Mélange de Segmentations

Vincent Brault [1] & Émilie Devijver[2] & Charlotte Laclau[3]

[1] *Univ. Grenoble Alpes, CNRS, Grenoble INP*, LJK, 38000 Grenoble, France.*
vincent.brault@univ-grenoble-alpes.fr
[2] *CNRS, Univ. Grenoble Alpes, Grenoble INP*, LIG, 38000 Grenoble, France.*
emilie.devijver@univ-grenoble-alpes.fr
[3] *Univ Lyon, UJM-Saint-Etienne, Laboratoire Hubert Curien*
charlotte.laclau@univ-st-etienne.fr

Résumé. Lorsque les observations d'un phénomène semblent provenir de plusieurs lois différentes, il existe deux approches dans la littérature : soit l'ordre des observations a un sens (comme dans le cas de séries temporelles) et nous chercherons dans ce cas les moments de ruptures séparant les lois (Carlstein et al., 1994), soit il n'en a pas et nous utiliserons les modèles de mélange (McLachlan, 1982). Dans chacun des cas, les procédures sont alors assez différentes. Dans le cas de tableaux, il est possible de chercher des comportements différents à la fois sur les lignes et les colonnes et, là encore, il existe deux techniques suivant si l'ordre des lignes et des colonnes a un sens (Brault et al., 2017, 2018) ou pas (Govaert et Nadif, 2003).

Toutefois, et à notre connaissance, il n'existe pas encore de procédure lorsque l'ordre sur les colonnes a un sens mais pas celui sur les lignes (ou inversement). Pour l'instant, les modèles utilisés considèrent les colonnes sans ordre en espérant que les groupes formés à la fin soient connexes et/ou cohérents (voir par exemple Corneli et al., 2019).

Dans ce travail, nous tentons de concilier les outils des deux communautés pour étudier l'apport de bien considérer l'ordre ou non. Nous présenterons différentes procédures issues du croisement de ces communautés et comparerons les résultats avec les procédures faisant le choix de ne pas prendre en compte l'ordre.

Mots-clés. Modèle de mélanges, Segmentation

Abstract. In the literature, two different approaches exist when one can assume that observations of a phenomenon come from different distributions: (1) if the order of these observations is meaningful, as it is the case for time series for instance, then we search for change points to differentiate the distributions (Carlstein et al., 1994); (2) if the order of the observations is not relevant then mixture models (McLachlan, 1982) are a suitable approach. Furthermore, in the case of data matrices one can also search for distinct patterns between the rows and the columns, and once again, two techniques can be used depending on whether the order of the rows and/or the columns is relevant (Brault et al., 2017, 2018) or not (Govaert et Nadif, 2003). However, to the best of our knowledge, there is no existing procedure able to handle the case where the order is only important

*Institute of Engineering Univ. Grenoble Alpes

for one of the two dimensions (i.e. either the rows or the columns). Traditionally, the models introduced in this context are considering the columns without constraining the order of the obtained group, hoping that they are connected (see Corneli et al., 2019).

In this work, we study the possibility of a hybrid approach between the aforementioned communities of research. We will introduce different procedures, where the order can be imposed only on the rows or on the columns, and will compare to previous works that were not taking order into account.

Keywords. Mixture Model, Change Point

1 Introduction

Dans cet exposé, nous nous intéressons au cas où nous observons plusieurs courbes (par exemple, des consommations électriques) issues de plusieurs phénomènes ayant des ruptures. Sur la figure 1, nous avons représenté 1000 courbes issues de deux phénomènes différents (schématisés ici par les couleurs noires et rouges) ainsi que les emplacements des ruptures pour chaque groupe (en traits pointillés bleus sur les figures centrales pour le groupe noir et les figures de droites pour le groupe rouge).

Nous pouvons observer que les groupes n'ont pas forcément le même nombre de ruptures.

Dans cet exposé, nous proposons une modélisation de ce phénomène et deux algorithmes d'estimations des paramètres.

2 Modèle

Dans cet article, nous nous intéressons aux mélanges de segmentations. Pour cela, nous supposons qu'il existe n courbes $(\boldsymbol{Y}_i)_{1\leq i\leq j}$ p dimensionnelles de longueurs d appartenant à K groupes de façon indépendante et avec une probabilité π_k. Pour chaque groupe k, nous supposons qu'il existe L_k ruptures $0 = T_{k,0} < T_{k,1} < \cdots < T_{k,L_k} < T_{k,L_k+1} = d$, $L_k + 1$ moyennes $\left(\boldsymbol{\mu}_{k,\ell}\right)_{1\leq\ell\leq L_k+1} \in \mathbb{R}^{(L_k+1)\times p}$ et $L_k + 1$ variances $\left(\sigma^2_{k,\ell}\right)_{1\leq\ell\leq L_k+1}$ de telle sorte que pour tout $k \in \{1,\ldots,K\}$, pour tout $\ell \in \{0,\ldots,L_k\}$ et pour tout $j \in \{T_{k,\ell}+1, T_{k,\ell}+2,\ldots,T_{k,\ell+1}\}$, nous ayons :

$$\boldsymbol{Y}_{i,j} \overset{\text{ind}}{\sim} \mathcal{N}\left(\boldsymbol{\mu}_{k,\ell}, \sigma_{k,\ell}\mathbb{I}_r\right)$$

où $\mathbb{I}_r$ est la matrice identité de dimension $r \times r$. Avec ces conditions et en notant $\boldsymbol{T}$ l'ensemble des ruptures et $\boldsymbol{\theta}$ l'ensembles des paramètres, nous obtenons donc la vraisemblance suivante :

$$p\left(\boldsymbol{Y}; K, \boldsymbol{T}, \boldsymbol{\theta}\right) = \prod_{i=1}^{n}\sum_{k=1}^{K}\pi_k\prod_{\ell=0}^{L_k}\prod_{j=T_{k,\ell}+1}^{T_{k,\ell+1}}\prod_{r=1}^{p}\left[\frac{1}{\sqrt{2\pi\sigma_{k,\ell}}}e^{-\frac{1}{2\sigma_{k,\ell}}\left(Y_{k,\ell,r}-\mu_{k,\ell,r}\right)^2}\right].$$

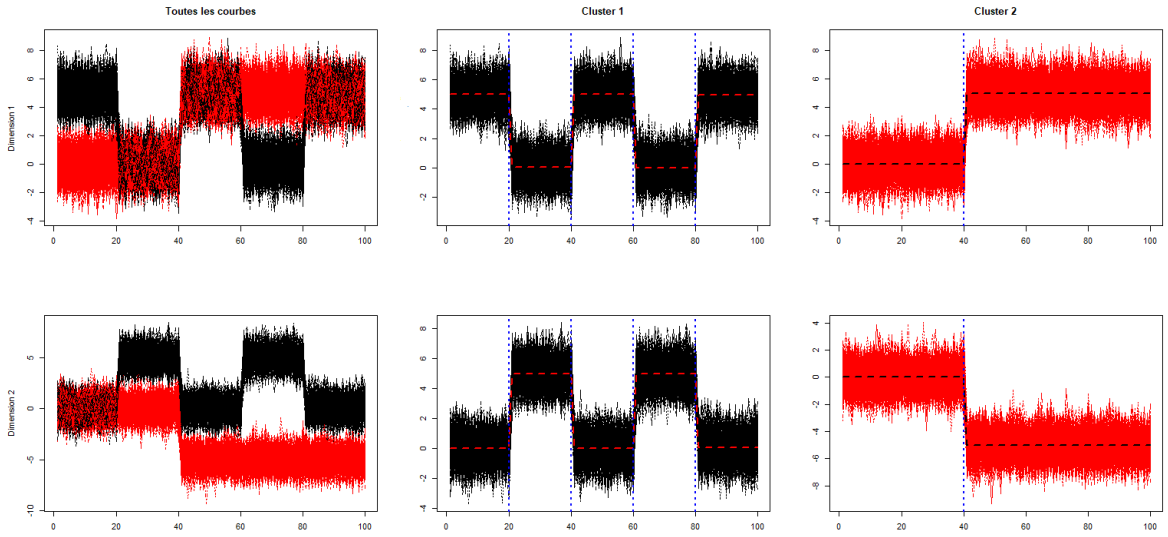


Figure 1: Représentation de $n = 1000$ courbes de dimension $d = 2$ et de longueurs $d = 100$ (à gauche) issues de deux groupes différents (le noir et le rouge ; groupes estimés par l'algorithme présenté en section 3). Au centre, nous avons conservé uniquement les courbes du groupe noir et, à droite, les courbes issues du groupe rouge ; les traits pointillés noirs et rouges correspondent aux moyennes estimées des plateaux et les traits pointillés bleus symbolisent les instants de ruptures estimés.

Ainsi, nous reconnaissons la partie propre aux modèles de mélanges et celle se rapportant à la segmentation. En plus des paramètres K, $\boldsymbol{T}$ et $\boldsymbol{\theta}$, nous cherchons également à estimer la partition $\boldsymbol{z}$ que nous noterons indifféremment sous sa forme binaire ($z_{i,k} = 1$ si et seulement la courbe i appartient à la classe k) et sa forme vectorielle ($z_i = k$ si et seulement si la courbe i appartient à la classe k).

3 Estimation

Pour l'estimation, nous supposons connaître le nombre de groupes K et les nombres de ruptures $L_1, \ldots, L_K$ et nous allons alors utiliser l'algorithme *Espérance Maximisation* proposé par Dempster et al. (1977) dont le but est de maximiser l'espérance conditionnelle de la logvraisemblance complète étant donné des paramètres $\boldsymbol{\theta}^{(c)}$ et des ruptures $\boldsymbol{T}^{(c)}$ fixés :

$$\mathbb{E}_{\boldsymbol{z}|\boldsymbol{T}^{(c)},\boldsymbol{\theta}^{(c)}} \left[\log p\left(\boldsymbol{Y}, \boldsymbol{z}; \boldsymbol{T}, \boldsymbol{\theta}\right)\right].$$

L'algorithme procède en deux étapes consistant à calculer l'espérance (étape E) étant donnés $\boldsymbol{T}^{(c)}$ et $\boldsymbol{\theta}^{(c)}$ et à trouver les paramètres $\boldsymbol{T}$ et $\boldsymbol{\theta}$ la maximisant puis de recommencer

jusqu'à une convergence locale du critère.

Si l'étape E ne pose pas de difficultés techniques, nous avons un problème pour la maximisation des instants de ruptures $\boldsymbol{T}$ et des paramètres $\boldsymbol{\theta}$. En effet, l'algorithme classique de segmentation, appelé algorithme de *programmation dynamique*, possède une complexité naïve de $\mathcal{O}(nd^2r)$ dans notre cas qui doit être appliqué à chaque itération pour chaque groupe. Afin de contourner ce problème de vitesse, certains auteurs comme Brault et al. (2017) proposent de transformer le problème en une estimation équivalente d'une régression linéaire dont le paramètre d'intérêt serait sparse et donc d'utiliser les procédures LASSO (Least Absolute Shrinkage and Selection Operator). L'inconvénient de ces méthodes est la gestion du paramètre de régularisation de la méthode LASSO et la multiplication du nombre de ruptures estimées.

4 Conclusions

Dans cet exposé, nous expliciterons et comparerons les deux méthodes (*Programmation Dynamique* et *LASSO*) sur leurs qualités d'estimation des ruptures et leurs rapidités.

Bibliographie

V. Brault, J. Chiquet, et C. Lévy-Leduc. Efficient block boundaries estimation in block-wise constant matrices: An application to hic data. *Electron. J. Statist.*, 11(1):1570-1599, 2017. ISSN 1935-7524. doi: 10.1214/17-EJS1270.

V. Brault, S. Ouadah, L. Sansonnet, et C. Lévy-Leduc. Nonparametric multiple change-point estimation for analyzing large hi-c data matrices. *Journal of Multivariate Analysis*, 165: 143 - 165, 2018. ISSN 0047-259X. doi: `https://doi.org/10.1016/j.jmva.2017.12.005`. URL `http://www.sciencedirect.com/science/article/pii/S0047259X17307753`.

E. Carlstein, H. Müller, et D. Siegmund. Change-point problems. institute of mathematical statistics lecture notes-monograph series 23. *IMS, Hayward, CA. MR1477909*, 1994.

M. Corneli, C. Bouveyron, P. Latouche, et F. Rossi. The dynamic stochastic topic block model for dynamic networks with textual edges. *Statistics and Computing*, 29(4), 677-695, 2019.

A. P. Dempster, N. M. Laird, et D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1), 1-22, 1977.

G. Govaert et M. Nadif. Clustering with block mixture models. *Pattern Recognition*, 36(2):463-473, 2003.

G. J. McLachlan. The classification and mixture maximum likelihood approaches to cluster analysis. *Handbook of statistics*, 2:199-208, 1982.

Clustering on multilayer graphs with missing values

Guillaume Braun [1] & Christophe Biernacki [2] & Hemant Tyagi [3]

[1] *Inria, Université de Lille, CNRS, Laboratoire de mathématiques Painlevé 59650 Villeneuve d'Ascq, France, guillaume.braun@inria.fr*
[2] *Inria, Université de Lille, CNRS, Laboratoire de mathématiques Painlevé 59650 Villeneuve d'Ascq, France, christophe.biernacki@inria.fr*
[3] *Inria, Université de Lille, CNRS, Laboratoire de mathématiques Painlevé 59650 Villeneuve d'Ascq, France, hemant.tyagi@inria.fr*

Résumé. La classification sur les graphes à couches multiples rencontre un intérêt croissant depuis une décennie en raison de leurs nombreux champs d'applications. Différentes méthodes ont été proposées, mais elles reposent toutes sur l'hypothèse que la totalité des interactions entre les noeuds du réseau sont observées. Nous proposons un cadre statistique afin d'étudier le cas souvent plus réaliste dans lequel des noeuds ne sont pas observés sur certaines couches. Une méthode spécifique permettant d'estimer les paramètres du modèle et d'imputer les valeurs d'interactions manquantes est également proposée.

Mots-clés. Graphes à couches multiples, Classification, Stochastic Block Model, données manquantes.

Abstract. Multilayer graphs clustering have gained increasing interest this last decade due to numerous applications in various fields. Several clustering methods have been proposed, but they rely all on the assumption that the network is fully observed. We propose a statistical framework to handle nodes that are missing on some layers as well as a method to estimate the model parameters and to impute missing edge values.

Keywords. Multilayer graph, Clustering, Stochastic Block Model, missing data.

1 Introduction

Simple graphs are often used to model relationships between different agents: each agent is associated to a node and the link between two agents is represented by an edge between the corresponding nodes. Weighted and directed edges can be used to express the intensity and the direction of a link.

However, relationships between agents can have multiple aspects that are better represented by multilayer graphs where each layer corresponds to one aspect of the relationship. For example, the social network of a person involves different relationship types such as emails exchanges, telephone calls, offline personal interactions, professional links and so

on. All these layers are dependent to some extent since a modification in one layer can impact the others. For example a change of employer could have repercussions on both online and offline networks. A layer might also correspond to a snapshot of the entire network at a sampled time instant. These networks are sometimes referred to as temporal or dynamic networks to emphasise that they are time dependent.

Because of their broad field of applications including Biology, Sociology, Genetics and Ecology, multilayer graph analysis has gained increasing attention in the last decade, especially for clustering purposes (see [Kim and Lee, 2015] for a survey). Although various clustering methods have been proposed, they rely on the common assumption that the network is fully observed. However missing values often occur in practice. For instance, various missing schemes for edges have been already studied for unilayer graphs in [Tabouy et al., 2019].

We propose a new sampling design to deal with missing nodes in some layers. Until now only missing edges have been considered in the literature. In the unilayer case, since there is no information about missing nodes, it is impossible to make an inference concerning their links with the other nodes. But in the multilayer case, a missing node in one layer can be present in other layers and the additional information provided by the layers where the node is observed can be used. Moreover, efficiently clustering missing nodes in some layer would have interesting applications in link prediction.

Our originality is to propose some algorithms for estimating the model parameters and the latent partition when some nodes are missing between layers. Our proposal fully relies on the existing Multi-Layer Stochastic Block Model [Lei et al., 2019] embedded in an explicit missing data mechanism.

In Section 2 we introduce notations, present the classical Stochastic Block Model for unilayer graphs, its extension to the multilayer case and also some non-response mechanisms. Section 3 describes the estimation procedure based on approximations of the maximum likelihood estimator. Finally, Section 4 exposes our future work directions which include ongoing numerical experiments.

2 A model for multilayer networks with missing nodes

2.1 Multilayer networks notations

A network can be recorded by an adjacency matrix. Thus a multilayer network can be represented by a collection of $l = 1, \dots, L$ adjacency matrices: each layer l is associated to an adjacency matrix $Y^l = (Y^l_{ij})_{1 \leq i,j \leq n}$ and the whole multilayer network is denoted by $Y = (Y^l)_{1 \leq l \leq L}$. A multilayer graph is said to be *pillar*, if the set of nodes $\mathcal{N} = \{1, \dots n\}$ is the same in each layer.

We restrict our study to undirected graphs that don't have loops. This implies that all the adjacency matrices are symmetric and have zero diagonal.

Let $R_{ij}^l = 1$ if the edge between the nodes i and j in the layer l is observed and 0 else. We denote $\mathcal{D}^{o,l} = \{(i,j) \text{ such that } R_{ij}^l = 1\}$ the indexes of the observed edges in layer l, $Y^{o,l} = (Y_{ij}^l)_{(i,j)\in\mathcal{D}^{o,l}}$ the corresponding values of the observed edges in layer l and $Y^o = (Y^{o,l})_{1\leq l\leq L}$. Symmetrically, we denote $\mathcal{D}^{m,l} = \{(i,j) \text{ such that } R_{ij}^l = 0\}$ the set of missing edges in layer l, $Y^{m,l} = (Y_{ij}^l)_{(i,j)\in\mathcal{D}^{m,l}}$ the corresponding values of the observed edges in layer l and $Y^m = (Y^{m,l})_{1\leq l\leq L}$. We say that a node i is not observed in the layer l if any of the edges between i and the other nodes are unobserved in this layer. We denote by $\mathcal{N}^{o,l}$ the corresponding set of observed nodes. The set of nodes that appear in at least one layer is denoted by $\mathcal{N}^o = \cup_l \mathcal{N}^{o,l}$.

2.2 The Multi-Layer Stochastic Block Model (ML-SBM)

The ML-SBM is an extension to multilayer networks of the Stochastic Block Model (SBM) devoted to unilayer networks. In the ML-SBM each layer is generated from a SBM, with eventually different connectivity parameters, and every node has the same block membership in every layer. A first introduction of this model can be founded in [Lei et al., 2019].

Let $\mathcal{K} = \{1, \ldots, K\}$ the different block (or cluster) numbers. The block-membership of the node i is encoded by Z_i, and $(Z_i)_{i\in\mathcal{N}}$ are i.i.d. random variables distributed over $\mathcal{K}$, independently of the layer number. The whole partition is now denoted by $Z = (Z_i)_{i\in\mathcal{N}}$. The distribution of Z_i is completely determined by $\mathbb{P}(Z_i = k) = \alpha_k$ for $k = 1 \ldots K$. It is sometimes convenient to associate Z_i to a binary vector $(Z_{i1}, \ldots, Z_{iK})$ where $Z_{ik} = 1$ if $Z_i = k$ and $Z_{ik'} = 0$ for $k' \neq k$. The realization of the random variable Z_i is denoted in lowercase by z_i, and the whole mixing proportion parameter is denoted by $\alpha = (\alpha_1, \ldots, \alpha_K)$.

Considering now the adjacency matrices Y^l $(l = 1, \ldots, L)$, ML-SBM considers that there is no loops (for all i, $Y_{ii}^l = 0$), the adjacency matrix is symmetric ($Y_{ij} = Y_{ji}$ for all i and j), and conditional distribution of edges depends on the layer. This latter is given by

$$Y_{ij}^l | (Z_i = k, Z_j = k') \overset{\text{i.i.d.}}{\sim} \mathcal{B}(\pi_{kk'}^l), \forall i < j$$

where $\mathcal{B}(p)$ denotes a Bernoulli distribution of parameter p and $\pi_{kk'}^l$ is the probability for node in community k to be linked with a node in community k'. Let finally denotes by $\pi^l = (\pi_{kk'}^l)_{1\leq k,k'\leq K}$ a symmetric matrix of probabilities and also $\theta = (\alpha, \pi^1, \ldots, \pi^L)$ the whole mixture model parameter.

2.3 Missing values

The mechanisms for missingness can be traditionally classified as MCAR (missing completely at random), MAR (missing at random) or MNAR (missing not at random). We assume that nodes are MAR and that the missingness mechanism is ignorable. Such a

classical assumption avoids to define (and then estimate) the related distribution of the R_{ij}^l's since it will have no consequence on the estimation of the parameter of interest in the clustering context, namely θ, [Little and Rubin, 2002].

3 Maximum likelihood estimation (MLE)

3.1 Observed and complete log-likelihood

The complete data log-likelihood of the model is given by

$$\mathcal{L}_c(\theta; z, Y) = \sum_i \log(\alpha_{z_i}) + \sum_{\substack{i,j,l \\ i<j}} (Y_{ij}^l \log \ \pi_{z_i,z_j}^l + (1 - Y_{ij}^l) \log(1 - \pi_{z_i,z_j}^l)).$$

However the block membership is usually not observed and some edge values are also missing. So we rather use the observed log-likelihood $\mathcal{L}(\theta; Y^o)$ obtained by integration over the latent variables Z and the missing edges Y^m, given by

$$\mathcal{L}(\theta; Y^o) = \log \left(\sum_z \int_{Y^m} \exp(\mathcal{L}_c(\theta; z, Y)) dY^m \right).$$

The associated MLE is computationally intractable in general, even with a classical EM algorithm [Dempster et al., 1977]. But there exists approximate EM algorithms and we explore one of them below. Once we have an estimate $\hat{\theta}$ for θ, the partition and the missing edges can be respectively estimated by

$$\hat{Z} = \arg\max_z \mathbb{P}_{\hat{\theta}}(z|Y^o) \quad \text{and} \quad \hat{Y}^m = \arg\max_{Y^m} \mathbb{P}_{\hat{\theta}}(Y^m|Y^o). \tag{1}$$

3.2 Variational EM (VEM)

The VEM algorithm has been developped in [Daudin et al., 2008] for avoiding computational difficulties implied by the EM algorithm when estimating the SBM. Instead of trying to maximize $\mathcal{L}(\theta; Y^o)$, VEM intends to maximize a lower bound $J_{\theta,\tau}(Y^o) := \mathcal{L}(\theta; Y^o) - KL(\mathbb{P}_\tau(Z)||\, \mathbb{P}_\theta(Z|Y^o)) = \mathbb{E}_{\mathbb{P}_\tau(Z)}(\log(\mathbb{P}_\theta(Z, Y^o)))$ of this quantity. If we consider the set of all possible distributions $\mathbb{P}_\tau$ for Z, the maximum is attained when $\mathbb{P}_\tau(Z) = \mathbb{P}_\theta(Z|Y^o)$ but this last quantity is difficult to compute. So we restrict the set in which $\mathbb{P}_\tau$ belongs to. We are only looking on probability that can be factorized such that $\mathbb{P}_\tau(Z) = \prod_i \mathcal{M}_{\tau_i}(Z_i)$ where $\mathcal{M}_{\tau_i}$ is the multinomial distribution with parameters $\tau_i = (\tau_{i1}, \ldots, \tau_{iK})$ and $\tau = (\tau_1, \ldots, \tau_n)$.

After having chosen an initial parameter $\theta^{(0)}$, we construct iteratively a sequence $(\theta_{t\geq 1}^{(t)}$ by repeating the following steps until $||\theta^{(t+1)} - \theta^{(t)}|| < \epsilon$, where ϵ is a parameter fixed by the user.

- **VE step:** compute $\tau^{(t)} := \arg\max_\tau J_{\theta^{(t)},\tau}(Y^o)$;
- **M step:** update $\theta^{(t)}$ by $\theta^{(t+1)} := \arg\max_\theta J_{\theta,\tau^{(t)}}(Y^o)$.

These two maximization problems are solved straightforwardly:

1. The variational parameters $\tau^{(t)}$ maximizing $J_{\theta^{(t)},\tau}(Y^o)$ when $\theta^{(t)}$ is held fixed are obtained with the following fixed point relation:

$$\tau_{ik}^{(t)} \propto \alpha_k^{(t)} \prod_{\substack{l \leq L \\ (i,j)\in\mathcal{D}^{o,l} \\ i<j}} \prod_{k'=1}^{K} b(Y_{ij}^l; \pi_{kk'}^{l(t)})^{\tau_{jk'}^{(t)}},$$

   where $b(y,\pi) = \pi^y(1-\pi)^{1-y}$.

2. The parameters $\theta^{(t+1)}$ maximizing $J_{\theta,\tau^{(t)}}(Y^o)$ when $\tau^{(t)}$ is held fixed are:

$$\alpha_k^{(t+1)} = \frac{\sum_{i\in\mathcal{N}^o} \tau_{ik}^{(t)}}{|\mathcal{N}^o|}, \qquad \pi_{kk'}^{l(t)} = \frac{\sum_{\substack{(i,j)\in\mathcal{D}^{o,l} \\ i<j}} \tau_{ik}^{(t)}\tau_{jk'}^{(t)} Y_{ij}^l}{\sum_{\substack{(i,j)\in\mathcal{D}^{o,l} \\ i<j}} \tau_{ik}^{(t)}\tau_{jk'}^{(t)}}.$$

3.3 Missing value estimation: partition and edges

The VEM algorithm provides an estimation of the ML-SBM parameters that can be used to solve the maximization problems (1) in order to get an estimator for the partition and missing edges. Unfortunately, these optimization problems are computationally intractable. We propose an approximated solution based on Gibbs sampling. Let's describe the algorithm.

1. **Initialisation.** Let $z^{(0)} = (z_1^{(0)}, \ldots, z_n^{(0)})$ and $Y^{m^{(0)}}$ random initial choices for the partition and missing values.
2. **Sampling partition.** Generate the partition with:

$$\mathbb{P}_{\hat\theta}(z_i^{(t)} = k | Y^o, Y^{m,(t-1)}, z^{-i,(t-1)}) \propto \hat\alpha_k \times \prod_{l,j,k'} b(y_{ij}^{l'}; \hat\pi_{kk'}^l)^{z_{jk'}^{(t-1)}}$$

   where $z^{-i,(t-1)}$ correspond to $z^{(t-1)}$ with the ith component removed.
3. **Missing values imputation.** Generate the missing data $Y^{l^{m,(t)}} = (y_{ij}^{l(t)})_{\substack{l\in[L] \\ i,j\in Y^m \\ i<j}}$ according to:

$$\mathbb{P}_{\hat\theta}(y_{ij}^{l(t)} | Y^o, z^{(t)}) = \prod_{k,k'} b(y_{ij}^{l(t)}; \hat\pi_{kk'}^l)^{z_{ik}^{(t)} z_{jk'}^{(t)}}.$$

Gibbs algorithm is iterated several times. The samples obtained during the burn-in period are disregarded. Finally, the final partition and missing observations are estimated using the mode of their marginal sampled distribution.

4 Directions for future works

We intend to perform numerical experiments on simulated and real data. We are also aiming to extend the proposed model and develop a clustering algorithm for multilayer graphs based on spectral methods because these methods have good performance on unilayer graphs and are more scalable than those based on maximum likelihood approximations.

References

[Daudin et al., 2008] Daudin, J. J., Picard, F., and Robin, S. (2008). A mixture model for random graphs. *Statistics and Computing*, 18(2):173–183.

[Dempster et al., 1977] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38.

[Kim and Lee, 2015] Kim, J. and Lee, J.-G. (2015). Community detection in multi-layer graphs: A survey. *ACM SIGMOD Record*, 44:37–48.

[Lei et al., 2019] Lei, J., Chen, K., and Lynch, B. (2019). Consistent community detection in multi-layer network data. *Biometrika*, 107(1):61–73.

[Little and Rubin, 2002] Little, R. and Rubin, D. (2002). *Statistical analysis with missing data*. Wiley series in probability and mathematical statistics. Probability and mathematical statistics. Wiley.

[Tabouy et al., 2019] Tabouy, T., Barbillon, P., and Chiquet, J. (2019). Variational inference for stochastic block models from sampled data. *Journal of the American Statistical Association*, pages 1–20.

Profondeur de Tukey: Ensembles de niveau empiriques et théoriques

Victor-Emmanuel Brunel [1]

[1] *CREST-ENSAE*
5, Av. Le Chatelier, 91120 Palaiseau
victor.emmanuel.brunel@ensae.fr

Résumé. La profondeur de Tukey est une notion statistique qui suscite beaucoup d'intérêt en statistique multivariée, car elle permet d'ordonner les données en généralisant, d'une certaine manière, la notion de quantile. Etant donnés une loi de probabilité et des données i.i.d. suivant cette loi, nous étudions la convergence des ensembles de niveau de la profondeur empirique vers les ensembles de niveau théoriques. Ces ensembles de niveau peuvent être interprétés comme des quantiles multivariés. Sous des hypothèses raisonnables, nous montrons la concentration des ensembles de niveau empiriques à la vitesse paramétrique, en utilisant des outils de géométrie convexe, de processus empiriques et de programmation linéaire semi-infinie.

Mots-clés. Pronfondeur de Tukey, ensembles de niveau, quantiles multivariés, convexité, fonction de support, distance de Hausdorff, programmation linéaire semi-infinie.

Abstract. Tukey depth has attracted much attention in multivariate statistics, because it allows to order the data, while extending the notion of univariate quantiles. Given a probability distribution and i.i.d. data from this distribution, we study the convergence of the level sets of the empirical depth towards those of the population depth. These level sets can be interpreted as multivariate quantiles. Under reasonable assumptions, we prove that the empirical level sets concentrate at the parametric rate, borrowing some tools from convex geometry, empirical process theory and semi-infinite linear programming.

Keywords. Tukey depth, level sets, multivariate quantiles, convexity, support function, Hausdorff distance, semi-infinite linear programming.

1 Introduction

Pour des données multivariées, il n'existe pas d'ordre canonique, contrairement au cas univarié, où beaucoup de procédures statistiques sont basées sur les statistiques d'ordre (tests de rang, valeurs extrêmes, etc.). C'est pourquoi plusieurs notions de *profondeur statistique* ont été proposées, afin d'ordonner les données de sorte que, dans un nuage de point multivarié, les points les plus centraux soient considérés comme profonds, et que les points périphériques soient considérés comme peu profonds. La notion de profondeur

statistique a été définie de manière rigoureuse par [14]. Le cas particulier de la profondeur de Tukey, définie par [12] dans un but général de décrire les données multivariées, est celui auquel nous nous intéressons ici. Cette notion de profondeur a été le sujet de beaucoup de travaux de recherche, avec des applications à la visualisation des données [7], la statistique robuste [3, 1], l'inférence non paramétrique [8], le bootstrap [13], la classification supervisée [5, 6], etc. Certaines propriétés asymptotiques de la profondeur de Tukey empirique ont été établies par [10], et la définition même de cette profondeur a été généralisée au cas de données fonctionnelles, ou non Euclidiennes [4]. Ici, nous nous intéressons aux ensembles de niveau de la profondeur de Tukey empirique et à leurs propriétés non asymptotiques, notamment leur concentration, au sens de la distance de Hausdorff, autour des ensembles de niveau théoriques. Nous montrons aussi pourquoi ces ensembles peuvent être interprétés comme des quantiles multivariés, et établissons un lien avec un objet important en géométrie convexe, appelé *corps flottant* ("floating body").

Cette présentation est basée sur le travail [2].

1.1 Définitions

Dans toute la suite, μ une mesure de probabilité sur $\mathbb{R}^d$ ($d \geq 1$ est fixé dans toute la suite). La profondeur de Tukey associée à μ est la fonction D_μ définie par

$$D_\mu(x) = \inf_{H \in \mathcal{H}: H \ni x} \mu(H), \quad \forall x \in \mathbb{R}^d,$$

où $\mathcal{H}$ est l'ensemble des demi-espaces fermés de $\mathbb{R}^d$. La profondeur d'un point est la plus petite masse d'un demi-espace fermé contenant ce point. Lorsque $X_1, \ldots, X_n$ sont des données dans $\mathbb{R}^d$ (où $n \geq 1$), la profondeur empirique est la fonction de profondeur $D_{\hat{\mu}_n}$, associée à la mesure empirique $\hat{\mu}_n$. Elle se réécrit de manière plus simple

$$D_{\hat{\mu}_n}(x) = \inf_{H \in \mathcal{H}: H \ni x} n^{-1} \#\{i = 1, \ldots, n : X_i \in H\}, \quad \forall x \in \mathbb{R}^d,$$

et s'interprète comme le nombre minimal de données incluses dans un demi-espace fermé contenant x.

On fixe un nombre $\alpha \in (0, 1)$ et on s'intéresse aux ensembles de niveau

$$G_\mu = \{x \in \mathbb{R}^d : D_\mu(x) \geq \alpha\} \quad \text{et} \quad G_{\hat{\mu}_n} = \{x \in \mathbb{R}^d : D_{\hat{\mu}_n}(x) \geq \alpha\}.$$

En particulier, que peut-on dire sur ces deux ensembles, et sur leur distance ? La distance que nous considérons est la distance de Hausdorff, définie par

$$\mathsf{d_H}(A, B) = \max\left(\max_{a \in A} \mathsf{d}(a, B), \max_{b \in B} \mathsf{d}(b, A)\right),$$

pour tous ensembles compacts non vides A et B dans $\mathbb{R}^d$: il s'agit de la plus grande distance d'un point d'un des deux ensembles à l'autre ensemble.

La fonction de support d'un ensemble convexe compact (non vide) G est la fonction

$$h_G(u) = \max_{x \in G} u^\top x, \quad \forall u \in \mathbb{R}^d.$$

Lorsque $u \neq 0$, il s'agit simplement de la plus grande distance signée de l'origine à un hyperplan tangent à G orthogonal à u. Une propriété essentielle de la distance de Hausdorff est que si les ensembles A et B sont convexes, alors

$$\mathsf{d_H}(A, B) = \max_{u \in \mathbb{S}^{d-1}} |h_A(u) - h_B(u)|, \tag{1}$$

où $\mathbb{S}^{d-1}$ est la sphère Euclidienne unité de $\mathbb{R}^d$.

Pour tout $u \in \mathbb{R}^d \setminus \{0\}$, on note F_u la fonction de répartition de la projection de μ le long de u, i.e., $F_u(t) = \mathbb{P}[u^\top X \leq t]$, pour tout $t \in \mathbb{R}$. On note les quantiles gauche et droit d'ordre $1-\alpha$ de la projection de μ le long de u comme suit:

$$q_u^\flat = \inf\{t \in \mathbb{R} : \mathbb{P}[u^\top X \leq t] \geq 1-\alpha\} \quad \text{et} \quad q_u^\sharp = \sup\{t \in \mathbb{R} : \mathbb{P}[u^\top X \geq t] \geq \alpha\}.$$

On peut à présent définir les quantiles multivariés gauche et droit d'ordre $(1-\alpha)$ de μ:

$$G^\eta_{\mathsf{MQ}} = \{x \in \mathbb{R}^d : \langle u, x\rangle \leq q_u^\eta, \forall u \in \mathbb{S}^{d-1}\}, \quad \eta \in \{\flat, \sharp\}. \tag{2}$$

Enfin, on rappelle la définition du corps flottant d'ordre α de μ:

$$G_{\mathsf{FB}} = \bigcap_{H \in \mathcal{H}: \mu(H) \geq 1-\alpha} H.$$

Cette définition apparaît essentiellement en géométrie convexe, dans la cas où μ est la mesure uniforme sur un ensemble convexe compact K: dans ce cas, G_{FB}, aussi appelé *corps flottant d'ordre α de K*, consiste en l'intersection de toutes les calottes de K (intersection de K avec un demi-espace fermé) de volume au moins égal à une fraction $1-\alpha$ du volume total de K [11].

1.2 Le cas univarié

En dimension 1, la profondeur de Tukey se décrit de manière très simple. Pour tout $x \in \mathbb{R}$, $D_\mu(x) = \min\left(\mu\left((-\infty, x]\right), \mu\left([x, \infty)\right)\right)$. En particulier, si μ est absolument continue par-rapport à la mesure de Lebesgue, $D_\mu(x) = \min\left(F(x), 1-F(x)\right)$, où F est la fonction de répartition de μ. L'ensemble G_μ n'est autre que $[q_\alpha^\flat, q_{1-\alpha}^\sharp]$, où, pour tout $\beta \in (0,1)$, $q_\beta^\flat$ (resp. $q_\beta^\sharp$) est le quantile gauche (resp. droit) d'ordre β de μ. Remarquons que le quantile gauche d'ordre α de μ coïncide avec l'opposé du quantile droit d'ordre $1-\alpha$ de $-X$, où $X \sim \mu$. Si $X_1, \ldots, X_n$ sont des données réelles, $G_{\hat{\mu}_n} = [X_{(\lceil n\alpha \rceil)}, X_{(\lfloor n(1-\alpha)\rfloor+1)}]$, où $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ correspond à l'échantillon réordonné. Dans ce cas, on observe immédiatement le lien entre l'ensemble de niveau de la profondeur de Tukey et la notion de quantiles.

1.3 Premières propriétés des ensembles de niveau

Le premier résultat que nous exposons ici montre le lien entre les quantiles multivariés, le corps flottant et l'ensemble de niveau de la profondeur de Tukey.

Lemme 1 $G_{FB} = G_{MQ}^{\flat} \subseteq G_{MQ}^{\sharp} = G_{\mu}$.

En particulier, si μ satisfait des conditions de continuité, ces quatre ensembles coïncident. Une autre conséquence de ce lemme est que l'ensemble G_{μ} (et, de même pour $G_{\hat{\mu}_n}$) est convexe, puisque, vu comme $G_{\mathsf{MQ}}^{\sharp}$, il est donné par des contraintes linéaires. Ainsi, et comme on le verra plus tard, sa fonction de support est donnée comme la valeur maximale d'une forme linéaire sous un nombre infini de contraintes linéaires:

$$h_{G_{\mu}}(u) = \max\left\{u^{\top}x : v\top x \leq q_v^{\sharp}, \forall v \in \mathbb{S}^{d-1}\right\}.$$

En particulier, $h_{G_{\mu}} \leq q_{\cdot}^{\sharp}$, avec égalité garantie dès lors que la fonction $q_{\cdot}^{\sharp}$ est convexe, ce qui n'est pas le cas en général (cf. [2, Proposition 1]).

1.4 Concentration des ensembles de niveau empiriques

Dans la suite, on considère un échantillon de données i.i.d. de loi μ, noté $X_1, \ldots, X_n$ et on cherche à borner, avec grande probabilité, $\mathsf{d}_{\mathsf{H}}(G_{\hat{\mu}_n}, G_{\mu})$. Pour $u \in \mathbb{R}^d$, on note $\hat{q}_u$ le quantile empirique droit des données projetées $u^{\top}X_1, u^{\top}X_2, \ldots, u^{\top}X_n$.

On fait les hypothèses suivantes sur μ. Les nombres strictement positifs ε, L, r et R sont fixés et satisfont $\varepsilon < r \leq R$.

Hypothèse 1

- *Pour tout* $u \in \mathbb{S}^{d-1}$, F_u *est continue sur* $[q_u^{\sharp} - \varepsilon, q_u^{\sharp} + \varepsilon]$.
- $\left|F_u(t) - F_u(q_u^{\sharp})\right| \geq L\left|t - q_u^{\sharp}\right|$, *pour tout* $u \in \mathbb{S}^{d-1}$ *et* $t \in [q_u^{\sharp} - \varepsilon, q_u^{\sharp} + \varepsilon]$.
- *Il existe* $a \in \mathbb{R}^d$ *tel que* $B(a, r) \subseteq G_{\mu} \subseteq B(a, R)$.

La seconde hypothèse assure que $q_u^{\flat} = q_u^{\sharp}$ et que F_u n'est pas trop plate autour de $q_u^{\sharp}$, ce qui assure la concentration du quantile empirique à la vitesse paramétrique, pour tout $u \in \mathbb{S}^{d-1}$. Les deux premières hypothèses sont vérifiées pour une très grande classe de distributions. Par exemple, si μ admet une densité strictement positive par-rapport à la mesure de Lebesgue.

La troisième hypothèse implique, en particulier, que G_{μ} n'est pas vide. Il est bien connu que dès lors que $\alpha \leq 1/(d+1)$, G_{μ} ne peut pas être vide, mais que si $\alpha > 1/(d+1)$, il existe des distributions μ pour lesquelles $G_{\mu} = \emptyset$ [3]. En revanche, si μ est log-concave, alors on a la garantie que $G_{\mu} \neq \emptyset$ lorsque $\alpha \leq 1/e$ [9, Lemme 5.12].

Notre résultat principal est le suivant.

Théorème 1 *Supposons les hypothèses ci-dessus satisfaites et soit* $x \geq 0$ *tel que* $\frac{10\sqrt{5(d+1)}}{L} \leq x < \varepsilon\sqrt{n}$, *où* $C = \frac{R}{r}\frac{1+\varepsilon/r}{1-\varepsilon/r}$ *et soit* $A = e^{-250(d+1)}$. *Avec probabilité au moins* $1 - Ae^{-L^2x^2/2+10\sqrt{5(d+1)}x}$,

$$\mathsf{d_H}(G_{\hat{\mu}_n}, G_\mu) \leq \frac{Cx}{\sqrt{n}}.$$

En particulier, cela prouve que $\mathsf{d_H}(G_{\hat{\mu}_n}, G_\mu) = O_{\mathbb{P}}(n^{-1/2})$. La démonstration de ce théorème repose essentiellement sur les deux arguments suivants.

Lemme 2 *Pour tout* $z \in \mathbb{R}$ *satisfaisant* $\frac{10\sqrt{5(d+1)}}{L\sqrt{n}} \leq z < \varepsilon$,

$$\mathbb{P}\left[\sup_{u\in\mathbb{S}^{d-1}} |\hat{q}_u - q^\sharp_u| \leq z\right] \geq 1 - A\exp\left(-L^2z^2n/2 + 10\sqrt{5(d+1)}Lz\sqrt{n}\right).$$

Ce premier lemme repose sur la théorie des processus empiriques et sur le fait que, grâce à la seconde hypothèse,

$$\mathbb{P}\left[\sup_{u\in\mathbb{S}^{d-1}} |\hat{q}_u - q^\sharp_u| \leq z\right] \geq \mathbb{P}\left[\sup_{H\in\mathcal{H}_0} |\mu_n(H) - \mu(H)| \leq Lz\right],$$

où $\mathcal{H}_0$ est un sous-ensemble de la famille de tous les demi-espaces fermés de $\mathbb{R}^d$, dont la dimension de Vapnik-Chervonenkis est égale à $d+1$.

Le second lemme permet de contrôler $\sup_{u\in\mathbb{S}^{d-1}} |h_{G_{\hat{\mu}_n}}(u) - h_{G_\mu}(u)|$ à partir du contrôle de $\sup_{u\in\mathbb{S}^{d-1}} |\hat{q}_u - q^\sharp_u|$. C'est l'argument clef de la démonstration, qui s'appuie sur la programmation linéaire semi-infinie. En effet, d'après le Lemme 1, $h_{G_\mu}(u)$ (et de même pour $h_{G_{\hat{\mu}_n}}(u)$) est la valeur maximale de la forme linéaire $u^\top x$ sous les contraintes linéaires $v^\top x \leq q^\sharp_v, v \in \mathbb{S}^{d-1}$.

Pour toute fonction $\zeta : \mathbb{S}^{d-1} \to \mathbb{R}$, on note $K_\zeta = \{x \in \mathbb{R}^d : u^\top x \leq \zeta(u), \forall u \in \mathbb{S}^{d-1}\}$.

Lemme 3 *Soient* ϕ *et* $\hat{\phi}$ *deux fonctions continues sur* $\mathbb{S}^{d-1}$. *Supposons que* K_ϕ *and* $K_{\hat{\phi}}$ *sont d'intérieurs non vides et que* $B(a,r) \subseteq K_\phi \subseteq B(a,R)$, *pour un certain* $a \in \mathbb{R}^d$. *Soit* $\eta = \max_{u\in\mathbb{S}^{d-1}} |\hat{\phi}(u) - \phi(u)|$. *Alors, si* $\eta < r$, *on a*

$$\mathsf{d_H}(K_{\hat{\phi}}, K_\phi) \leq \frac{\eta R}{r}\frac{1+\eta/r}{1-\eta/r}.$$

On applique enfin ce lemme déterministe à $G_\mu = K_{q^\sharp_\cdot}$ et $G_{\hat{\mu}_n} = K_{\hat{q}^\sharp_\cdot}$, en remarquant que la fonction $\hat{q}^\sharp_\cdot : u \mapsto \hat{q}^\sharp_u$ est μ-presque sûrement continue, et que la fonction $q^\sharp_\cdot : u \mapsto q^\sharp_u$, quant à elle, est continue grâce aux deux premières hypothèses ci-dessus.

References

[1] Miguel A. Arcones, Zhiqiang Chen, and Evarist Giné. Estimators related to U-processes with applications to multivariate medians: asymptotic normality. *Ann. Statist.*, 22(3):1460–1477, 1994.

[2] Victor-Emmanuel Brunel. Concentration of the empirical level sets of tukey's halfspace depth. *Probability Theory and Related Fields*, 173(3-4):1165–1196, 2019.

[3] David L. Donoho and Miriam Gasko. Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *Ann. Statist.*, 20(4):1803–1827, 1992.

[4] Subhajit Dutta, Anil K. Ghosh, and Probal Chaudhuri. Some intriguing properties of Tukey's half-space depth. *Bernoulli*, 17(4):1420–1434, 2011.

[5] Anil K. Ghosh and Probal Chaudhuri. On data depth and distribution-free discriminant analysis using separating surfaces. *Bernoulli*, 11(1):1–27, 2005.

[6] Anil K. Ghosh and Probal Chaudhuri. On maximum depth and related classifiers. *Scand. J. Statist.*, 32(2):327–350, 2005.

[7] Regina Y. Liu, Jesse M. Parelius, and Kesar Singh. Multivariate analysis by data depth: descriptive statistics, graphics and inference. *Ann. Statist.*, 27(3):783–858, 1999. With discussion and a rejoinder by Liu and Singh.

[8] Regina Y. Liu and Kesar Singh. A quality index based on data depth and multivariate rank tests. *J. Amer. Statist. Assoc.*, 88(421):252–260, 1993.

[9] László Lovász and Santosh Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures Algorithms*, 30(3):307–358, 2007.

[10] Jean-Claude Massé. Asymptotics for the Tukey depth process, with an application to a multivariate trimmed mean. *Bernoulli*, 10(3):397–419, 2004.

[11] C. Schütt and E. Werner. The convex floating body. *Math. Scand*, 66:275–290, 1990.

[12] John W. Tukey. Mathematics and the picturing of data. *Proceedings of the International Congress of Mathematicians (Vancouver, B. C., 1974), Vol. 2*, pages 523–531, 1975.

[13] Arthur B. Yeh and Kesar Singh. Balanced confidence regions based on Tukey's depth and the bootstrap. *J. Roy. Statist. Soc. Ser. B*, 59(3):639–652, 1997.

[14] Yijun Zuo and Robert Serfling. General notions of statistical depth function. *Ann. Statist.*, 28(2):461–482, 2000.

MLDA-TCT : UNE METHODE D'ANALYSE DE TABLEAUX DE CONTINGENCE A TROIS ENTREES

Philippe CASIN

Université de Lorraine, LCOMS , 7 rue Marconi,57070 Metz philippe.casin@univ-lorraine.fr

Résumé. L'objet de cette communication est de présenter la méthode MLDA-TCT (Multiblock Linear Discriminant Analysis of Three-way Contingency Tables). Cette méthode permet de décrire un ensemble de tableaux de contingence observés dans différentes circonstances et possédant tous le même nombre de lignes et de colonnes. MLDA-TCT calcule une ou plusieurs composantes pour chaque tableau : ces composantes décrivent d'une part les relations entre les différents tableaux, et d'autre part les relations entre les lignes et les colonnes de ces tableaux. Un exemple d'application est donné : la comparaison des tableaux de contingence décrivant la relation entre le type d'objet volé et la classe d'âge, selon les genres.

Mots-clés. Analyse discriminante, analyse canonique généralisée, analyse des correspondance multi-blocs, tableau de données bi-partitionné.

Abstract. The purpose of this communication is to introduce MLDA-TCT (Multiblock Linear Discriminant Analysis of Three-way Contingency Tables), a new method for analyzing a set of contingency tables which have been observed in different occasions and have the same number of rows and the same number of columns. MLDA-TCT computes one or several components for each data table : these components take into account canonical correlation between data tables and the partition given by the occasions in one hand, and in the other hand take into account relationships between rows and columns of the contingency tables in the other hand. An example of application is given: it concerns relationship between stolen object and age classes, by gender.

Keywords. Linear discriminant analysis, generalized canonical analysis, multiblock correspondence analysis, bi-partitioned data table.

1. Les données et le problème

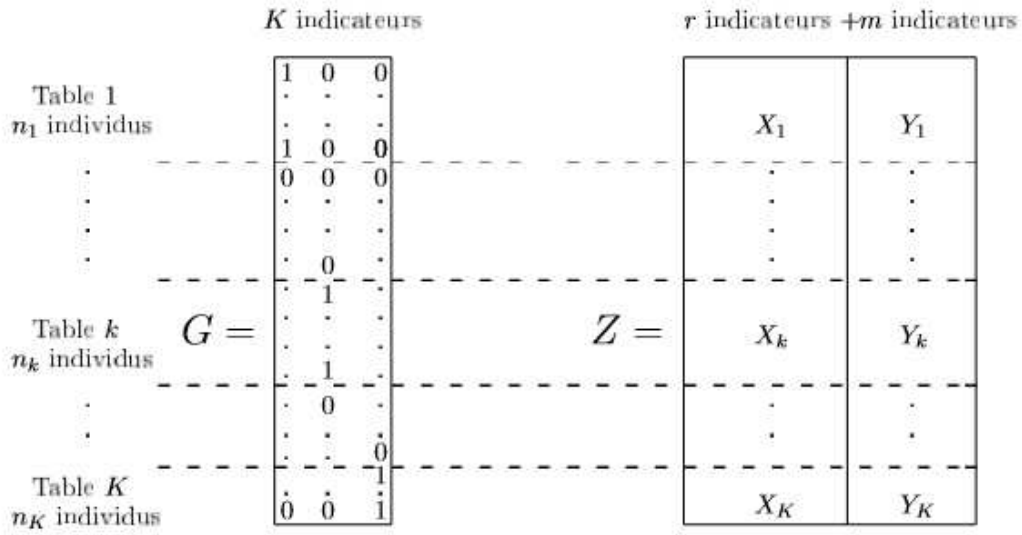
1.1 Les données

On considère une suite de K tableaux de contingence C_k, pour $k = 1, \dots, K$. Chaque tableau C_k comporte r lignes et m colonnes.

X_k (resp. Y_k) est le tableau disjonctif complet à n_k lignes et r (resp. m colonnes) qui indique la modalité des lignes (resp. colonnes) prise par chacun des n_k individus du tableau C_k , pour $k = 1, \dots, K$. Il s'ensuit que $C_k = X_k' Y_k$.

X (resp. *Y*) désigne le tableau disjonctif complet à $n = \sum_{k=1}^{K} n_k$ lignes et *r* (resp. *m*) colonnes qui indique la ligne (resp. la colonne) à laquelle appartient chacun des *n* individus des *K* tableaux.

G est le tableau disjonctif complet à *n* lignes et *K* colonnes qui indique à quel tableau appartient chacun des *n* individus, tandis que Z est le tableau juxtaposant X et Y.



1.2 Le problème

Il s'agit à la fois :

- de décrire chacun des tableaux de contingence C_k, pour $k=1,...,K$, au sens d'une analyse des correspondances (et donc d'une analyse canonique entre X_k et Y_k), c'est à dire finalement de décrire les relations entre *X* et *Y*.
- d'expliquer la partition des individus en *K* classes, au sens d'une analyse discriminante pour laquelle *G* est la variable à expliquer et *X* et *Y* sont les variables explicatives.

2. La méthode

2.1 Les critères à maximiser

MLDA-TCT (Multiblock Linear Discriminant Analysis of Three-way Contingency Tables, Casin (2019)) est la méthode MLDA (Casin (2018)) appliquée à une suite de tableaux de contingence.

A l'étape 1, MLDA-TCT détermine la variable synthétique normée z^1 telle que :

- $R^2(z^1, z_X^1)$ (resp. $R^2(z^1, z_Y^1)$) ait une valeur maximale, z_X^1 (resp. z_Y^1) désignant la projection de z^1 sur l'espace engendré par les colonnes de *X* (resp. *Y*) : il s'agit du critère d'analyse canonique généralisée de Carroll (1968).
- η_X^2 (resp. η_Y^2) ait une valeur maximale, η_X^2 (resp. η_Y^2) désignant le pouvoir discriminant de la variable z_X^1 (resp. z_Y^1) par rapport à la variable à expliquer *G* : il s'agit du critère

d'analyse discriminante de Fisher (1936).

La maximisation séparée de chacun de ces deux critères ne conduit pas aux mêmes solutions, et on choisit donc de maximiser un compromis entre ces deux critères.
A l'étape 1, il s'agit de déterminer la composante normée z^1 telle que $R^2(z^1, z_X^1)\eta_X^2 + R^2(z^1, z_Y^1)\eta_Y^2$ ait une valeur maximale.
A l'étape *k*, z^k est la composante normée telle que $R^2(z^k, z_X^k)\eta_X^k + R^2(z^k, z_Y^k)\eta_Y^k$ soit maximal sous les contraintes d'orthogonalisation $R(z_X^k, z_X^j)$ pour $j = 1, \dots, k-1$ et $R(z_Y^k, z_Y^j)$ pour $j = 1, \dots, k-1$.

2.2 La solution

Soit P_X (resp. P_Y, P_G) le projecteur sur l'espace engendré par les colonnes de *X* (resp. *Y*, *G*), alors z^1 est le premier vecteur propre normé de $P_X P_G P_X + P_Y P_G P_Y$. On en déduit $z_X^1 = P_X z^1$ et $z_Y^1 = P_Y z^1$.
A l'étape *k*, P_{X_k} (resp. P_{Y_k}) désignant le projecteur sur l'espace engendré par les résidus de la régression des colonnes de *X* (resp. *Y*) par les variables $z_X^1, \dots . z_X^{k-1}$ (resp. $z_Y^1, \dots . z_Y^{k-1}$), z^k est le premier vecteur propre normé de $P_{X_k} P_G P_{X_k} + P_{Y_k} P_G P_{Y_k}$. Il s'ensuit que $z_X^k = P_{X_k} z^k$ et $z_Y^k = P_Y z^k$.
En raison des contraintes d'orthogonalisation, les vecteurs z_X^k (resp. z_Y^k) sont deux à deux orthogonaux, ce qui permet d'obtenir une base orthogonale de l'espace engendré par les colonnes de *X* (resp. *Y*). Les variables z^ksont aussi deux à deux orthogonales (Casin (2019)), ce qui permet d'obtenir une représentation simultanée des variables *X* et *Y*.

3. Un exemple d'application : la relation entre le type d'objet volé et la classe d'âge, selon le genre

3.1 Les données

Chacun des deux tableaux de données (Israels, (1987)), l'un pour les 20 819 hommes, l'autre pour les 12 282 femmes, comporte 9 classes d'âge (de 0-11 à 65+) et 13 types d'objet volés dans des magasins (CLOThes, CLothing, Accessories, TOBAcco, WRITing accessories, BOOKs, RECOrds, HOUSehold accessories, SWEEts, TOYS, JEWEllery, PERFums, HOBBies an OTHErs).

3.2 Les résultats numériques

Les résultats pour les deux premières étapes de calcul sont donnés par le tableau ci-dessous :

	λ^j	μ_z^j	$R^2(z^j, z_X^j)$	μ_X^j	$R^2(z^j, z_Y^j)$	μ_Y^j
Etape j=1	.183	.190	.957	.176	.017	.051
Etape j=2	.012	.012	.034	.001	.003	.012

3.3 Les sorties graphiques

L'orthogonalité entre elles des variables z^k d'une part, et l'orthogonalité des variables z_X^k (resp. z_Y^k) entre elles d'autre part, va permettre d'obtenir trois graphiques : une représentation simultanée des objets volés et des classes d'âge (figure 1 : les composantes z^k), une représentation des objets volés (figure 2 : les composantes z_X^k) et une représentation des classes d'âge (figure 3 : les composantes z_Y^k) .

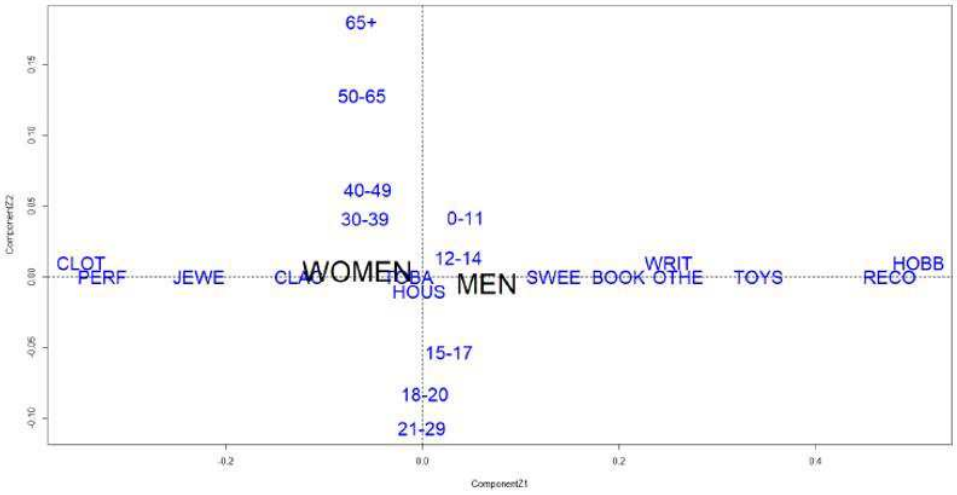


Figure 1 : représentation simultanée des objets volés et des classes d'âge, axes 1 et 2

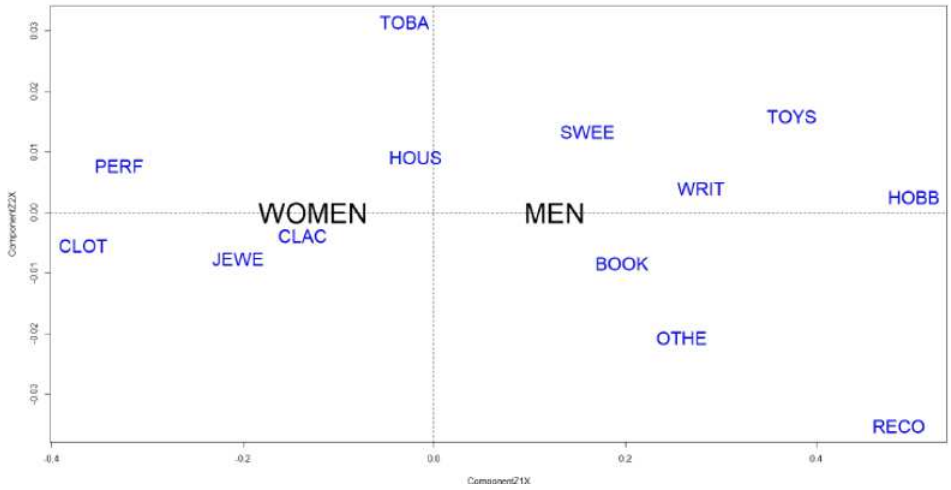


Figure 2 : représentation des objets volés, axes 1 et 2

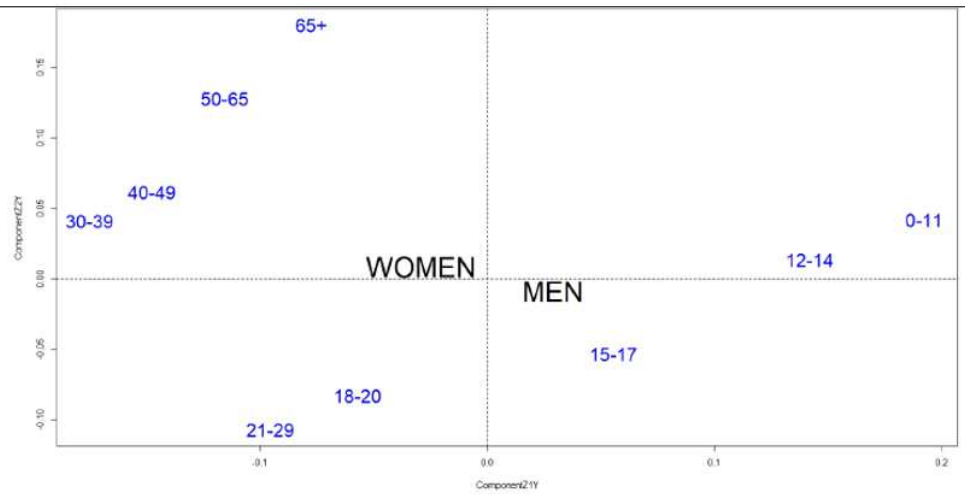


Figure 3 : représentation des classes d'âge, axes 1 et 2

3.4 L'interprétation des résultats

En ce qui concerne le premier axe, la position (à gauche) de CLOC, CLAC, PERF, JEWE, 30-39,40-49,50-59 correspond à une sur-représentation de ces objets volés pour les femmes et en particulier les plus âgées d'entre elles. RECO, WRIT, SWEET, BOOK, OTHE et HOBB, 0-11, 12-14, 15-17 se trouvent du coté droit : ces objets sont volés par des hommes, et plus particulièrement des hommes jeunes.
Le pouvoir discriminant du second axe est beaucoup moins élevé, et indique surtout que les femmes qui dérobent des objets sont en général plutôt âgées alors que les hommes sont plutôt jeunes.

Bibliographie

Carroll, J.D. (1968), Generalization of canonical correlation analysis to three or more sets of variables, *Proceedings of the 76th annual convention of the American Psychological Association,* vol.3, pp 227-228.
Casin, Ph. (2018), Categorical Multiblock Linear Discriminant Analysis, *Journal of applied statistics*, 45(8), pp. 1396-1409
Casin, Ph. (2019), Une extension de l'analyse discriminante multi-blocs aux tableaux de contingence ternaires, *Journal de la Société Française de Statistique*, 160 (2), pp. 67-82.
Fisher, R.A (1936), the use of multiple measurements in taxonomic problems, *Annals of eugenics,* 3, pp. 179-188.
Israels, A. (1987), *Eigenvalue technics for qualitative data*, DSWO Press

PRÉVISION DANS LE MODÉLE LINÉAIRE FONCTIONNEL EN PRÉSENCE DE DONNÉES MANQUANTES DANS LA RÉPONSE ET LA COVARIABLE

Christophe Crambes[1] & Chayma Daayeb[1,2] & Ali Gannoun[1] & Yousri Henchiri[2,3]

[1]*Institut Montpelliérain Alexander Grothendieck, Université de Montpellier, France.*
[2]*Université de Tunis El Manar, Laboratoire de Modélisation Mathématique et Numérique dans les Sciences de l'Ingénieur (ENIT-LAMSIN), Tunisie.*
[3]*Université du Québec à Montréal, Département de Mathématiques, 201 Avenue du Président Kennedy, H2X 3Y7, Montréal, Canada.*
E-mail : christophe.crambes@umontpellier.fr, chayma.daayeb@etu.umontpellier.fr, ali.gannoun@umontpellier.fr, yousri.henchiri@umontpellier.fr

Résumé. Les valeurs manquantes sont un des problèmes qui surviennent fréquemment dans le processus d'observation ou d'enregistrement des données. Dans ce travail, nous considérons le modèle de régression linéaire fonctionnelle, lorsque la variable d'intérêt, réelle, et la variable explicative, fonctionnelle, contiennent des valeurs manquantes. Nous utilisons un opérateur de reconstruction qui vise à reconstruire les parties manquantes dans les courbes, puis nous nous intéressons à la méthode d'imputation par régression des données manquantes sur la variable réponse, en utilisant la régression fonctionnelle sur composantes principales pour estimer le coefficient fonctionnel du modèle. Nous étudions le comportement asymptotique de l'erreur de prévision commise lorsque les valeurs manquantes sont remplacées par les valeurs imputées. Le comportement de la méthode est également étudié en pratique sur des données simulées et réelles.

Mots-clés. Modèle linéaire fonctionnel, Données manquantes, Composantes Principales Fonctionnelles, Missing At Random, Missing Completely At Random, Imputation par régression.

Abstract. Dealing with missing values is an important issue in data observation or data recording process. In this paper, we consider a functional linear regression model, when some observations of the real response and the functional covariate are missing. We use a reconstruction operator that aims at recovering the missing parts of the explanatory curves, then we are interested in regression imputation method of missing data on the response variable, using functional principal component regression to estimate the functional coefficient of the model. We study the asymptotic behaviour of the prediction error we commit when missing data are replaced by the imputed values. The practical behaviour of the method is also studied on simulated and real datasets.

Keywords. Functional linear model, Missing data, Functional Principal Components, Missing At Random, Missing Completely At Random, Regression imputation.

1 Introduction

L'analyse des données fonctionnelles a connu un développement très important ces dernières années, comme en attestent les nombreux ouvrages sur le sujet : Ramsay et Silverman (2005), Ferraty et Vieu (2006), Hsing et Eubank (2015) constitue une liste non exhaustive de monographies donnant une vision d'ensemble sur ce thème. Un des modèles les plus populaires en analyse de données fonctionnelles est le modèle linéaire fonctionnel, qui établit une relation de dépendance entre une variable réelle Y et une variable aléatoire fonctionnelle $X = (X(t), t \in [a, b])$. La variable X est à valeurs dans l'espace $H := L^2([a, b])$ des fonctions de carré intégrable sur l'intervalle compact $[a, b]$. Nous supposons dans la suite que $\mathbb{E}(\|X\|^2) < +\infty$ où $\|.\|$ désigne la norme usuelle de H associée au produit scalaire $\langle ., . \rangle$, défini par $\langle f, g \rangle = \int_a^b f(t)g(t)\mathrm{d}t$ pour toutes fonctions f et g de H. Le modèle linéaire fonctionnel a été étudié par de nombreux auteurs, par exemple Cardot et al. (1999), Cai et Hall (2006), Hall et Horowitz (2007), Crambes et al. (2009). Ce modèle est défini par

$$Y = \theta_0 + \int_a^b \theta(t)X(t)\mathrm{d}t + \varepsilon, \tag{1.1}$$

où $\theta_0 \in \mathbb{R}$ et $\theta \in H$ sont les paramètres à estimer. L'erreur du modèle ε est une variable aléatoire réelle centrée indépendante de X avec une variance finie $\mathbb{E}(\varepsilon^2) = \sigma_\varepsilon^2$. Le modèle (1.1) peut s'écrire sous la forme

$$Y = \theta_0 + \Theta X + \varepsilon, \tag{1.2}$$

où $\Theta : H \to \mathbb{R}$ est l'opérateur linéaire continu défini par $\Theta u = \langle \theta, u \rangle$ pour toute fonction $u \in H$. Dans la suite, nous considérons un échantillon $(X_i, Y_i)_{i=1,\ldots,n}$ indépendant et identiquement distribué de même loi que le couple (X, Y). Pour estimer θ ou Θ, nous considérons la régression fonctionnelle sur composantes principales. Il s'agit d'une régression des moindres carrés de la réponse Y sur les variables réelles qui sont les coordonnées de la projection de X sur l'espace engendré par les fonctions propres associées aux plus grandes valeurs propres de l'opérateur de covariance de X (voir Cardot et al., 1999). Soit $(k_n)_{n\geq 1}$ une suite de nombres entiers, l'estimateur $\widehat{\Theta}$ de Θ proposé par Cardot et al. (1999) est défini par

$$\widehat{\Theta} = \langle \widehat{\theta}, . \rangle = \widehat{\Pi}_{k_n} \widehat{\Delta}_n (\widehat{\Pi}_{k_n} \widehat{\Gamma}_n \widehat{\Pi}_{k_n})^{-1}, \tag{1.3}$$

où $\widehat{\Delta}_n$ est l'opérateur de covariance croisée empirique donné par $\widehat{\Delta}_n u = \frac{1}{n} \sum_{i=1}^n \langle X_i, u \rangle Y_i$ pour tout $u \in H$, $\widehat{\Gamma}_n$ est l'opérateur de covariance empirique défini par $\widehat{\Gamma}_n u = \frac{1}{n} \sum_{i=1}^n \langle X_i, u \rangle X_i$ pour tout $u \in H$ et $\widehat{\Pi}_{k_n} = \sum_{j=1}^{k_n} \langle \widehat{\phi}_j, u \rangle \widehat{\phi}_j$ est l'opérateur de projection orthogonale sur le sous-espace engendré par les fonctions propres $(\widehat{\phi}_1, \ldots, \widehat{\phi}_{k_n})$ associées aux k_n plus grandes valeurs propres $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_{k_n}$ de l'opérateur $\widehat{\Gamma}_n$. En supposant que $\widehat{\lambda}_1 > \ldots > \widehat{\lambda}_{k_n} > 0$ p.s., l'estimateur de θ est donné par

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k_n}\frac{\langle X_i, \widehat{\phi}_j\rangle Y_i}{\widehat{\lambda}_j}\widehat{\phi}_j. \tag{1.4}$$

En outre, l'estimateur de $\theta_0 = \mathbb{E}(Y) - \int_a^b \theta(t)\mathbb{E}(X(t))\mathrm{d}t$ s'écrit sous la forme $\widehat{\theta}_0 = \overline{Y} - \int_a^b \widehat{\theta}(t)\overline{X}(t)\mathrm{d}t$ où $\overline{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$ et $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$.

Le cadre de ce travail est la situation où des valeurs manquantes affectent à la fois la réponse et la variable explicative. L'objectif est ***(i)*** de reconstituer les courbes X manquantes et d'imputer les données manquantes sur Y, ***(ii)*** d'estimer θ ou Θ avec le jeu de données reconstitué, ***(iii)*** de prédire une nouvelle valeur de la réponse Y étant donnée une nouvelle observation test sur la variable explicative X.

2 Données manquantes

Pour le mécanisme de données manquantes dans la réponse, nous considérons une variable aléatoire binaire $\delta^{[Y]}$ et un échantillon $(\delta_i^{[Y]})_{i=1,\dots,n}$ tel que $\delta_i^{[Y]} = 1$ si la valeur Y_i est observée et $\delta_i^{[Y]} = 0$ si la valeur Y_i est manquante, pour tout $i = 1, \dots, n$. Nous considérons les données manquantes de la réponse "Missing At Random" (MAR) : le fait que la valeur Y est manquante ne dépend pas de la réponse du modèle, mais peut éventuellement dépendre de la covariable, c'est-à-dire

$$\mathbb{P}(\delta^{[Y]} = 1 \mid X, Y) = \mathbb{P}(\delta^{[Y]} = 1 \mid X).$$

Dans ce qui suit, le nombre de valeurs manquantes parmi $Y_1, \dots, Y_n$ est noté

$$m_n^{[Y]} = \sum_{i=1}^{n} \mathbb{1}_{\left\{\delta_i^{[Y]}=0\right\}}.$$

Dans Crambes et Henchiri (2019), une méthodologie d'imputation des données manquantes par régression est donnée, sous cette hypothèse MAR, mais la covariable est censée être complètement observée, ce qui n'est plus le cas ici. Nous considérons une variable fonctionnelle $\delta^{[X]}$ et un échantillon $(\delta_i^{[X]})_{i=1,\dots,n}$ tel que, pour $t \in [a, b]$, $\delta_i^{[X]}(t) = 1$ si $X_i(t)$ est observé et $\delta_i^{[X]}(t) = 0$ si $X_i(t)$ est manquant. Nous considérons les données manquantes de la covariable "Missing Completely At Random" (MCAR) : le fait que X contient des données manquantes ne dépend pas de la covariable du modèle, ni de la réponse, c'est-à-dire, pout tout $t \in [a, b]$

$$\mathbb{P}(\delta^{[X]}(t) = 1 \mid X, Y) = \mathbb{P}(\delta^{[X]}(t) = 1).$$

D'autre part, le nombre de courbes où des valeurs manquantes apparaissent est donné par

$$m_n^{[X]} = \sum_{i=1}^{n} \mathbb{1}_{\left\{\exists\, t\in\,[a,b],\, \delta_i^{[X]}(t)=0\right\}}.$$

Dans la suite, nous présentons la méthodologie de reconstruction des courbes, puis l'imputation par régression d'une valeur manquante pour la variable d'intérêt. Enfin, nous donnons l'estimateur de θ à partir du jeu de données reconstitué, et la prédiction d'une valeur de la réponse suite à la donnée d'une nouvelle observation test pour la variable explicative.

3 Reconstruction des covariables manquantes

Nous appliquons dans cette partie la méthodologie introduite dans Kneip et Liebl (2020) pour reconstruire la partie manquante d'une courbe. Soit $(O_i)_{i=1,\dots,n}$ l'échantillon des périodes d'observation des courbes, c'est-à-dire $O_i = \left\{ t \in [a,b], \delta_i^{[X]}(t) = 1 \right\}$ pour tout $i = 1, \dots, n$. En outre, notons $M_i = [a,b] \setminus O_i$ pour tout $i = 1, \dots, n$. Dans la suite, nous utilisons O et M pour désigner une production donnée de O_i et M_i. De plus, nous notons la partie observée de X_i par X_i^O et X_i^M pour la partie manquante. Nous considérons la décomposition de Karhunen-Loève (KL) pour X_i^O dans $L^2(O)$

$$X_i^O(t) = \sum_{k=1}^{+\infty} \xi_{ik}^O \phi_k^O(t), \tag{3.1}$$

pour $t \in O$, où $(\phi_k^O)_{k\geq 1}$ désigne la suite des fonctions propres de l'opérateur de covariance de X_i^O et $(\xi_{ik}^O)_{k\geq 1}$ est une suite de variables aléatoires centrées et décorrélées avec $\mathbb{E}(\xi_{ik}^O) = \lambda_k^O$, la suite $(\lambda_k^O)_{k\geq 1}$ étant la suite des valeurs propres de l'opérateur de covariance de X_i^O. La partie manquante de la courbe s'écrit, pour $t \in O$ et $s \in M$

$$X_i^M(s) = L(X_i^O(t)) + Z_i(s), \tag{3.2}$$

où $L : L^2(O) \to L^2(M)$ est un opérateur linéaire défini par

$$L(X_i^O(t)) = \sum_{k=1}^{+\infty} \frac{\mathbb{E}[\xi_{ik}^O X_i^M(s)]}{\lambda_k^O},$$

dont le but est de reconstruire les parties manquantes $X_i^M \in L^2(M)$ à partir des observations $X_i^O \in L^2(O)$. La variable $Z_i \in L^2(M)$ est l'erreur de reconstruction. Nous cherchons à minimiser l'erreur quadratique moyenne $\mathbb{E}[(X_i^M(t) - L(X_i^O)(t))^2]$ avec $t \in M$, pour obtenir l'opérateur de reconstruction linéaire optimal. Sous des hypothèses classiques dans ce contexte, des vitesses de convergence uniforme de la courbe reconstruite vers la vraie courbe sont données dans Kneip et Liebl (2020). Dans la suite, nous notons $\widehat{X}$ une courbe X reconstruite et $X^\star = \delta^{[X]} X + (1 - \delta^{[X]})\widehat{X}$.

4 Imputation par régression

Nous nous intéressons ici à l'imputation des données manquantes sur la réponse Y, suivant la méthode présentée dans Crambes et Henchiri (2019). Nous définissons l'opérateur de covariance avec les courbes reconstruites par

$$\widehat{\Gamma}_{n,rec}^{obs} = \frac{1}{n - m_n^{[Y]}} \sum_{i=1}^{n} \langle X_i^\star, . \rangle \delta_i^{[Y]} X_i^\star.$$

Soit ℓ un nombre entier compris entre 1 et n tel que Y_ℓ soit manquante, c'est-à-dire avec $\delta_\ell^{[Y]} = 0$. La valeur imputée par régression pour Y_ℓ est définie par

$$Y_{\ell,imp} = \widetilde{\theta}_0 + \langle \widetilde{\theta}, X_\ell^\star \rangle, \tag{4.1}$$

avec

$$\widetilde{\theta} = \frac{1}{n - m_n^{[Y]}} \sum_{i=1}^{n} \sum_{j=1}^{k_n} \frac{\langle X_i^\star, \widehat{\phi}_{j,rec}^{obs} \rangle \delta_i^{[Y]} Y_i}{\widehat{\lambda}_{j,rec}^{obs}} \widehat{\phi}_{j,rec}^{obs} \quad \text{et} \quad \widetilde{\theta}_0 = \overline{Y}_{obs} - \int_a^b \widetilde{\theta}(t) \overline{X}^\star(t) \mathrm{d}t,$$

où $(\widehat{\lambda}_{j,rec}^{obs})_{j \geq 1}$ et $(\widehat{\phi}_{j,rec}^{obs})_{j \geq 1}$ sont les éléments propres de l'opérateur $\widehat{\Gamma}_{n,rec}^{obs}$ et les moyennes empiriques sont $\overline{Y}_{obs} = \frac{1}{n} \sum_{i=1}^{n} \delta_i^{[Y]} Y_i$ et $\overline{X}^\star = \frac{1}{n} \sum_{i=1}^{n} X_i^\star$.

Sous des hypothèses analogues à celles de Kneip et Liebl (2020) et Crambes et Henchiri (2019), nous obtenons des vitesses de convergence pour la valeur imputée $Y_{\ell,imp}$ de façon similaire à Crambes et Henchiri (2019) (cas d'une covariable fonctionnelle complètement observée et d'une réponse affectée par des données manquantes).

5 Prédiction

Une fois la base de données reconstruite, nous estimons le coefficient fonctionnel θ et l'intercept θ_0, respectivement, par

$$\widehat{\theta}^\star = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k_n} \frac{\langle X_i^\star, \widehat{\phi}_{j,rec} \rangle Y_i^\star}{\widehat{\lambda}_{j,rec}} \widehat{\phi}_{j,rec} \quad \text{et} \quad \widehat{\theta}_0^\star = \overline{Y}^\star - \int_a^b \widehat{\theta}^\star(t) \overline{X}^\star(t) \mathrm{d}t,$$

avec $Y_i^\star = Y_i \delta_i^{[Y]} + Y_{i,imp}(1 - \delta_i^{[Y]})$, pour tout $i = 1, \ldots, n$, et $\overline{Y}^\star = \frac{1}{n} \sum_{i=1}^{n} Y_i^\star$. Nous pouvons à présent définir la prédiction d'une nouvelle valeur Y_{new} associée à l'observation X_{new} de la variable explicative par

$$\widehat{Y}_{new} = \widehat{\theta}_0^\star + \langle \widehat{\theta}^\star, X_{new}^\star \rangle. \tag{5.1}$$

Une étude asymptotique de l'erreur de prévision de $\widehat{Y}_{new}$ a été réalisée. Le comportement de la méthode en pratique a également été évalué sur des données simulées et réelles. À titre d'exemple, nous avons simulé 400 réplications sur le modèle (1.1) avec $[a;b] = [0;1]$, $\theta_0 = 3$, et $\theta(t) = \sum_{j=1}^{50} b_j \Phi_j(t)$ pour tout $t \in [0;1]$, où $b_1 = 0.3$, $b_j = 4(-1)^{j+1} j^{-2}$ pour $j > 1$ et $\Phi_1(t) = 1$, $\Phi_j(t) = \sqrt{2}\cos(j\pi t)$ pour $t \in [0;1]$ et $j > 1$. Le bruit ε est simulé suivant la loi $N(0;\sigma_\varepsilon^2)$ avec $\sigma_\varepsilon^2 = 0.04$ et la variable X est simulée en écrivant $X(t) = \sum_{j=1}^{150} \xi_j \lambda_j \Phi_j(t)$ où $\lambda_j = (-1)^{j+1} j^{-2}$ pour $j \geq 1$ et ξ_j suit la loi uniforme sur l'intervalle $[-\sqrt{3};\sqrt{3}]$. Nous avons pris une taille d'échantillon $n = 360$ et les courbes ont été discrétisées à l'aide de $p = 100$ points de mesure. Concernant les données manquantes, les parties $[0;1/8]$ et $[7/8;1]$ ont été retirées aléatoirement de 14.8% des courbes X et 12% de données ont été retirées de Y. Sur les 400 réplications, nous avons calculé une erreur quadratique moyenne de prévision de 8.46×10^{-2} avec un écart-type de 8.66×10^{-2} lorsque nous avons reconstruit les courbes X et imputé les données manquantes sur Y. L'erreur quadratique moyenne de prévision devient 9.24×10^{-2} avec un écart-type de 8.82×10^{-2} lorsque l'on se contente de supprimer les observations pour lesquelles X ou Y est manquant. Cela montre, sur cet exemple l'intérêt de reconstituer le jeu de données plutôt que de simplement enlever les données manquantes.

Bibliographie

CAI, T.T. and HALL, P. (2006). Prediction in functional linear regression. *Annals of Statistics*, **34**, 2159-2179.

CARDOT, H., FERRATY, F. and SARDA, P. (1999). Functional linear model. *Statistics and Probability Letters*, **45**, 11-22.

CRAMBES, C. and HENCHIRI, Y. (2019). Regression imputation in the functional linear model with missing values in the response. *Journal of Statistical Planning and Inference*, **201**, 103-119.

CRAMBES, C. KNEIP, A. and SARDA, P. (2009). Smoothing splines estimators for functional linear regression. *Annals of statistics*, **37**, 35-72.

FERRATY, F. and VIEU, P. (2006). *Nonparametric functional data analysis : Theory and practice*. Springer-Verlag, New York.

HALL, P. and HOROWITZ, J.L. (2007). Methodology and convergence rates for functional linear regression, *The Annals of Statistics*, **35**, 70-91.

HSING, T. and EUBANK, R. (2015). *Theoretical foundations of functional data analysis, with an introduction to linear operators*. Wiley series in probability and statistics, John Wiley & Sons, Chichester.

KNEIP, A and LIEBL, D. (2020). On the optimal reconstruction of partially observed functional data. *Annals of Statistics*, to appear.

RAMSAY, J.O. and SILVERMAN, B.W. (2005). *Functional Data Analysis* (Second edition). Springer-Verlag, New York.

Handling dependence in significance tests of high-dimensional parameter

David Causeur[1]

[1] *Agrocampus Ouest, CNRS, IRMAR - UMR 6625, F-35000 Rennes, France - david.causeur@agrocampus-ouest.fr*

Résumé. Malgré quelques avancées récentes concernant la prise en compte de la dépendance entre les statistiques de test marginales dans les procédures de test portant sur un vecteur de paramètres en grande dimension, il reste pourtant difficile d'en déduire des recommandations claires pour les utilisateurs. Cette problématique, dite de test global, définit un cadre général pour un grand nombre de situations, telles que celles de l'Analyse de la Variance fonctionnelle (fANOVA) et des tests d'association entre une variable réponse phénotypique et une région du génome dans les études d'association à l'échelle du génome (GWAS). Il est intéressant de remarquer que, pour ces deux cas, les méthodes de test les plus utilisées sont basées sur une aggrégation triviale des statistiques de test marginales, ignorant leur dépendance mutuelle. Prendre en compte la dépendance consiste souvent au mieux en des modifications *ad hoc* de méthodes standards, construites sous l'hypothèse d'indépendance. Or, garantir l'efficacité de ces corrections pour un large spectre de formes de dépendance peut réduire la puissance des procédures. Cette observation a motivé l'émergence de nouvelles méthodes s'appuyant sur une prise en compte de la dépendance dans la conception même des statistiques de test, le plus souvent en introduisant une étape préalable de décorrélation des statistiques de test marginales. Dans un premier temps, on montre qu'aucune des deux approches, consistant à ignorer la dépendance ou au contraire à la prendre en compte par décorrélation, n'est uniformément optimale sur l'ensemble des combinaisons entre structure de dépendance et forme du signal d'association. Une nouvelle classe de statistiques de test par aggrégation est proposée, offrant la possibilité d'une décorrélation partielle, adaptée à chacune de ces combinaisons. Les performances du test optimal dans cette classe sont discutées à la fois dans le cadre d'une étude par analyse de la variance fonctionnelle de grands dispositifs de potentiels évoqués (électroencéphalogrammes en réponse à un stimulus) et dans une étude d'association phénotype-génotypes par blocs de marqueurs à l'échelle du génome.

Mots-clés. Analyse de variance fonctionnelle, Décorrélation, Dépendance, GWAS, Grande dimension, Test global.

Abstract. The best way to handle dependence across features when testing a high dimensional parameter has raised many discussions with unclear final recommendations. The former global testing issue arises in a wide scope of applications, including functional Analysis of Variance (fANOVA) and association tests between a region of the genome formed by contiguous Single Nucleotide Polymorphisms (SNP) and a case/control response variable in Genome Wide Association Studies (GWAS). Interestingly, in the two

former fields of applications, many popular methods are just based on simple aggregation of pointwise test statistics ignoring their dependence. Addressing the dependence issue often consists in observing its detrimental impact on the performance of standard methods designed to be optimal under independence, and deduce ad-hoc improvements. To be valid for arbitrarily complex dependence patterns, such approaches can lead to poorly powerful procedures. Therefore, a new generation of methods have emerged, advocating for an adaptive handling of dependence based on a preliminary whitening of the data. After a general discussion on the performance of testing methods ignoring dependence or whitening the pointwise test statistics, we show that none of those two extreme choices is uniformly powerful over the variety of dependence and association patterns. A new class of aggregation methods is therefore introduced, spanning the range between total ignorance of dependence and complete decor- relation. Its performance is discussed both in fANOVA for large Event-Related Potentials (evoked ElectroEncephaloGrams) designs and in SNPset approaches of GWAS.

Keywords. Decorrelation, Dependence, GWAS, Functional ANOVA, Global test, High dimension.

Dependence in high-dimensional global testing

In many research fields, signal detection is viewed as the simultaneous tests of pointwise null hypotheses, e.g. over a time interval in functional Analysis of Variance (fANOVA), over a specific segment of the genome in Genome Wide Association Studies (GWAS) or over a two-dimensional region of an image in functional Magnetic Resonance Imaging (fMRI). In the former situations where the number of features is usually large, sometimes larger than the sample size, such testing issues are generally addressed by deriving a global test statistic for the conjunction of null hypotheses from the aggregation of the corresponding pointwise test statistics. The diversity of existing aggregation methods (see reviews by Zhang and Liang (2014) for fANOVA and Derkach *et al.* (2014) for GWAS issues) reflects the difficulty to identify a method that would show a good detection performance in a wide scope of situations. As reported by Cai *et al* (2014) for the two-group mean comparison issue in high-dimension, the possibly strong dependence across pointwise test statistics turns out to be a crucial point in the comparative studies of aggregation procedures.

However, the most popular whole-interval or whole-region testing methods, both in fANOVA and in GWAS, are based on simple aggregations of pointwise test statistics, not especially designed to be optimal under dependence. For example, Ramsay *et al.* (2009) suggest using the maximum absolute pointwise test statistics, which turns out to be analogous to the `minP` procedure, proposed by Conneely and Boehnke (2007) to test for significant relationship between genotypes of a given set of Single Nucleotide

Polymorphisms (SNPs) and a case/control group membership in the context of GWAS. A functional F-type test statistic based on the squared L_2-norm of the vector of pointwise test statistics is also introduced by Zhang (2013), whereas similar weighted or unweighted L_2-norm statistics are recommended by many authors (Wu *et al.*, 2011, Derkach *et al.*, 2014) for GWAS issues.

The choice of an appropriate method to aggregate pointwise test statistics falls into the general context of global testing as defined by Arias-Castro *et al.* (2011), who especially focuses on the impact of the sparsity rate of the true association signal in a wide variety of correlation patterns on the choice between the L_2-norm based test statistics of standard Analysis of Variance and the Higher Criticism (Donoho and Jin, 2004). The former Higher Criticism (HC) can be viewed as an alternative way of aggregating the pointwise test statistics, by taking a Kolmogorov-Smirnov type distance between the standardized empirical distribution of the pointwise p-values and the theoretical uniform null distribution.

In this general framework, it is commonly observed that, whatever the aggregation method, detection performance for a given association signal can be affected by dependence across pointwise test statistics. A growing number of studies therefore suggests that signal detection procedures can be improved by aggregating decorrelated pointwise test statistics, as for instance in Hall and Jin (2008, 2010) for HC and Ahdesmäki and Strimmer (2010) for the slightly different feature selection issue in two-group classification models. However, as discussed in Bickel and Levina (2004), also in the two-group classification issue or Barnett *et al.* (2017) for global testing using HC in the GWAS context, the potential gain in detection performance that can be expected from decorrelation remains unclear.

Adaptive decorrelation

Both Hébert *et al.* (2020) and Causeur *et al.* (2019) show that, whatever the dependence structure across pointwise test statistics, neither ignoring dependence nor on the contrary fully whitening those marginal test statistics results in a uniformly most powerful global testing procedure over all possible patterns of association between the response and the explanatory variables. Causeur *et al.* (2019) focuses on this dependence issue in functional Analysis of Variance of large designs of Event-Related Potentials (ERP) study, in which time dependence is both strong and more complex than the autoregressive models usually assumed for such data. Some flexibility in the whitening of the pointwise test statistics is introduced by assuming a low-rank factor decomposition of the time correlation matrix, the number of factors being used as a tuning parameter to span the range between ignoring dependence and full decorrelation. For many typical association signal observed in ERP studies, the resulting generalized Likelihood-Ratio test turns out to show improvements with respect to the standard functional ANOVA testing procedures.

In blockwise approach of Genome-Wide Association Studies, whereas the structures of within-block spatial dependence along the genome are rather similar, it is hard to conjecture a general shape, even a sparsity rate, of the association signal between the response and the genetic markers within a SNPset. In the former situation, Hébert *et al.* (2020) propose a new approach, whose aim is to adapt the aggregation of pointwise tests to both the within-block correlation structure and the pattern of the association signal. For that purpose, a general class of test statistics defined as a weighted sum of the squared decorrelated statistics is introduced, which enables a flexible handling of correlation in the aggregation procedure and prevents from the dilution of the signal that can be induced from a complete whitening of the raw pointwise statistics. A closed-form expression of optimal weights is derived by maximizing a Cumulant Generating Function-based distance between the null and non-null distributions of the test statistics.

The presentation will first demonstrate that ignoring dependence can show improvements under some combinations of a dependence pattern and a true association signal with respect to using decorrelation techniques, whereas under other combinations, testing procedures based on whitened test statistics clearly outperform the standard global testing methods. Both the testing approaches proposed in Hébert *et al.* (2020) and Causeur *et al.* (2019) will be introduced as alternatives between ignoring dependence and fully whitening the pointwise test statistics. Based on a large panel of data-driven simulations and illustrations in the context of interdisciplinary research projects, it will demonstrate that the former method provides a good detection performance in a wide variety of situations.

Bibliographie

Ahdesmäki, M., Strimmer, K. (2010). Feature selection in omics prediction problems using cat scores and false nondiscovery rate control. *Ann. Appl. Stat.* 4, pp. 503–519.

Arias-Castro, E., Candès, E.J., Plan, Y. (2011). Global testing under sparse alternatives: ANOVA, multiple comparisons and the higher criticism. *The Annals of Statistics*. 39, 5, pp. 2533–2556. doi:10.1214/11-AOS910. https://projecteuclid.org/euclid.aos/1322663467

Barnett, I., Mukherjee, R., Lin, X. (2017). The generalized higher criticism for testing SNP-set effects in genetic association studies. *Journal of the American Statistical Association*. 112, pp. 64–76.

Bickel, P.J., Levina, E. (2004). Some theory for Fisher's linear discriminant function, 'naive bayes', and some alternatives when there are many more variables than observations. *Bernoulli*. 10, pp. 989–1010.

Cai, T., Liu, W., Xia, Y. (2014). Two-sample test of high dimensional means under dependence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 76, pp. 349–372.

Causeur, D., Sheu, C.-F., Perthame, E. and Rufini, F. (2019). A functional generalized F-test for signal detection with applications to event-related potentials significance analysis. *Biometrics*. 1–11. https://doi.org/10.1111/biom.13118

Conneely, K.N., Boehnke, M. (2007). So many correlated tests, so little time! rapid adjustment of p values for multiple correlated tests. *The American Journal of Human Genetics*. 81, 1158–1168.

Donoho, D., Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Annals of Statistics*. 32, 3, pp. 962–994.

Derkach, A., Lawless, J.F., Sun, L. (2014). Pooled association tests for rare genetic variants: a review and some new results. *Statistical Science*, 29, 2, pp. 302–321.

Hall, P., Jin, J. (2008). Properties of higher criticism under strong dependence. *The Annals of Statistics*. 36, pp. 381–402.

Hall, P., Jin, J. (2010). Innovated higher criticism for detecting sparse signals in correlated noise. *The Annals of Statistics*. 38, pp. 1686–1732.

Hébert F., Causeur, D. and Emily, M. (2020). An Adaptive Decorrelation Procedure for Signal Detection. *Under revision*.

Ramsay, J., Hooker, G., Graves, S. (2009). *Functional Data Analysis with R and MATLAB*. Use R!, Springer New York. URL:https://books.google.fr/books?id=fNKHa8eV7WYC

Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*. 89, pp. 82–93.

Zhang, J.T. (2013). *Analysis of variance for functional data*. CRC Press.

Zhang, J.T., Liang, X. (2014). One-way anova for functional data via globalizing the pointwise F-test. *Scandinavian Journal of Statistics*. 41, pp. 51–71.

Un modèle à blocs stochastiques pour les réseaux multiniveaux

Saint-Clair Chabert-Liddell [†,1] & Pierre Barbillon [†,2] & Sophie Donnet [†,3] & Emmanuel Lazega [⋆,4]

[†] *UMR MIA-Paris, AgroParisTech, INRAe, Université Paris-Saclay, 75005, Paris, France*

[⋆] *Institut d'Études Politiques de Paris, France*

[1] *saint-clair.chabert-liddell@agroparistech.fr*

[2] *pierre.barbillon@agroparistech.fr*

[3] *sophie.donnet@agroparistech.fr*

[4] *emmanuel.lazega@sciencespo.fr*

Résumé. Nous définissons un réseau multiniveau comme la jonction de deux réseaux d'interaction, l'un représentant les interactions entre individus et l'autre les interactions entre organisations. Ces niveaux sont reliés par une relation d'appartenance, chaque individu appartenant à une unique organisation. Le modèle à blocs stochastiques (SBM) est un modèle à variables latentes qui permet de modéliser l'hétérogénéité des connexions d'un réseau en classifiant les nœuds suivants leurs profils de connectivité. Nous étendons le SBM au cas des réseaux multiniveaux et prouvons l'identifiabilité de ce nouveau modèle. Les paramètres de notre modèle sont estimés par des méthodes variationnelles (algorithme VEM) et nous développons un critère de vraisemblance complète intégrée (ICL) pour sélectionner non seulement le nombre de blocs mais également pour décider de la dépendance ou non entre les structures des deux niveaux. Nous justifions via des simulations l'intêret de notre approche ainsi que la robustesse de notre méthode d'estimation de paramètres et de notre critère de sélection de modèle. Nous appliquons notre modèle sur des données collectées lors d'un salon audiovisuel. Le niveau inter-organisationnel représente le réseau des relations économiques entre entreprises et le niveau inter-individuel celui des relations informelles entre leurs représentants sur ce salon.

Mots-clés. Modèle à variables latentes, modèle hiérarchique, réseaux sociaux, inférence variationnelle

Abstract. We define a multilevel network as the junction of two interaction networks, one level representing the interactions between individuals and the other one the interactions between organizations. The levels are linked by an affiliation relationship, each individual belonging to a unique organization. We design a Stochastic block model (SBM) suited to multilevel networks. SBM is a latent variable model for networks, where the connections between nodes depend on a latent clustering (blocks), thus modeling some connection heterogeneity. We prove the identifiability of our model. The parameters of the model are estimated with a variational EM algorithm. An Integrated Completed

Likelihood criterion is developed not only to select the number of blocks but also to detect whether the two levels (individuals and organizations) are dependent or not. In a comprehensive simulation study, we exhibit the benefit of considering our approach, illustrate the robustness of our parameter estimation and highlight the reliability of our model selection criterion. Our approach is applied on a sociological dataset collected during a television program trade fair. The inter-organizational level is the economic network between companies and the inter-individual level is the informal network between their representatives.

Keywords. Latent variable model, Hierarchical modeling, Social networks, Variational inference

1 A multilevel stochastic block model (MLVSBM)

In what follows, a multilevel network is defined as the collection of an inter-individual network, an inter-organizational network and the affiliation of the individuals to the organizations. Besides, we assume that the individuals belong to a unique organization. All the results are given for undirected networks.

Let us consider n_I individuals involved in n_O organizations. We encode the networks into adjacency matrices as follows. Let X^I be the binary $n_I \times n_I$ matrix representing the inter-individual network. X^I is such that : $\forall (i, i') \in \{1, \dots, n_I\}^2$:

$$X^I_{ii'} = \begin{cases} 1 & \text{if there is an interaction from individual } i \text{ to individual } i', \\ 0 & \text{otherwise.} \end{cases}$$

X^O is the binary $n_O \times n_O$ matrix representing the inter-organizational network, $\forall (j, j') \in \{1, \dots, n_O\}^2$:

$$X^O_{jj'} = \begin{cases} 1 & \text{if there is an interaction from organization } j \text{ to organization } j', \\ 0 & \text{otherwise.} \end{cases}$$

Let A be the affiliation matrix. A is a $n_I \times n_O$ matrix such that:

$$A_{ij} = \begin{cases} 1 & \text{if individual i belongs to organization j,} \\ 0 & \text{otherwise} \end{cases}.$$

A is such that $\forall i = 1, \dots, n_I$, $\sum_{j=1}^{n_O} A_{ij} = 1$ since we assume that any individual belongs to a unique organization.

We propose a joint modeling of the inter-individual and inter-organizational networks based on an extension of the stochastic block model (SBM; Snijders and Nowicki, 1997).

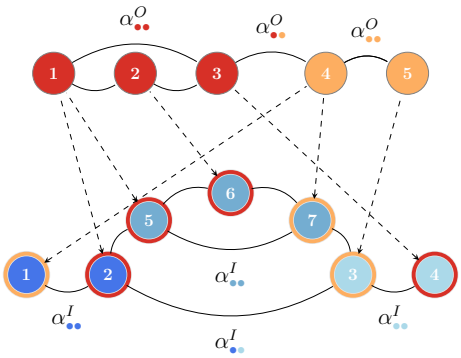


Figure 1: Representation of a multilevel network with inter-organizational level on the top and inter-individual level on the bottom. .

More precisely, assume that the n_O organizations are divided into Q_O blocks and that the n_I individuals are divided into Q_I blocks. Let $Z^O = (Z_1^O, \ldots, Z_{n_O}^O)$ and $Z^I = (Z_1^I, \ldots, Z_{n_I}^I)$ be such that $Z_j^O = l$ if organization j belongs to cluster l ($l \in \{1, \ldots, Q_O\}$) and $Z_i^I = k$ if individual i belongs to cluster k ($k \in \{1, \ldots, Q_I\}$).

Given these clusterings, we assume that the interactions between organizations and interactions between the individuals are independent and distributed as follows:

$$\begin{aligned} \mathbb{P}(X_{jj'}^O = 1 | Z_j^O, Z_{j'}^O) &= \alpha^O_{Z_j^O Z_{j'}^O} \\ \mathbb{P}(X_{ii'}^I = 1 | Z_i^I, Z_{i'}^I) &= \alpha^I_{Z_i^I Z_{i'}^I}. \end{aligned} \tag{1}$$

As a consequence, the blocks gather nodes sharing the same profiles of connectivity.
In order to take into account the fact that organizations may shape the individual behaviors, we assume that the memberships of the individuals (Z^I) depend on the cluster of the organizations (Z^O) they are affiliated to. More precisely, we set:

$$\mathbb{P}(Z_i^I = k | Z_j^O, A_{ij} = 1) = \gamma_{kZ_j^O} \quad \forall i \in \{1, \ldots, n_I\} \quad \forall k \in \{1, \ldots, Q_I\} \tag{2}$$

where γ is a $Q_I \times Q_O$ matrix such that $\sum_{k=1}^{Q_I} \gamma_{kl} = 1$ for any $l \in \{1, \ldots, Q_O\}$. The (Z_j^O) are assumed to be independent random variables distributed as

$$\mathbb{P}(Z_j^O = l) = \pi_l^O, \qquad \forall j \in \{1, \ldots, n_O\} \quad \forall l \in \{1, \ldots, Q_O\} \tag{3}$$

with $\sum_{l=1}^{Q_O} \pi_l^O = 1$.
A small multilevel network is depicted in Figure 1.

Independence. We derive conditions for the structural independence between levels in term of parameters equality.

Proposition 1. *In the MLVSBM, the two following properties are equivalent: [1.]: Z^I is independent on Z^O, [2.]: $\gamma_{kl} = \gamma_{kl'} \quad \forall l, l' \in \{1, \ldots, Q_O\}$ and imply that: [3.]: X^I and X^O are independent.*

Identifiability. We adapt the proof given in Celisse et al. (2012) to obtain the identifiability of the MLVSBM.

Proposition 2. *The MLVSBM is identifiable up to label switching under the following assumptions:*

A1. All coefficients of $\alpha^I \cdot \gamma \cdot \pi^O$ *are distinct and all coefficients of* $\alpha^O \cdot \pi^O$ *are distinct.*

A2. $n_I \geq 2Q_I$ *and* $n_O \geq \max(2Q_O, Q_O + Q_I - 1)$.

A3. At least $2Q_I$ *organizations contain one individual or more.*

Likelihood. From Equations (1), (2) and (3), we derive the complete log-likelihood for a directed MLVSBM where θ denotes all the model parameters:

$$\begin{aligned}
&\log \ell_\theta \left(X^I, X^O, Z^I, Z^O | A\right) \\
&\quad = \log \ell_{\pi^O}(Z^O) + \log \ell_\gamma(Z^I | Z^O, A) + \log \ell_{\alpha^I}(X^I | Z^I) + \log \ell_{\alpha^O}(X^O | Z^O) \\
&\quad = \sum_{j,l} \mathbb{1}_{Z_j^O = l} \log \pi_l^O + \sum_{i,k} \mathbb{1}_{Z_i^I = k} \sum_{j,l} A_{ij} \mathbb{1}_{Z_j^O = l} \log \gamma_{kl} \qquad (4) \\
&\quad + \frac{1}{2} \sum_{i' \neq i} \sum_{k,k'} \mathbb{1}_{Z_i^I = k} \mathbb{1}_{Z_{i'}^I = k'} \log \phi(X_{ii'}^I, \alpha_{kk'}^I) + \frac{1}{2} \sum_{j' \neq j} \sum_{l,l'} \mathbb{1}_{Z_j^O = l} \mathbb{1}_{Z_{j'}^O = l'} \log \phi(X_{jj'}^O, \alpha_{ll'}^O),
\end{aligned}$$

where $\phi(x, a) = a^x (1 - a)^{1-x}$.

2 Statistical Inference, Simulations and Applications

Variational method for maximum likelihood estimation Due to the latent variables, the estimation of the parameters is a complex task. The likelihood of $\mathbf{X} = \{X^I, X^O\}$ $\ell_\theta(\mathbf{X}|A)$ is obtained by integrating out the latent variables $\mathbf{Z} = \{Z^I, Z^O\}$ in the complete data likelihood (4). However, this calculus becomes not computationally tractable as the number of nodes and blocks grow.

The Expectation-Maximisation algorithm (EM) (Dempster et al., 1977) is a popular solution to maximize the likelihood of models with latent variables, but it requires the computation of $\mathbb{P}(\mathbf{Z}|\mathbf{X}, A)$ which is also not tractable in our case. Hence, as Daudin et al. (2008) did for the SBM, we resort to a variational version of the EM algorithm.

The variational EM algorithm aims to maximize a lower bound of $\log \ell_\theta(\mathbf{X}|A)$ by iterating two steps. Step VE maximizes the lower bound with respect to the parameters of an approximate distribution of $\mathbb{P}_\theta(\mathbf{Z}|\mathbf{X}, A)$. Step M maximizes the lower bound with respect to the model parameters θ.

Model selection Following Biernacki et al. (2000) and Daudin et al. (2008), we propose an ad-hoc version of the Integrated Complete Likelihood (ICL) criterion to choose the number of blocks. It is an asymptotic approximation of the complete likelihood integrated over its parameters and latent variables given for the MLVSBM by:

$$ICL(Q_I, Q_O) = \log \ell_{\hat{\theta}}(X^I, X^O, \widehat{Z^I}, \widehat{Z^O} | A, Q_I, Q_O) - pen(Q_I, Q_O),$$

where

$$\begin{aligned} pen(Q_I, Q_O) = & \frac{1}{2} \frac{Q_I(Q_I+1)}{2} \log \frac{n_I(n_I-1)}{2} + \frac{Q_O(Q_I-1)}{2} \log n_I + \\ & \frac{1}{2} \frac{Q_O(Q_O+1)}{2} \log \frac{n_O(n_O-1)}{2} + \frac{Q_O-1}{2} \log n_O \end{aligned}$$

where $\widehat{Z^O}$ and $\widehat{Z^I}$ are the imputed latent variables using the maximum a posteriori (MAP) of $\mathbb{P}_{\hat{\theta}}(\mathbf{Z}|\mathbf{X}, A; Q_I, Q_O)$.

We also use the ICL criterion to assess whether the two levels of interactions are independent or not by comparing the ICL of the MLVSBM with the sum of the ICL of two independent SBMs, one for each level.

We provide a stepwise procedure for model selection which seeks for the optimal number of blocks at a reasonable cost. As a by-product, it states whether the two levels are independent or not. To simulate and infer the MLVSBM, we developed our own `R` package available at https://chabert-liddell.github.io/MLVSBM/.

Simulation studies We study the performances of the inference procedure for the MLVSBM including the ability to recover blocks, the selection of the numbers of blocks and the independence detection.

For $Q_I = Q_O = 3$, on three standard topologies and for different values of density, we set $\gamma_{kk} = \delta$ and $\gamma_{kk'} = .5(1-\delta)$ for $k \neq k'$, $k, k' \in \{1, 2, 3\}$. δ is a parameter for the strength of the dependence between levels ranging from 0 to 1. When $\delta = 1/3$ the levels are independent.

We compare the SBM and the MLVSBM in their abilities to recover the clusters of the inter-individual level (Figure 2 A), the true number of blocks ($Q_I = 3$) and to capture the interdependence between the two levels (Figure 2 B and C).

Application We apply our model to a sociological dataset collected during a television program trade fair. The inter-organizational level is the economic network between companies and the inter-individual level is the informal network between their representatives. Our results exhibit the structural interdependence between the two levels and in particular the heterogeneity of individuals who replicate to different extent their organizations's ties.

A preprint is available at `https://hal.archives-ouvertes.fr/hal-02353711v1`.

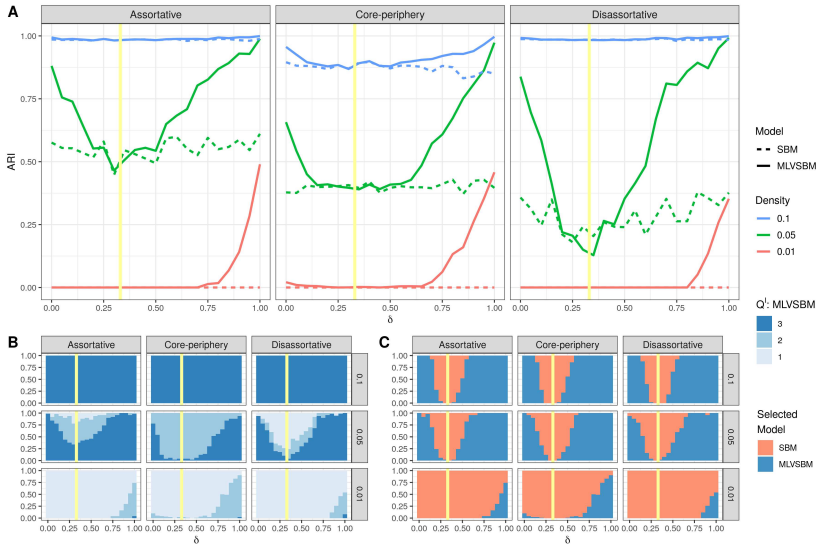


Figure 2: Clustering and model selection accuracy for 3 different topologies and densities on the inter-individual level, as function of δ. The yellow vertical lines corresponds to $\delta = 1/3$ ensuring the independence between the two levels. **A**: ARI (Adjusted Rand Index) for the inter-individual level, comparing the MLVSBM with two independent SBMs. **B**: Number of blocks for the inter-individual level chosen by the ICL criterion for the MLVSBM. **C**: Model selected by the ICL for the inter-level dependence (either MLVSBM or two independent SBMs).

References

Biernacki, C., G. Celeux, and G. Govaert (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE transactions on pattern analysis and machine intelligence 22*(7), 719–725.

Celisse, A., J.-J. Daudin, and L. Pierre (2012). Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electronic Journal of Statistics 6*, 1847–1899.

Daudin, J.-J., F. Picard, and S. Robin (2008). A mixture model for random graphs. *Statistics and computing 18*(2), 173–183.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological) 39*(1), 1–22.

Snijders, T. A. and K. Nowicki (1997). Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of classification 14*(1), 75–100.

Analyse statistique de données anatomiques longitudinales de patients traités. Application au suivi de chimiothérapie

Juliette Chevallier [1] & Stéphanie Allassonnière [2]

[1] *Inria Sophia-Antopolis, Équipe Maassai, Laboratoire J.A. Dieudonné, CNRS, Université Côte d'Azur, juliette.chevallier@inria.fr*

[2] *Centre de Recherche des Cordeliers, Université Paris-Descartes, stephanie.allassonniere@parisdescartes.fr*

Résumé. Les études longitudinales permettent une meilleure compréhension de l'évolution temporelle de phénomènes biologiques ou naturels. Par exemple, le suivi de chimiothérapie repose de plus en plus sur la compréhension de la progression globale de la maladie et non sur des états de santé ponctuels. Les modèles à effets mixtes ont prouvé leur efficacité dans l'étude des données longitudinales, notamment dans le cadre d'applications médicales. Nous proposons ici un modèle non-linéaire à effets mixtes dans lequel les trajectoires d'évolution individuelles sont vues comme des déformations spatio-temporelles d'une trajectoire géodésique par morceaux et représentative de l'évolution de la population. Ce modèle fournit un cadre générique et consistant pour l'étude de données longitudinales à dynamiques multiples.

L'estimation des paramètres du modèle géométrique est réalisée par un estimateur du maximum a posteriori dont nous démontrons l'existence et la consistance sous des hypothèses standards. Numériquement, du fait de la non-linéarité de notre modèle, l'estimation est réalisée par une approximation stochastique de l'algorithme EM, l'algorithme SAEM.

Ce modèle a été développé avec en visée des applications au suivi de chimiothérapie.

Mots-clés. Géométrie riemannienne, données longitudinales, modèles non-linéaires à effets mixtes, algorithmes de type EM, analyse spatio-temporelle

Abstract. Longitudinal studies are powerful tools to better understand the temporal progressions of biological or natural phenomenons. For instance, efforts to monitor chemotherapy rely more and more on the understanding of the global disease progression and not only on punctual states of health. Mixed-effects models have proved their efficiency in the study of longitudinal data sets, especially for medical purposes. We propose here a nonlinear mixed-effects model that allows to estimate a representative piecewise-geodesic trajectory of the global progression and together with spatial and temporal inter-individual variability. This model provides a generic and coherent framework for studying longitudinal manifold-valued data.

Estimation is formulated as a well-defined and consistent maximum a posteriori. Numerically, due to the non-linearity of our model, the estimation of the parameters is performed through a stochastic version of the EM algorithm, namely the SAEM algorithm.

This model have been developed with the idea of an application to chemotherapy monitoring.

Keywords. Riemannian geometry, longitudinal data, nonlinear mixed effect models, EM-like algorithms, spatio-temporal analysis

1 Introduction

L'étude de la variabilité des formes anatomiques a été développée surtout dans le cadre d'études transversales visant à séparer différents types anatomiques à un stade donné de la maladie. Toutefois, il semble plus adapté pour comprendre l'évolution d'une maladie, de comparer l'évolution des formes anatomiques au cours du temps, plutôt que ces formes à un âge fixé ou stade déterminé de la maladie. Cette étude requiert la définition de modèles statistiques pour l'analyse des bases de données longitudinales dans lesquelles les observations de plusieurs individus sont acquises à plusieurs instants.

Pour des données scalaires, comme le volume d'une tumeur par exemple, l'étude de données longitudinales se fait naturellement par l'intermédiaire de modèles linéaires à effets mixtes (Laird and Ware, 1982). Cependant, ces modèles font souvent l'hypothèse que l'origine temporelle du processus observé est connue. Or, notamment dans le cas des maladies, nous sommes amenés à observer des sujets qui sont à des stades de développement différents, et pour lesquels le processus biologique sous-jacent a commencé à des âges variables et inconnus. Cela n'a donc aucun sens de comparer la valeur du volume chez deux sujets aux mêmes âges, comme cela est fait dans le cas des modèles linéaires à effets mixtes "classiques". A l'inverse, nous devons introduire le concept de recalage temporel comme variable cachée du modèle, qui permet de comparer des données à des instants qui correspondent au même stade d'avancement du processus normal ou pathologique.

D'un point de vue statistique, l'introduction de ces nouveaux modèles avec recalage temporel nécessite de revisiter les propriétés bien connues des modèles linéaires à effets mixtes. Un premier modèle non linéaire à effets mixtes qui répond à ces contraintes pour des pathologies de type maladies neuro-dégénératives qui ne font que s'aggraver (comme la maladie d'Alzheimer pour laquelle les traitements ne font que ralentir partiellement les effets) a été proposé par Schiratti et al. (2015). Ce modèle prend également en compte des observations multidimensionnelles vivant sur une variété riemannienne ce qui permet d'appliquer le même modèle à des données de type scalaires, images ou tenseurs. L'estimation des paramètres du modèle (scénario de vieillissement moyen et variations des individus par rapport à cette moyenne) se fait par un algorithme de type Espérance-Maximisation stochastique.

Ce modèle possède intrinsèquement une grande applicabilité mais une hypothèse très forte sur la dynamique du phénomène observé est faite, ce qui en réduit du même coup

la portée : tous les individus sont alignés selon une dynamique unique et supposée monotone, ce qui est en pratique peu réaliste. Mathématiquement, cette hypothèse se traduit par la construction d'une trajectoire représentative géodésique. Il ne permet en particulier pas l'analyse de données anatomiques dans le cas de pathologies traitées telles que les chimiothérapies. En effet, le volume de la tumeur sous l'action de la chimiothérapie peut diminuer, continuer à grossir ou rester stable, les différentes phases de la progression de la maladie se succédant au cours du temps. Dans (Chevallier et al., 2017), nous avons introduit un modèle non-linéaire à effets mixtes permettant l'étude de données longitudinales dont la dynamique d'évolution connaît des phases de progression variables. Après avoir rappelé les fondements de ce modèle, nous illustrons son applicabilité au suivi de chimiothérapie et plus particulièrement à la prise en charge du cancer du rein métastatique en nous appuyant sur des données de l'Hôpital Européen Georges Pompidou (HEGP).

2 Un modèle générique pour l'étude de données longitudinales à valeurs sur des variétés riemanniennes

Soit un jeu de données obtenu par l'observation, pour chaque individu i, de k_i mesures multivariées $y_i = (y_{i,j})_{j\in[\![1,k_i]\!]} \in \mathbb{R}^d$ aux temps $t_i = (t_{i,j})_{j\in[\![1,k_i]\!]} \in \mathbb{R}$.

2.1 Un modèle spatio-temporel à effets mixtes

En se basant sur les travaux de Schiratti et al. (2015), nous avons proposé un modèle non-linéaire à effets mixtes pour l'étude de données longitudinales à dynamiques d'évolution multiples (Chevallier et al., 2017). Ce modèle repose sur la discrimination de déformations temporelles, liées à l'acquisition des données et au rythme de progression du phénomène observé, de déformations spatiales, liées à la géométrie intrinsèque des formes observées.

Plus précisément, à considérer que l'on observe des échantillons bruités le long de trajectoires d'évolution individuelles, on suppose que ces trajectoires dérivent du scénario représentatif de l'évolution au niveau macroscopique par des transformations spatio-temporelles. Contrairement au modèle de Schiratti et al. (2015), la trajectoire représentative de l'évolution de la population n'est plus supposée géodésique mais géodésique par morceaux. Autrement dit, étant donné une subdivision $t_R = \left(-\infty < t_R^1 < \ldots < t_R^{m-1} < +\infty\right)$ de $\mathbb{R}$ on construit la trajectoire γ_0 représentative de l'évolution de la population en posant

$$\forall t \in \mathbb{R}, \quad \gamma_0(t) \;=\; \gamma_0^1(t)\,\mathbb{1}_{]-\infty,t_R^1]}(t) \;+\; \sum_{\ell=2}^{m-1} \gamma_0^\ell(t)\,\mathbb{1}_{]t_R^{\ell-1},t_R^\ell]}(t) \;+\; \gamma_0^m(t)\,\mathbb{1}_{]t_R^{m-1},+\infty[}(t)\,,$$

où chacun des tronçons γ_0^ℓ est construit de manière à être géodésique sur le segment $]t_R^{\ell-1}, t_R^\ell]$ correspondant. Les différents t_R^ℓ sont appelés temps de rupture (de la dynamique

d'évolution) et on impose un recollement au moins continue de la trajectoire d'évolution en ceux-ci. En construisant une trajectoire représentative géodésique par morceaux, on conserve la grande applicabilité du modèle de Schiratti et al. (2015). En effet, une telle hypothèse facilite la paramétrisation de cette dernière, indépendamment du type de données étudiées et, de fait, l'estimation numérique.

À partir de cette trajectoire représentative, on construit pour chacun des individus i une trajectoire d'évolution individuelle en posant

$$\forall t \in \mathbb{R}, \quad \gamma_i(t) \;=\; \gamma_i^1(t)\,\mathbb{1}_{]-\infty,t_{R,i}^1]}(t) \;+\; \sum_{\ell=2}^{m-1} \gamma_i^\ell(t)\,\mathbb{1}_{]t_{R,i}^{\ell-1},t_{R,i}^\ell]}(t) \;+\; \gamma_i^m(t)\,\mathbb{1}_{]t_{R,i}^{m-1},+\infty[}(t)\,,$$

où les γ_i^ℓ sont des déformations spatio-temporelles de la courbe γ_0^ℓ associée. Les déformations temporelles consistent en des reparamétrisations temporelles affines, autorisant ainsi chacun des individus, mais aussi chacune des différentes phases d'évolution de la maladie, à progresser selon leur propre rythme. Autrement dit, on considère des déformations temporelles de la forme

$$\psi_i\colon\; t \mapsto \alpha_i^\ell(t - t_R^{\ell-1} - \tau_i^\ell) + t_R^{\ell-1}\,, \quad \text{où} \quad (\alpha_i^\ell, \tau_i^\ell) \in \mathbb{R}^+ \times \mathbb{R}\,,$$

et où, pour garantir la continuité des trajectoires individuelles, on impose que $\psi_i^\ell(t_{R,i}^{\ell-1}) = t_R^{\ell-1}$. Concernant les déformations spatiales, la seule contrainte mathématique est la continuité des trajectoires d'évolution. Leur construction est induite par le type de données étudiées et la question pratique considérée.

Finalement, nos observations étant considérées comme des échantillons bruités le long des différentes courbes γ_i, on écrit

$$\forall (i,j) \in [\![1,n]\!] \times [\![1,k_i]\!], \quad y_{i,j} \;=\; \gamma_i(t_{i,j}) \;+\; \varepsilon_{i,j}\,, \quad \text{où} \quad \varepsilon_{i,j} \sim \mathcal{N}(0, \sigma^2 I_d)\,.$$

2.2 Estimation par maximum a posteriori

Afin de garantir une certaine efficacité numérique, notre modèle est supposé paramétrique : autrement dit, on suppose que la trajectoire d'évolution γ_0 (resp. γ_i) peut être entièrement décrite par un nombre fini de variables $z_{\text{pop}} \in \mathbb{R}^{p_{\text{pop}}}$ (resp. $z_i \in \mathbb{R}^{p_{\text{ind}}}$). Nous souhaitons réaliser l'estimation des paramètres de notre modèle par le biais d'une approximation stochastique de l'algorithme EM : l'algorithme SAEM qui a déjà fait ses preuves pour ce type de problématiques (Lavielle and Mentré, 2007). Pour autant la convergence de l'algorithme SAEM n'a été démontrée que dans le cas des familles exponentielles (Allassonnière et al., 2010; Delyon et al., 1999). On recourt donc à l'astuce de Kuhn and Lavielle (2005) et on interprète les variables z_{pop} comme des gaussiennes très piquées dont on estime la moyenne. De plus, on suppose les variables z_i comme dérivant de gaussiennes centrées dont on estime la matrice de covariance, ceci afin de proposer un modèle génératif.

En résumé, notre modèle s'écrit hiérarchiquement :

$$y \,|\, z, \theta \sim \bigotimes_{i=1}^{n} \bigotimes_{j=1}^{k_i} \mathcal{N}\big(\gamma_i(t_{i,j}) \,,\, \sigma^2\big) \quad \text{et} \quad z \,|\, \theta \sim \mathcal{N}\big(\overline{z_{\text{pop}}} \,,\, D_{\text{pop}}^{-1}\big) \bigotimes_{i=1}^{n} \mathcal{N}(0, \Sigma) \,,$$

où $\theta = (\overline{z_{\text{pop}}}, \Sigma, \sigma) \in \mathbb{R}^{p_{\text{pop}}} \times \mathscr{S}^+_{p_{\text{ind}}}(\mathbb{R}) \times \mathbb{R}^+$.

Dans (Chevallier et al., 2019), nous avons démontré la consistance de l'estimateur du maximum a posteriori (MAP), en ne supposant en particulier *pas* que les observations sont distribuées selon le "bon" modèle mais seulement selon une distribution à décroissance polynomiale en dehors d'un compact. Nous avons également démontré l'existence du MAP, à condition d'imposer des *a priori* de type inverse Wishart pour les variances Σ et σ.

3 Application au suivi de chimiothérapie

La méthode décrite ci-dessus étant très générique, elle peut être appliquée à un grand nombre de situations. Nous appliquons ce modèle au suivi de chimiothérapie et plus particulièrement au suivi du cancer métastatique du rein. En effet, dans ce contexte, la compréhension du rythme de progression du cancer est au cœur de la prise en charge médicale.

Les patients atteints du cancer du rein métastatique prennent un traitement quotidien et viennent régulièrement à l'hôpital contrôler la l'évolution de leur(s) tumeur(s). Au cours des dernières années, la prise en charge médicale pour ce type de pathologies a profondément changé : une nouvelle classe de médicaments antiangiogéniques ciblant les vaisseaux tumoraux plutôt que les cellules tumorales a fait son apparition, permettant ainsi de multiplier par trois le taux de survis des patients (Escudier et al., 2016). Cependant, ces nouveaux traitements ne guérissent pas le cancer et ne visent qu'à retarder la croissance tumorale, ce qui nécessite le recours à différentes thérapies successives, qui se doivent d'être poursuivies ou interrompues au moment le plus opportun pour le patient, en fonction de sa réponse au traitement considéré. Ce nouveau protocole médical est donc à la fois un nouveau défi scientifique : comment choisir la pharmacothérapie la plus efficace compte tenu des premiers temps d'observation du profil de réponse d'un patient donné ?

3.1 Modèle logistique par morceaux : Suivi de chimiothérapie par le biais dux score RECIST

La première application concerne le suivi de scores RECIST pour response evaluation criteria in solid tumors en anglais. Ces scores étant des données scalaires, on réalise une instanciation du modèle générique pour des données réelles bornées en se plaçant sur le segment $[0, 1]$ munit de la métrique logistique. Ce modèle a été élaboré en collaboration avec des oncologues et radiologues de l'Hôpital Européen Georges Pompidou (HEGP). Des expériences numériques sur données réelles et synthétiques en valident la pertinence.

3.2 Modèle géodésique par morceaux pour les formes : Suivi de chimiothérapie à travers les formes anatomiques

La seconde application porte sur le suivi de formes anatomiques 3D, toujours pour l'évaluation de la réponse tumorale. Ce modèle repose sur la notion de grandes déformations et s'applique aussi bien aux courants (Vaillant and Glaunès, 2005) qu'aux varifolds (Charon and Trouvé, 2013), qui sont des espaces de forme standards pour l'analyse de formes anatomiques. Des expériences numériques ont également été conduites dans ce cadre.

4 Conclusion et perspectives

Nous avons proposé un modèle non linéaire à effets mixtes pour l'analyse statistique de données longitudinales à dynamiques multiples et à valeurs sur des variétés riemanniennes. Ce modèle a été conçu avec en tête des applications à l'anatomie computationnelle.

En particulier, nous avons appliqué ce modèle au suivi de chimiothérapie, situation typique dans laquelle la dynamique d'évolution est amenée à changer. En effet, la mise en place d'un nouveau traitement induit généralement trois phases d'évolution distinctes pour un même patient : une phase de réponse au traitement dans laquelle la taille de ses tumeurs va décroître, une phase dite stable où la taille des tumeurs reste inchangée et, dans la plupart des cas, une phase de progression de la maladie dans laquelle les tumeurs vont de nouveau croître, ce qui nécessite la mise en place d'un nouveau traitement dans les délais les plus brefs. Pouvoir estimer avec précision le temps d'échappement au traitement est donc crucial dans ce contexte.

Plus précisément, dans le cadre du suivi chimiothérapeutique, nous avons proposé deux instanciations du modèle générique : d'une part, pour le suivi de scores RECIST et, d'autre part, pour le suivi de formes anatomiques en 3 dimension que l'on apparente à des tumeurs segmentées. Ce premier modèle pour les scores est le fruit d'une collaboration avec des oncologues et radiologues de l'HEGP.

Cependant, et malgré son caractère très générique, le modèle tel que proposé en l'état ne permet pas l'étude de populations dans lesquelles les comportements de certaines sous-populations diffèrent. Pour cela, il conviendrait, en se basant sur des modèles de mélanges usuels, d'introduire une sur-couche dans la modélisation que nous proposons ici afin d'aboutir à un modèle de mélange pour l'étude statistique de données longitudinales à valeurs sur des variétés riemanniennes. Les travaux récents de Debavelaere et al. (2019) vont dans ce sens.

Références

Stéphanie Allassonnière, Estelle Kuhn, and Alain Trouvé. Construction of Bayesian deformable models via a stochastic approximation algorithm : A convergence study. *Bernoulli*, 16(3) :641–678, 2010.

Nicolas Charon and Alain Trouvé. The varifold representation of nonoriented shapes for diffeomorphic registration. *SIAM Journal on Imaging Sciences*, 6(4) :2547–2580, 2013.

Juliette Chevallier, Stéphane Oudard, and Stéphanie Allassonnière. Learning spatiotemporal piecewise-geodesic trajectories from longitudinal manifold-valued data. In *Neural Information Processing Systems*, Long Beach, CA, USA, 2017.

Juliette Chevallier, Vianney Debavelaere, and Stéphanie Allassonnière. A coherent framework for learning spatiotemporal piecewise-geodesic trajectories from longitudinal manifold-valued data. Preprint, Hal-01646298, 2019.

Vianney Debavelaere, Alexandre Bône, Stanley Durrleman, and Stéphanie Allassonnière. Clustering of longitudinal shape data sets using mixture of separate or branching trajectories. to appear in MICAI 2019, Hal-02103355, 2019.

Bernard Delyon, Marc Lavielle, and Eric Moulines. Convergence of a stochastic approximation version of the EM algorithm. *The Annals of Statistics*, 27(1) :94–128, 1999.

Bernard Escudier, Camillo Porta, Mélanie Schmidinger, Nathalie Rioux-Leclercq, Axel Bex, Vincent S. Khoo, Viktor Gruenvald, and Alan Horwich. Renal cell carcinoma : ESMO clinical practice guidelines for diagnosis, treatment and follow-up. *Annals of Oncology*, 27(suppl 5) :v58–v68, 2016.

Estelle Kuhn and Marc Lavielle. Maximum likelihood estimation in nonlinear mixed effects models. *Computational Statistics & Data Analysis*, 49(4) :1020–1038, 2005.

Nan M. Laird and James H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38(4) :963–974, 1982.

Marc Lavielle and France Mentré. Estimation of population pharmacokinetic parameters of saquinavir in HIV patients with the MONOLIX software. *Journal of pharmacokinetics and pharmacodynamics*, 34(2) :229–249, 2007.

Jean-Baptiste Schiratti, Stéphanie Allassonnière, Olivier Colliot, and Stanley Durrleman. Learning spatiotemporal trajectories from manifold-valued longitudinal data. In *Neural Information Processing Systems*, number 28 in Advances in Neural Information Processing Systems, Montréal, Canada, 2015.

Marc Vaillant and Joan Glaunès. Surface matching via currents. In *Information Processing in Medical Imaging*, volume 3565, Glenwood Springs, USA, 2005.

Estimation spectrale du processus de Hawkes : alpha-mélange et théorème central limite

Felix Cheysson [1] & Gabriel Lang [2]

[1] *AgroParisTech, 16 rue Claude Bernard 75005 Paris, felix.cheysson@agroparistech.fr*
[2] *AgroParisTech, 16 rue Claude Bernard 75005 Paris, gabriel.lang@agroparistech.fr*

Résumé. Les processus de Hawkes sont une famille de processus stochastiques pour lesquels l'occurrence d'un événement modifie temporairement la probabilité d'occurrence des événements futurs. Nous nous intéressons à la série temporelle générée par le comptage des événements du processus de Hawkes stationnaire. À partir des propriétés de "cluster" du processus, nous établissons une borne supérieure sur le coefficient d'alpha-mélange de sa série de comptage, sous l'hypothèse d'existence de moments du noyau de reproduction. Cette borne permet d'établir un théorème central limite pour un estimateur spectral du processus de Hawkes à partir de données de comptages. Des jeux de simulations illustrent le théorème central limite et les performances de l'estimation, en particulier des paramètres de reproduction du processus.

Mots-clés. Processus de Hawkes, Alpha-mélange, Estimation de Whittle, Série temporelle

Abstract. Hawkes processes are a family of stochastic processes for which the occurrence of any event increases the probability of further events occurring. In this talk, we study the time series generated by the event counts of the stationary Hawkes process. Using its cluster properties, we derive an upper bound for the count series' alpha-mixing coefficient provided mild conditions on the moments of the reproduction kernel. This result leads to a central limit theorem for the estimation of the parameters of the Hawkes process from its count data, using a spectral approach derived from Whittle's likelihood. Simulated datasets illustrate the performances of the estimation, notably, of the Hawkes reproduction mean and kernel, even with relatively large binsizes.

Keywords. Hawkes process, Alpha-mixing, Whittle estimation, Time series

1 Introduction

Les processus auto-excitants, éponymes de Hawkes [8], forment une famille de processus stochastiques pour lesquels l'occurrence d'un événement modifie temporairement la probabilité d'occurrence d'événements futurs. Par ailleurs, les processus de Hawkes présentent des propriétés d'agrégation : ce sont des processus d'agrégats, pour lesquels chaque agrégat est un arbre de Galton-Watson sous-critique avec une loi de reproduction Poissonnienne

[9]. Grâce à ces deux propriétés, d'auto-excitation et d'agrégation, attrayantes pour la modélisation, l'utilisation des processus de Hawkes s'est répandue à nombreuses disciplines, d'abord presque exclusivement en séismologie [1, 13], puis en neurophysiologie [3], en finance [2], en génomique [15] et en épidémiologie [11].

Lorsque les temps d'occurrence des événements sont observés, l'estimation des processus de Hawkes a été étudiée d'abord par des approches spectrales [1], puis par maximum de vraisemblance [12, 13, 14]. Cependant, dans le cas où le comptage des événements est observé en temps discret (*i.e.* la ligne temporelle est partitionnée en intervalles disjoints et le nombre d'événements est compté sur chaque intervalle), il n'est pas possible d'appliquer ces méthodes directement.

Au cours de cette présentation, nous revisitons l'approche spectrale d'Adamopoulos à l'estimation paramétrique de processus de Hawkes. À partir du spectre de Bartlett du processus (*i.e.* la densité spectrale de la mesure de covariance du processus), Adamopoulos proposa comme estimateur le minimiseur de la log-vraisemblance spectrale, introduite par Whittle [17]. Nous étendons ces résultats aux processus dont le comptage des événements est observé en temps discret uniquement.

Rosenblatt [16] introduisit le coefficient d'alpha-mélange afin de formaliser une mesure de dépendance entre variables aléatoires, coefficient qui intervient dans la preuve de nombreuses inégalités de moment et de théorèmes central limite [5]. Dans cette présentation, nous établissons une borne supérieure sur le coefficient d'alpha-mélange du processus de Hawkes et de sa série de comptage. Ces résultats nous permettent de proposer un théorème central limite pour l'estimateur spectral des paramètres du processus, à partir des travaux de Dzhaparidze [6] sur la méthode de Whittle. Enfin, nous illustrons ce théorème par des simulations de processus de Hawkes.

2 Le processus de Hawkes

L'*intensité conditionelle* $\lambda(\cdot)$ d'un processus ponctuel N peut être définie (*cf.* [4] pour une définition rigoureuse) par la limite, si elle existe, de l'espérance du nombre d'événements dans un intervalle $(t, t+\Delta]$ dont la taille tend vers zéro :

$$\lambda(t) = \lim_{\Delta \to 0} \ \Delta^{-1} \, \mathbb{E}\left[N\big((t, t+\Delta]\big) \mid \mathcal{F}_t \right],$$

où $\mathcal{F}_t$ est la *filtration naturelle* associée à N et représente l'information disponible jusqu'au temps t.

Le processus de Hawkes sur la ligne temporelle $\mathbb{R}$ est un processus ponctuel N dont l'intensité conditionelle $\lambda(t)$ dépend des événements $\{T_i\}$ du processus par la relation suivante :

$$\begin{aligned} \lambda(t) &= \eta(t) + \int h(t-u) N(\mathrm{d}u) \\ &= \eta(t) + \sum h(t - T_i) \end{aligned}$$

où $\int \phi(u)N(\mathrm{d}u) = \sum \phi(T_i)$ désigne l'intégrale stochastique de Stieltjes. La constante $\eta > 0$ est appelée *intensité d'immigration* et la fonction mesurable $h : \mathbb{R} \to \mathbb{R}_+$ *noyau de reproduction*.

Par la suite, on appellera série de Hawkes une série temporelle du type $\{X_k\}_{k\in\mathbb{Z}} = \{N(k\Delta, (k+1)\Delta]\}_{k\in\mathbb{Z}}$, générée par le comptage du processus sur des intervalles de taille Δ (voir figure 1).

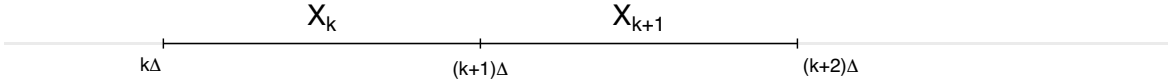


Figure 1: Une série de Hawkes $\{X_k\}_{k\in\mathbb{Z}}$ avec un pas d'observation Δ.

3 Propriétés d'alpha-mélange

Rappelons que, pour une série temporelle donnée $(X_k)_{k\in\mathbb{Z}}$, le coefficient d'alpha-mélange est défini par la mesure suivante [16] :

$$\alpha_X(r) \coloneqq \sup \left\{ |\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)| : -\infty < k < \infty, A \in \mathcal{F}_{-\infty}^{k}, B \in \mathcal{F}_{k+r}^{\infty} \right\},$$

où $\mathcal{F}_a^b$ est la tribu générée par $(X_k)_{a\leq k\leq b}$.

La série (X_k) est dite alpha-mélangeante si $\alpha_X(r) \to 0$ as $r \to \infty$. Intuitivement, l'alpha-mélange signifie que le passé et le futur du processus sont asymptotiquement indépendants.

Citons maintenant le résultat important de la présentation :

Théorème 1. *Soit N un processus de Hawkes sur $\mathbb{R}$ avec noyau de reproduction h, tel que $\int_{\mathbb{R}} h < 1$, et $(X_k)_{k\in\mathbb{Z}} = \big(N(k, k+1]\big)_{k\in\mathbb{Z}}$ sa série de comptage. Supposons qu'il existe $\beta > 0$ tel que le noyau de reproduction h possède un moment fini d'ordre $1+\beta$:*

$$\nu_{1+\beta} \coloneqq \int_{\mathbb{R}} t^{1+\beta} h(t)\mathrm{d}t < \infty.$$

Alors le coefficient d'alpha-mélange de (X_k) vérifie

$$\alpha_X(r) = \mathcal{O}\left(\frac{1}{r^\beta}\right).$$

La preuve de ce résultat s'appuie sur la structure d'agrégat du processus : par indépendance des agrégats, nous nous plaçons à l'échelle d'un unique arbre de Galton-Watson, puis utilisons ses propriétés pour calculer une borne supérieure sur le coefficient d'alpha-mélange de l'arbre. L'idée de la preuve est que, puisque le processus de Galton-Watson s'éteint presque surement et que le noyau de reproduction h possède un moment fini, alors la probabilité qu'il existe un point de l'arbre de génération k loin de la racine tend rapidement vers zéro lorsque k augmente.

4 Estimation des séries de comptage

Pour une série stationnaire (X_k), possédant une densité spectrale $f_\theta(\cdot)$, avec θ un vecteur de paramètres inconnus, Hosoya [10] et Dzhaparidze [7], à partir des travaux de Whittle [17], proposèrent comme estimateur de θ le minimiseur

$$\widehat{\theta}_n = \arg\min_{\theta\in\Theta} \mathcal{L}_n(\theta)$$

où

$$\mathcal{L}_n(\theta) = \frac{1}{4\pi}\int_{-\pi}^{\pi}\left(\log f_\theta(\omega) + \frac{I_n(\omega)}{f_\theta(\omega)}\right)\mathrm{d}\omega$$

est la log-vraisemblance spectrale du processus et $I_n(\omega) = (2\pi n)^{-1}\left|\sum_{k=1}^n X_k\, e^{-ik\omega}\right|^2$ est le périodogramme de $(X_k)_{1\leq k\leq n}$. Ils établirent également les propriétés asymptotiques de cet estimateur sous des conditions de régularité appropriées.

Dzhaparidze [6] étendit ces résultats à des cas plus généraux, et en particulier aux processus stationnaires alpha-mélangeants. Le théorème suivant est donc une adaptation de celui de Dzhaparidze [6, Theorem II.7.2] pour les processus de Hawkes stationnaires.

Théorème 2. *Soit N un processus de Hawkes comme défini dans le Théorème 1, et $(X_k)_{k\in\mathbb{Z}} = \big(N(k,k+1]\big)_{k\in\mathbb{Z}}$ sa série de comptage de fonction de densité spectrale f_θ. Supposons qu'il existe $\beta > 0$ tel que le noyau de reproduction h possède un moment fini d'ordre $2+\beta$. Alors, sous des conditions de régularité adéquates sur f_θ, l'estimateur $\widehat{\theta}_n$ est consistant et asymptotiquement normal, et*

$$\mathrm{Var}\left(\widehat{\theta}_n - \theta_0\right) \underset{n\to\infty}{=} \mathcal{O}\left(n^{-1}\right)$$

où θ_0 désigne le vrai vecteur de paramètres.

5 Simulations

Considérons un processus de Hawkes stationnaire avec fonction de reproduction exponentielle :

$$\lambda(t) = \eta + \mu\int \beta e^{-\beta(t-u)} N(\mathrm{d}u),$$

i.e. avec noyau de reproduction $h(t) = \mu\beta e^{-\beta t}$ pour $t \geq 0$. Notons que le processus vérifie les hypothèses du Théorème 2.

Nous avons simulé $S = 1.000$ réalisations du processus de Hawkes sur l'intervalle $[0, T]$ avec pour paramètres $\eta = 1$, $\mu = 0.5$ and $\beta = 1$. Pour chacune des simulations, nous avons créé quatre séries temporelles en comptant le nombre d'événements dans des intervalles de taille 0,25, 0,5, 1 et 2 respectivement. Nous avons ensuite estimé les paramètres η, μ et β comme dans la Section 4 pour chacune des quatres séries, et calculé l'erreur quadratique

moyenne, définie par MSE $= S^{-1}\sum(\widehat{\theta}_n - \theta_0)^2$, pour les estimations de chaque ensemble de $S = 1.000$ simulations pour chaque T et taille d'intervalle donnés. Nous comparons cette approche spectrale au maximum de vraisemblance (Figure 2) : comme le maximum de vraisemblance utilise toute l'information disponible sur la position des événements, ses estimateurs sont meilleurs que n'importe quel estimateur basé sur les séries de comptage, et nous informent donc sur le meilleur des cas pour les estimateurs spectraux, *i.e.* quand la taille de l'intervalle tend vers zéro.

On observe que, pour T élevé, la pente de l'erreur quadratique moyenne atteint -1 pour tous les paramètres et presque toutes les tailles d'intervalle, illustrant le taux de convergence $\mathcal{O}(n^{-1})$ du Théorème 2. Par ailleurs, pour des tailles d'intervalle raisonnables (≤ 1), les estimateurs spectraux du taux de reproduction μ ont un MSE comparable à celui du maximum de vraisemblance.

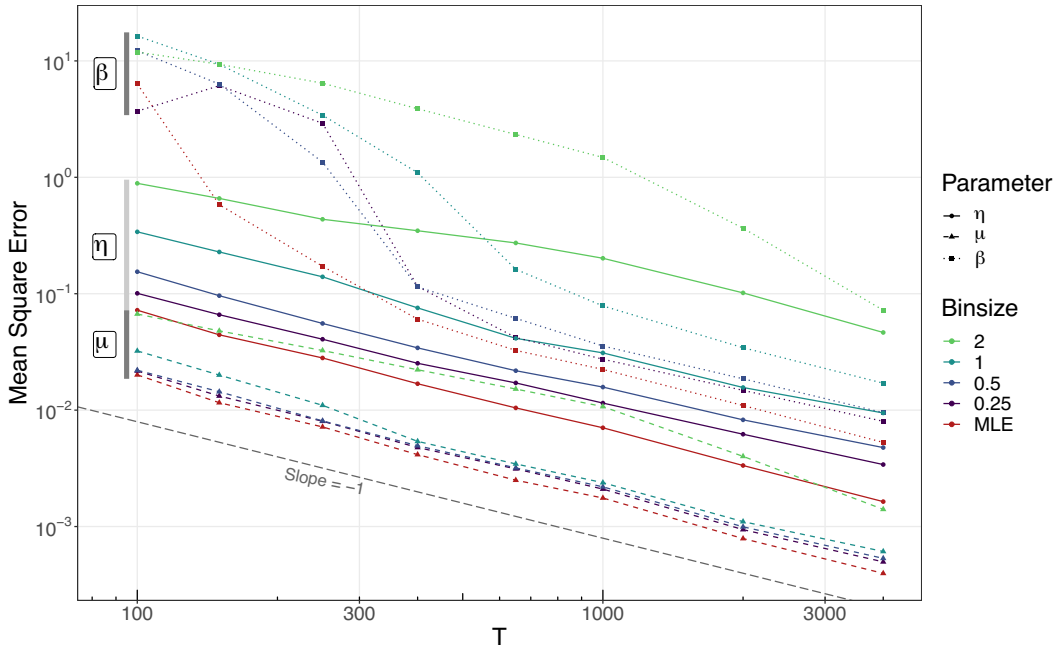


Figure 2: Erreur quadratique moyenne des estimations des paramètres η, μ et β pour 1.000 simulations du processus de Hawkes stationnaire avec noyau de reproduction $h(t) = \mu\beta e^{-\beta t}$ sur l'intervalle $[0, T]$, en échelle log-log. La ligne pointillée grise représente la pente idéale de -1, *i.e.* une vitesse de convergence de $\mathcal{O}(n^{-1})$.

References

[1] Adamopoulos, L. (1976). Cluster models for earthquakes: Regional comparisons. *Journal of the International Association for Mathematical Geology*, 8(4):463–475.

[2] Bacry, E., Mastromatteo, I., and Muzy, J.-F. (2015). Hawkes Processes in Finance. *Market Microstructure and Liquidity*, 01(01):1550005.

[3] Chornoboy, E. S., Schramm, L. P., and Karr, A. F. (1988). Maximum likelihood identification of neural point process systems. *Biological Cybernetics*, 59(4-5):265–275.

[4] Daley, D. J. and Vere-Jones, D. (2003). *An Introduction to the Theory of Point Processes, Volume I: Elementary Theory and Methods*. Probability and its Applications. Springer, New York.

[5] Doukhan, P. (1994). Mixing: Properties and Examples. *LECTURE NOTES IN STATISTICS -NEW YORK- SPRINGER VERLAG-*, 1(85):142 / 82.

[6] Dzhaparidze, K. (1986). *Parameter Estimation and Hypothesis Testing in Spectral Analysis of Stationary Time Series*, volume 27 of *Springer Series in Statistics*. Springer New York, New York, NY.

[7] Dzhaparidze, K. O. (1974). A New Method for Estimating Spectral Parameters of a Stationary Regular Time Series. *Theory of Probability & Its Applications*, 19(1):122–132.

[8] Hawkes, A. G. (1971). Spectra of Some Self-Exciting and Mutually Exciting Point Processes. *Biometrika*, 58(1):83–90.

[9] Hawkes, A. G. and Oakes, D. (1974). A cluster process representation of a self-exciting process. *Journal of Applied Probability*, 11(03):493–503.

[10] Hosoya, Y. (1974). *Estimation problems on stationary time series models*. Ph.d. dissertation, Yale University.

[11] Meyer, S., Elias, J., and Höhle, M. (2012). A Space-Time Conditional Intensity Model for Invasive Meningococcal Disease Occurrence. *Biometrics*, 68(2):607–616.

[12] Ogata, Y. (1978). The asymptotic behaviour of maximum likelihood estimators for stationary point processes. *Annals of the Institute of Statistical Mathematics*, 30(1):243–261.

[13] Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association*, 83(401):9–27.

[14] Ozaki, T. and Ogata, Y. (1979). Maximum likelihood estimation of Hawkes' self-exciting point processes. *Annals of the Institute of Statistical Mathematics*, 31(1):145–155.

[15] Reynaud-Bouret, P. and Schbath, S. (2010). Adaptive estimation for hawkes processes; Application to genome analysis. *Annals of Statistics*, 38(5):2781–2822.

[16] Rosenblatt, M. (1956). a Central Limit Theorem and a Strong Mixing Condition. *Proceedings of the National Academy of Sciences*, 42(1):43–47.

[17] Whittle, P. (1952). Some results in time series analysis. *Scandinavian Actuarial Journal*, 1952(1-2):48–60.

Utilisation de régressions *loess*, spline et monotone sur des données de protéomique quantitative de type *data-independent acquisition*

Marie Chion [1], Joanna Bons [2], Myriam Maumy-Bertrand [3], Christine Carapito [4] & Frédéric Bertrand [5]

[1] *Institut de Recherche Mathématique Avancée, UMR 7501, 7 rue René Descartes, 67084 Strasbourg Cedex et Laboratoire de Spectrométrie de Masse Bio-Organique, IPHC, UMR 7178, 25 rue Becquerel, 67087 Strasbourg Cedex, chion@math.unistra.fr*

[2] *Laboratoire de Spectrométrie de Masse Bio-Organique, IPHC, UMR 7178, 25 rue Becquerel, 67087 Strasbourg Cedex, joanna.bons@etu.unistra.fr*

[3] *Institut de Recherche Mathématique Avancée, UMR 7501, 7 rue René Descartes, 67084 Strasbourg Cedex, mmaumy@math.unistra.fr*

[4] *Laboratoire de Spectrométrie de Masse Bio-Organique, IPHC, UMR 7178, 25 rue Becquerel, 67087 Strasbourg Cedex, ccarapito@unistra.fr*

[5] *Laboratoire de Modélisation et Sûreté des Systèmes, Institut Charles Delaunay, Université de Technologie de Troyes, 12 Rue Marie Curie, 10300 Troyes Cedex, frederic.bertrand@utt.fr*

Résumé. L'analyse protéomique consiste à étudier l'ensemble des protéines contenues dans un système biologique donné, à un instant donné et dans des conditions données. Les méthodes de quantification globale permettent d'obtenir des informations quantitatives relatives pour l'ensemble des protéines détectées dans la série d'échantillons. Les méthodes de quantification ciblée permettent, en introduisant des standards synthétiques marqués correspondant aux peptides d'intérêt et préalablement sélectionnés dans l'échantillon biologique considéré, de connaître précisément la quantité d'une (de) protéine(s) spécifique(s) dans celui-ci. Une approche récente, appelée *Data-Independent Acquisition*, permet de combiner ces deux méthodes en une seule analyse. À partir des données d'intensité et de quantité obtenues en quantification ciblée, nous proposons d'ajuster et de comparer des modèles de régression *loess*, spline et monotone expliquant la quantité d'un peptide par son intensité dans l'échantillon considéré. Ces modèles nous permettent d'estimer les quantités de l'ensemble des peptides détectés grâce à l'utilisation des standards internes marqués pour un sous-ensemble des peptides.

Mots-clés. Régression *loess*, splines, régression monotone, prédiction et données protéomiques.

Abstract. Proteomic analysis consists in studying proteins from a given biological system, at a given time and under given conditions. Global quantification methods allow obtaining relative quantitative information on all proteins accross the sample series. Targeted quantification methods allow, by introducing labelled synthetic standards corresponding to previously selected peptides of interest, to know precisely the quantity of

a specific protein(s) in the biological sample considered. A recent approach, called Data-Independent Acquisition, makes it possible to combine these two methods in a single analysis. From the intensity and quantity data obtained in targeted quantification, we propose to fit and compare *loess*, spline and monotonic regression models explaining the quantity of a peptide by its intensity in the considered sample. These models allows us to estimate the amounts of all detected peptides thanks to the use of internal labelled standards for a subset of peptides.

Keywords. Loess regression, spline smoothing, monotonic regression, prediction and proteomics data.

1 Contexte

L'analyse protéomique quantitative permet d'identifier et de quantifier l'ensemble des protéines exprimées par une cellule, un tissu, un organe ou un organisme à un moment donné et sous des conditions données. Deux approches de quantification peuvent être distinguées : la quantification ciblée et la quantification globale.
Les méthodes de quantification ciblée permettent, en introduisant dans l'échantillon biologique considéré des standards synthétiques marqués correspondant aux peptides d'intérêt et préalablement sélectionnés, de connaître précisément la quantité d'une (de) protéine(s) spécifique(s) dans celui-ci. Les méthodes de quantification globale fournissent une information quantitative pour chaque protéine contenue dans les échantillons, et ce sans marquage. Elles permettent ainsi de comparer les niveaux d'expression des protéines à travers les différents échantillons biologiques considérés.
Une approche récente, appelée Data-Independent Acquisition (DIA) (Gillet *et al.* (2012)), combine ces deux approches (Borràs et Sabidó (2017), Ludwig *et al.* (2018)). En effet, elle permet d'obtenir en une seule analyse une quantification précise de protéines d'intérêt présentes dans un échantillon grâce à l'utilisation de peptides standards marqués et un profilage global de l'ensemble des protéines contenues dans ce même échantillon.

2 Problématique

Le travail décrit ici s'appuie sur l'expérience menée par Bonnet *et al.* (2020). Nous considérons 64 échantillons de 3 muscles bovins, pour lesquels 20 peptides correspondant aux 10 protéines potentiels biomarqueurs pour la tendreté ou le persillage de la viande de boeuf ont été analysés par l'approche DIA. Une première étape de quantification ciblée a permis, à partir de l'intensité mesurée par chromatographie liquide couplée à la spectrométrie de masse, de déterminer précisément la quantité des 20 peptides d'intérêt au sein de chacun des 64 échantillons considérés. Pour ce faire, la relation suivante a été

utilisée:

$$\frac{Quantité\ du\ peptide}{Quantité\ du\ peptide\ synthétique} = \frac{Intensité\ du\ peptide}{Intensité\ du\ peptide\ synthétique}.$$

Une seconde étape d'extraction des données de ces mêmes analyses a permis de mesurer l'intensité d'environ 5500 peptides par échantillon.
En protéomique quantitative, l'hypothèse forte est faite d'une relation de proportionnalité entre la quantité d'un peptide et son intensité au travers du facteur de réponse, propre à chaque peptide dans chaque échantillon. La démarche que nous présentons ici est la suivante. À partir des données d'intensité et de quantité obtenues par quantification ciblée grâce à des peptides standards internes marqués sur un sous-ensemble de peptides, nous ajustons des modèles de régression *loess*, spline et monotone expliquant la quantité d'un peptide par son intensité dans l'échantillon considéré. Ces modèles nous permettent ensuite d'estimer les quantités pour l'ensemble des peptides dont les intensités ont été mesurées durant l'analyse en mode DIA. Une comparaison de ces modèles est présentée.

3 Méthodes

3.1 La régression *loess*

La régression *loess* (Cleveland et Devlin, 1988) est une méthode de régression polynomiale considérant des sous-ensembles de données. Ainsi, au **voisinage** de chaque point x_i, une fonction polynomiale est ajustée en utilisant la méthode des moindres carrés ordinaires. Cet ajustement de polynôme se fait avec **pondération** : les points les plus proches de x_i ont davantage de poids dans l'ajustement. La taille du voisinage est déterminée par un paramètre noté α. Quand $\zeta < 1$, le voisinage inclut une proportion ζ des points du jeu de données. Quand $\zeta \geqslant 1$, tous les points sont utilisés. L'ajustement ne dépend alors plus de ζ. La détermination d'un paramètre `span` optimal peut se faire par rééchantillonnage (*bootstrap*, validation croisée, validation croisée généralisée) ou à l'aide d'un critère d'information (AICc). La figure 1 représente l'ajustement optimal par validation croisée bootstrap.

3.2 Interpolation par spline

Considérons que les données soient divisées en intervalles, tels que pour n points $x_i, i = 1, \ldots n$, il y ait $n-1$ intervalles $[x_i; x_{i+1}]$. Une spline est une fonction polynomiale définie par morceaux sur chaque intervalle. Les bornes des intervalles constituent des points de raccord pour chaque morceau de polynôme. Plus particulièrement, les splines non-négatives (M-splines) et les splines monotones (I-splines) ont été envisagées (Ramsay, 1988).

3.3 Régression non paramétrique monotone

Les équations des modèles de régression non paramétriques s'écrivent de la façon suivante :

$$Y_i = m(X_i) + \sigma(Xi) \times \varepsilon_i,$$

où m est la fonction de régression et σ la fonction de variance. La fonction `monreg` du *package* `monreg` du logiciel libre `R` effectue une estimation monotone de la fonction de régression inconnue m. Elle commence par estimer m par une méthode non paramétrique non contrainte, comme par exemple l'estimation classique de Nadaraya-Watson ou l'estimation linéaire locale. Dans une deuxième étape, l'inverse de la fonction de régression (monotone) est calculé, en monotonisant cette estimation non contrainte. Avec la notation ci-dessus et en notant $\widehat{m}$ l'estimation non contrainte, la deuxième étape s'écrit :

$$\widehat{m}_I^{-1} = \frac{1}{nh_d} \sum_{i=1}^{n} \int_{-\infty}^{t} K_d \left(\frac{\widehat{m}\left(\frac{i}{n}\right) - u}{h_d} \right) du,$$

où K_d désigne un noyau qui estime la fonction de densité et h_d la fenêtre liée également à l'estimation de la fonction de densité. L'estimation monotone est obtenue par l'inversion de $\widehat{m}_I^{-1}$ et est représentée à la figure 2.

4 Conclusion et perspectives

L'hypothèse de proportionnalité entre la quantité d'un peptide et son intensité conduit à privilégier la régression monotone. Un modèle de régression monotone pour chaque échantillon a été ajusté à partir des données de quantification ciblée. Les quantités correspondant aux intensités des peptides mesurées lors de l'étape de quantification globale ont ainsi pu être estimées, permettant *in fine* de mettre en avant des différences de propriétés métaboliques et contractiles des muscles bovins étudiés.

Bibliographie

M. Bonnet, J. Soulat, J. Bons, S. Léger, L. De Koning, C. Carapito et B. Piccard. (2020). Quantification of biomarkers for beef meat qualities using a combination of Parallel Reaction Monitoring- and antibody-based proteomics. *Food Chemistry*, **317**.

E. Borràs et E. Sabidó. (2017). What is targeted proteomics? A concise revision of targeted acquisition and targeted data analysis in mass spectrometry. *Proteomics*, **17**(17-18).

L. C. Gillet *et al.* (2012). Targeted Data Extraction of the MS/MS Spectra Generated by Data-independent Acquisition: A New Concept for Consistent and Accurate Proteome Analysis. *Molecular & Cellular Proteomics*, **11**(6).

C. Ludwig, L. Gillet, G. Rosenberger, S. Amon, B. C. Collins et R. Aebersold. (2018). Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial. *Molecular Systems Biology*, **14**(8).

K. Pilz et S. Titoff. (2020). monreg: Nonparametric Monotone Regression. R package version 0.1.4. https://CRAN.R-project.org/package=monreg.

W. S. Cleveland et S. J. Devlin. (1988). Locally-Weighted Regression: An Approach to Regression Analysis by Local Fitting, *Journal of the American Statistical Association*, **83**(403), 596-610.

J.O. Ramsay. (1988). Monotone Regression Splines in Action, *Statistical Science*, **3**(4), 425-461.

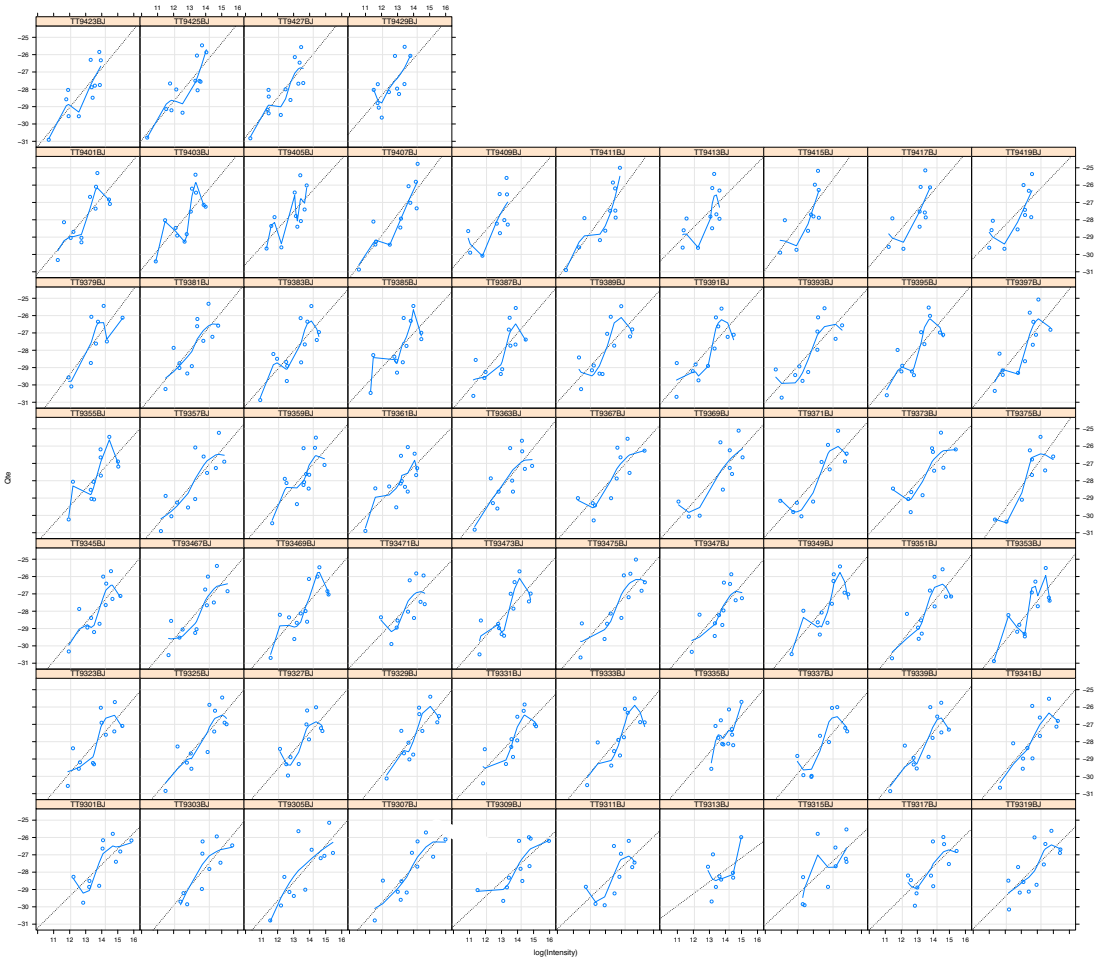


Figure 1: Représentations du logarithme de la quantité en fonction du logarithme de l'intensité, des régressions linéaires et des régressions *loess* optimales.

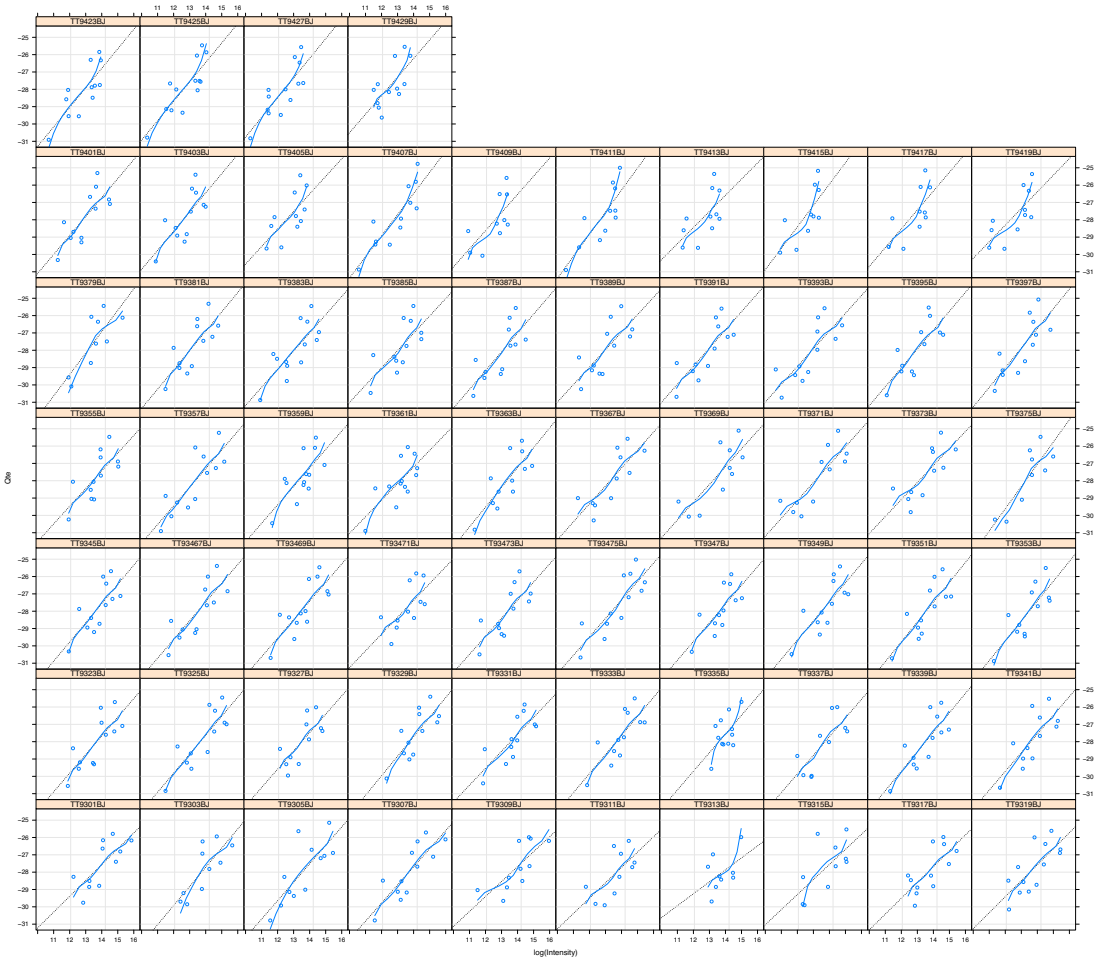


Figure 2: Représentations du logarithme de la quantité en fonction du logarithme de l'intensité, des régressions linéaires et des régressions non-paramétriques monotones.

A smooth, consistent regression tree and ensemble extensions through RF and GBT

Sami Alkhoury [1] & Emilie Devjver [2]& Marianne Clausel [3]& Myriam Tami [4]& Eric Gaussier [5]& Georges Oppenheim [6]

[1] *Univ. Grenoble Alpes, CNRS, LIG, France (sami.alkhoury@univ-grenoble-alpes.fr)*
[2] *Univ. Grenoble Alpes, CNRS, LIG, France (emilie.devijver@univ-grenoble-alpes.fr)*
[3] *Univ. Lorraine, CNRS, IECL, France (marianne.clausel@univ-lorraine.fr)*
[4] *Centrale-Supelec, Paris-Saclay, France (myriam.tami@centralesupelec.fr)*
[5] *Univ. Grenoble Alpes, CNRS, LIG, France (eric.gaussier@imag.fr)*
[6] *Laboratoire de Mathématiques d'Orsay, Université Paris Sud, France (georges.oppenheim@gmail.com)*

Résumé. Les méthodes d'ensemble basées sur les arbres de décision, comme les Forêts Aléatoires et les Gradient Boosting Trees, ont été utilisés avec succès pour des problèmes de regression dans des applications très diverses. Nous proposons ici une généralisation des arbres de décisions, les *smooth trees*, qui s'adaptent à la régularité de la fonction de lien. On considère ainsi qu'une observation, appartenant à une région particulière subit l'influence des autres régions avec un certain poids dépendant de sa distance à ces région. Nous montrons que les *smooth trees* sont consistants, une propriété qui n'avait pas été établie pour les modèles précédents comme les *soft trees*. On montre ensuite comment les *smooth-trees* peuvent être utilisés dans les méthodes d'ensemble comme comme les Forêts Aléatoires et les Gradient Boosting Trees. Des expériences numériques conduites sur plusieurs jeux de données illustrent le bon comportement de notre méthode.

Mots-clés. Méthodes d'ensembles, smooth trees consistants, forêts alétoires, gradient boosting

Abstract. Tree-based ensemble methods, as Random Forests and Gradient Boosted Trees, have been successfully used for regression problems in many applications and research studies. We propose here a generalization of regression trees, referred to as *smooth trees*, that adapt to the smoothness of the link function. By doing so, one considers that an observation, even though it belongs to a particular region, can still be associated to other regions with a certain weight that depends on the distance between the observation and the region. We show that smooth trees are consistent, a property that has not been established, as far as we know, on previous proposals as *soft trees*. We also show how smooth regression trees can be used in different ensemble methods, namely Random Forests and Gradient Boosted Trees. Experiments conducted on several data sets further illustrate the good behavior of the method.

Keywords. Ensemble methods, consistant smooth trees, random forest, gradient boosting

1 Our new model : Smooth Regression Trees

We consider the following regression problem

$$Y = f(\mathbf{X}; \Theta) + \varepsilon_Y,\ \varepsilon_Y \sim \mathcal{N}(0, \tilde{\sigma}^2), \tag{1}$$

where $\mathbf{X} = (X_1, \cdots, X_p)$ is a p-dimensional input random vector, Y be an output random variable and Θ is the set of parameters on which f relies.

We are interested in this study in approximations of f based on regression trees, either considered in isolation or aggregated in *sum-of-trees* models as Random Forests or Gradient Boosted Trees.

Standard regression trees (see Breiman et al. 1984) define a partition of $\mathbb{R}^p$ into K hyper-rectangles, referred to as regions and denoted $\mathcal{R}_k = [a_{k,1}, b_{k,1}] \times ... \times [a_{k,p}, b_{k,p}]_{1 \leq k \leq K}$, obtained by dyadic splits. A weight γ_k is associated to the k-th region $\mathcal{R}_k$ leading, for observations $\mathbf{x} \in \mathbb{R}^p$, to a predictor of the form: $f(\mathbf{x}; \Theta) = \sum_{k=1}^{K} \gamma_k \mathbb{1}_{\{\mathbf{x} \in \mathcal{R}_k\}}$. Both categorical and quantitative inputs can in theory be considered. For the sake of simplicity, we focus in this study on quantitative inputs.

As noted in previous studies (see Irsoy et al. 2012), as they are based on constant piecewise functions, standard regression trees may fail to accommodate the smoothness of the link function. To solve this problem, we introduce a prediction function that generalizes the one of standard regression trees by replacing its indicator function with a smooth function Ψ:

$$T_s(\mathbf{x}; \Theta) = \sum_{k=1}^{K} \gamma_k \Psi(\mathbf{x}; \mathcal{R}_k, \boldsymbol{\sigma}). \tag{2}$$

The set of parameters $\Theta = ((\mathcal{R}_k)_{1 \leq k \leq K}, \boldsymbol{\gamma}, \boldsymbol{\sigma})$ corresponds to the set of regions, associated weights represented by $\boldsymbol{\gamma} \in \mathbb{R}^K$ and noise in the input variables captured in $\boldsymbol{\sigma} \in \mathbb{R}_+^p$. We emphasis that when $\Psi(\mathbf{x}; \mathcal{R}_k, \boldsymbol{\sigma}) = \mathbb{1}_{\{\mathbf{x} \in \mathcal{R}_k\}}$, $\forall k,\ 1 \leq k \leq K$, one recovers classical standard regression trees.

The functions Ψ we consider relate, through a *sufficiently regular* probability density function ϕ, data points to different regions of the tree and smooth the predictions made. They are defined by: for all $\mathbf{x} \in \mathcal{X}$,

$$\Psi(\mathbf{x}; \mathcal{R}_k, \boldsymbol{\sigma}) = \frac{1}{\prod_{j=1}^{p} \sigma_j} \int_{\mathcal{R}_k} \phi\left(\left(\frac{u_j - x_j}{\sigma_j} \right)_{1 \leq j \leq p} \right) d\mathbf{u}. \tag{3}$$

One can for e.g. consider for ϕ the multivariate Gaussian distribution with a diagonal covariance matrix.

The above framework significantly departs from the ones of previously proposed for regression trees in that (a) the estimation of the parameters, in particular of the different regions, requires a dynamic update of the probability distributions of data points across

regions, which contrasts with the way regions are constructed in standard and soft regression trees, and (b) the consistency of the prediction function T_s cannot be derived from the one of standard regression trees as the latter strongly relies on the hard assignment of data points to regions.

2 Parameter estimation

Given a training set $\mathcal{D}_n = \left\{(\mathbf{x}^{(i)}, y^{(i)})_{1\leq i\leq n}\right\}$, with $\mathbf{x} \in \mathbb{R}^p$, $y \in \mathbb{R}$, and in accordance with the empirical risk minimization principle with a quadratic loss, the estimation procedure for smooth trees followed here aims at finding the parameters Θ solutions of:

$$\underset{\Theta}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y^{(i)} - \sum_{k=1}^{K} \gamma_k P_{ik} \right)^2, \tag{4}$$

with $P_{ik} := \Psi(\mathbf{x}^{(i)}; \mathcal{R}_k, \boldsymbol{\sigma})$. The $n \times K$ matrix $\boldsymbol{P}$ thus encodes the relations between each training example $\mathbf{x}^{(i)}$ and each region $\mathcal{R}_k$ and is such that $0 \leq P_{ik} \leq 1$ and $\forall i,\ 1 \leq i \leq n, \sum_{k=1}^{K} P_{ik} = 1$.

The estimation of the different parameters in Θ is done in the following way: for a fixed $\boldsymbol{\sigma}$, one alternates in between region and weight estimates, as in standard regression trees, till a stopping criterion is met[1]. During this process, the number of regions is increased and the matrix $\boldsymbol{P}$ and the weights $\boldsymbol{\gamma}$ are gradually updated.

Estimating $\boldsymbol{\gamma}$ – When fixing the regions $(\mathcal{R}_k)_{1\leq k\leq K}$ and the vector $\boldsymbol{\sigma}$, minimizing Eq. (4) with respect to $\boldsymbol{\gamma}$ leads to a weighted average of $y^{(1)}, \ldots, y^{(n)}$ if $\mathbf{P}^T\mathbf{P}$ is not singular:

$$\hat{\boldsymbol{\gamma}} = \underset{\boldsymbol{\gamma}\in\mathbb{R}^K}{\operatorname{argmin}} \sum_{i=1}^{n} (y^{(i)} - \sum_{k=1}^{K} \gamma_k P_{ik})^2 = \left(\boldsymbol{P}^T\boldsymbol{P}\right)^{-1} \boldsymbol{P}^T\boldsymbol{y}, \tag{5}$$

where $\boldsymbol{y}$ denotes the vector of n univariate outputs $y^{(1)}, \ldots, y^{(n)}$ (the derivation is direct and omitted here). The invertibility of $\mathbf{P}^T\mathbf{P}$ when ϕ is the multivariate Gaussian distribution can be proved under some mild assumptions.

Estimating $(\mathcal{R}_k)_{1\leq k\leq K}$ – Let us assume that K' regions, referred to as *current regions*, have already been identified, meaning that the current tree has K' leaves. As in standard regression trees, each current region $\mathcal{R}_k$, $1 \leq k \leq K'$, can be decomposed into two subregions wrt a coordinate $1 \leq j \leq p$ and a splitting point s_k^j that minimizes Eq. (4). Each split leads to the update of $\mathbf{P}$, that now belongs to $M_{n,K'+1}(\mathbb{R})$, and $\boldsymbol{\gamma}$, that now belongs to $\mathbb{R}^{K'+1}$. Substituting $\boldsymbol{\gamma}$ by its value given in Eq. (5), the best split for the current region $\mathcal{R}_k$ is given by:

$$\underset{1\leq j\leq p, s\in\mathcal{S}_k^j}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y^{(i)} - \sum_{l=1}^{K'+1} \left(\left(\boldsymbol{P}^T\boldsymbol{P}\right)^{-1} \boldsymbol{P}^T\boldsymbol{y} \right)_l P_{il} \right)^2, \tag{6}$$

[1]Any standard stopping criterion, as tree depth or number of examples in a leaf, can be used here.

where $\mathcal{S}_k^j$ denotes the set of splitting points for region $\mathcal{R}_k$ and variable j. More precisely, $\mathcal{S}_k^j$ is the set of middle points of the observations from $\mathcal{R}_k$ projected on the jth coordinate.
Estimating $\boldsymbol{\sigma}$ – Lastly, the vector $\boldsymbol{\sigma}$ can either be based on *a priori* knowledge or be learned through a grid search on a validation set. We rely on the latter in our experiments.

3 Illustration of the method on a toy example

To illustrate the behavior of smooth trees, we consider a toy example based on $Y = \cos(X) + \varepsilon$, with $X \sim \mathcal{U}([0, 5])$ and $\varepsilon \sim \mathcal{N}(0, 0.05^2)$. For the smooth tree, the function ϕ involved in the definition of Ψ is a multivariate Gaussian and the vector $\boldsymbol{\sigma}$ is fixed, on each dimension j, to $\sigma_j^2 = \widehat{\text{var}}((x_j{}^{(i)})_{1 \leq i \leq n})/2 = 0.74^2$. Figure 1 (left) compares the performance of both trees to the true regression function.

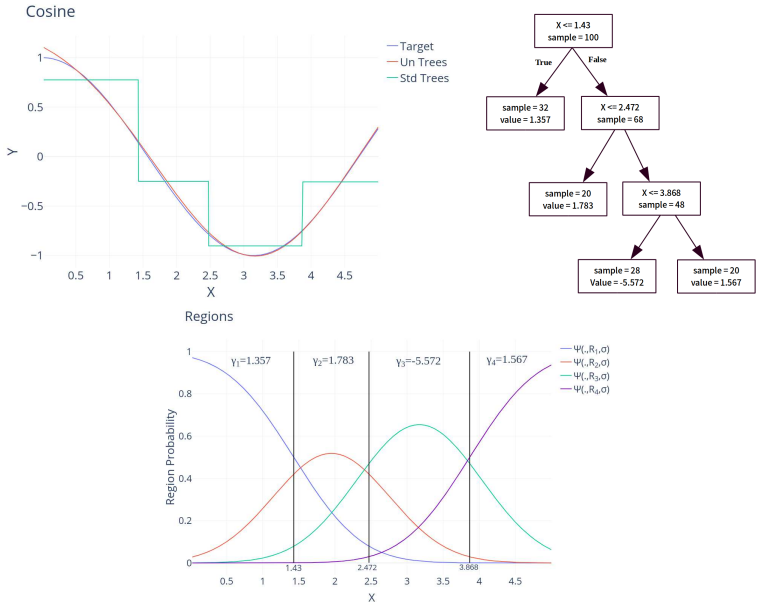


Figure 1: Left: plot of the true regression function (in blue), the smooth tree prediction (in red) and the standard regression tree prediction (in green). Middle: smooth tree learned from 100 observations where the stopping criteria consists in having at least 20% of the observations in each leaf. Right: description of the smooth tree parameters: regions (corresponding to intervals as $p = 1$), $\boldsymbol{\gamma}$ in each region and the functions Ψ associated to each region.

4 Consistency

We present in this section theoretical results on the consistency of smooth trees: they show that the smooth tree learned from a training set of size n, denoted $\hat{T}_s^{(n)}$, satisfies $\lim_{n\to+\infty} \mathbb{E}[\hat{T}_s^{(n)}(\mathbf{X}) - \mathbb{E}(Y|\mathbf{X})]^2 = 0$.

To emphasize the fact that the trees learned depend on a training set of size n, we denote the k-th region as $\mathcal{R}_k^{(n)}$. Furthermore, we consider that the observations lie in $[0,1]^p$. We also assume that the noise vector $\boldsymbol{\sigma}$ is fixed and known[2].

The function Ψ at the basis of smooth trees (Eq. (3)) relies on a probability density function ϕ that regulates how each data point is distributed across all regions of the tree. We assume here that ϕ satisfies the following conditions:

1. The support of its Fourier transform $\mathcal{F}\phi$ is $\mathbb{R}^p$.
2. $\exists r > 0$, $\sup_{\mathbf{v}\in\mathbb{R}^p} |\mathbf{v}|^{1+r+p/2}|\phi(\mathbf{v})| < \infty$.
3. $\phi \in \mathcal{C}^1$.

Note that the second condition ensures that $\phi \in L^2$. Let us now consider the Sobolev space of functions defined, for $s \in (1,2)$, by:

$$H^s([0,1]^p) = \{f \in L^2([0,1]^p),\ \exists g \in H^s(\mathbb{R}^p) \text{ s.t. } f = g|_{[0,1]^p}\},$$

where:

$$H^s(\mathbb{R}^p) = \{f \in L^2(\mathbb{R}^p),\ (1 + ||\cdot||_2^2)^{\frac{s}{2}}|\mathcal{F}f(\cdot)| \in L^2(\mathbb{R}^p)\}.$$

Let (K_n) be a non decreasing sequence of integers. Assume that:

$$K_n \xrightarrow[n\to+\infty]{} +\infty,\ \frac{K_n(\log n)^9}{n} \xrightarrow[n\to+\infty]{} 0,$$

which means that the number of regions needed to explain an infinite number of data points is also infinite, however still largely (by a factor $(\log n)^9$) dominated by the number of points. The main result of this section is the following one

Theorem 1. *Set $M > 0$. With K_n as above, assume that for some $s \in (1,2)$, $\mathbb{E}(Y|\mathbf{X}) \in H^s([0,1]^p)$ a.s. and that*

$$\max_{k=1,\cdots,K_n} \left[\text{diam}(\mathcal{R}_k^{(n)} \cap [-M,M]^p)\right] \xrightarrow[n\to+\infty]{} 0.$$

Then

$$\lim_{n\to+\infty} \mathbb{E}[\hat{T}_s^{(n)}(\mathbf{X}) - \mathbb{E}(Y|\mathbf{X})]^2 = 0.$$

The assumption on the diameter of the regions is reasonable for data points lying in a compact subspace: as the number of regions grows (to infinity) with the number of data points, their diameter will decrease, data points being more and more concentrated in each region.

[2]Note that this assumption is in line with the procedure we rely on to estimate $\boldsymbol{\sigma}$.

5 Experimental validation

Experiments conducted on classical data sets illustrate the good behavior of uncertain trees with respect to other classical approaches as standard decision trees or soft decision trees (Irsoy et al. 2012). We make use here of 10 data sets, namely Abalone (AB), Ailerons (AL), Boston (BO), Diabetes (DI), Facebook Comments (FC), Forest Fires (FF), Ozone (OZ), Skill (SK), Super Conductor (SC) and Video Transcoding (VT), all commonly used in regression tasks.

For soft trees, we use the implementation available in github[3] with the default parameters. For standard regression trees, we use the implementation from Scikit-Learn [**?**]. Concerning our method *smooth trees*, in order to reduce the overall complexity, we consider as potential variables to split the top V variables according to the splitting criterion of standard regression trees (this last step is negligible when the tree is of depth 2 or more). We give here the results for $V = 1$, $V = 3$ and $V = 5$.

Table 1: Results obtained with 10 stratified cross-validation on smooth, standard and soft trees. Best results are starred and results not significantly different from the best result are in bold.

Data set	Smooth tree			Std tree	Soft tree
	V=1	V=3	V=5		
OZ	**17.82(2.4)***	**18.66(3.65)**	**18.88(3.23)**	**18.9(3.39)**	**34.44(43.27)**
DI	**57.05(3.82)**	**55.92(3.97)**	**55.76(3.96)***	60.95(3.92)	64.18(4.15)
AB	**3.08(0.29)**	**3.11(0.27)**	**3.08(0.28)***	**3.15(0.29)**	**3.11(0.23)**
BO	4.7(0.88)	**4.47(1.04)**	**4.21(0.88)***	5.27(0.61)	4.54(0.97)
FF	**62.58(28.66)**	**62.69(28.5)**	**62.58(28.34)***	**62.72(27.77)**	**64.13(28.41)**
SK	1(0.04)	0.98(0.05)	0.98(0.03)	1.06(0.03)	**0.95(0.02)***
AL	**1.23(0.03)***	**1.24(0.02)**	**1.24(0.02)**	1.65(0.03)	1.57(0.7)
SC	18.79(0.21)	**17.89(0.28)***	18.1(0.37)	18.79(0.21)	21.96(1.27)
FC	**28.77(3.32)***	**28.87(3.32)**	**28.9(3.35)**	33.12(3.82)	**31.68(3.47)**
VT	11.44(0.16)	**11.08(0.18)***	11.16(0.19)	11.87(0.17)	15.98(0.25)

Bibliographie

Breiman L., Friedman J.H., Olshen R.A., Stone C.J. (1984), Classification and Regression Trees. *New York: Chapman and Hall*.

Györfi, L., Kohler, M., Krzyżak, A., and Walk, H. (2002), A Distribution-Free Theory of Nonparametric Regression. *Springer*.

Irsoy, O., Yildiz, O. T., and Alpaydin, E. (2012), Soft decision trees. *In International Conference on Pattern Recognition*.

[3]https://github.com/oir/soft-tree

Spatial sampling and spatial entropy
Échantillonnage spatial et entropie spatiale

Linda Altieri and Daniela Cocchi *

Resumé

L'échantillonnage spatial a pour objectif de collecter des sous-ensembles d'individus d'une population, dans l'espace bidimensionnel, afin d'estimer certaines caractéristiques de la population. Après avoir rappelé quelques aspects importants de la théorie de l'entropie spatiale, l'article compare les conséquences de l'adoption de deux systèmes de pondération différents qui considèrent l'espace dans une sorte d'échantillonnage séquentiel. Quatre structures spatiales sont proposées. La technique de poids maximal plus traditionnelle est préférable dans le cas d'un schéma spatial compact. Puisque cette technique est basée uniquement sur les distances, une forte corrélation positive entre les valeurs de la variable est cachée dans ce cas.

Keywords: Environmental sampling, spatially correlated Poisson sampling, sampling entropy.

1 Introduction

The objective of spatial sampling is to collect samples, i.e. subsets of individuals from a population, in the 2-dimensional space, in order to estimate some population characteristics. Spatial sampling is strongly linked to environmental sampling. Such expression states that the data spatial location is a fundamental information, and that sampling techniques for environmental data are mutuated from the theory of spatial sampling. Under the viewpoint of the reference population, environmental sampling focuses on natural populations. Examples can be found in biology, geography, landscape studies, forestry, and in the study of environmental dangers such as wildfires, earthquakes, polluting agents.

In finite population inference, the design-based context aims at estimating population quantities, considered as unknown but fixed. In this case the only source of randomness is the probability of the samples, which is related to the inclusion/extraction probabilities of each population element. Information may be available for moving such individual probabilities from equality. In particular, information related to space may be organized in this respect.

The link between sampling and entropy has been extensively debated in statistics (Shewry and Wynn, 1987; Lee, 2006). The search for sampling plans with high entropy is an important task in survey sampling design-based theory. Sample selection should follow the idea of randomization: a

*Department of Statistical Sciences, University of Bologna

sampling design should assign a non-null probability to as many samples as possible. A widely accepted measure of randomness of a sampling design is its entropy (Tillé and Haziza, 2010; Tillé and Wilhelm, 2017): a sampling design has high entropy when there is a high amount of uncertainty or surprise in the sample to select. Conditional Poisson sampling has been identified as the maximum entropy sampling design when the sample size is fixed (Hajek, 1981; Tillé, 2006; Tillé and Wilhelm, 2017). Maximum entropy sampling has been deepened in computer science and received important contributions in such field (Ko et al., 1995).

Under a different perspective, entropy is a popular heterogeneity measure since a long time, with reference to any kind of random variables. After being firstly introduced in information theory (Shannon, 1948), it rapidly became popular in many applied sciences to measure the degree of heterogeneity among observations. In its original proposal, entropy does not take space into account. A rather recent research field aims at accounting for space in entropy measures. In this spirit, a sequel of papers (Altieri et al., 2018a, 2019a,b) exploits the decomposition of bivariate distributions linked to entropy in order to quantify the contribution of spatial association to the entropy of a variable. Euclidean distances between spatial locations are employed for constructing the second variable. Such spatial entropy measures are employed in this exposition to improve a sequential spatial design.

In what follows we refer to the basic concept above as "spatial entropy", while the entropy of the sampling design is to be named, rather, "sampling entropy". The two entropies need to be distinguished, as they refer to different aspects of the data. Spatial entropy refers to the spatial correlation of the study variable. Sampling entropy is associated to the randomness of the potential samples; it regards the chances of selecting population units, irrespective of the value they possess for the variable.

The simulation study and all computations are implemented via the `R` software, with the help of the packages `SpatEntropy` (Altieri et al., 2018b) and `BalancedSampling` (Grafström and Lisic, 2018).

2 Recalling some theory

Spatially Correlated Poisson Sampling is a sequential (adaptive) technique that modifies the initial first order inclusion probabilities of the elements of a finite population according to the scheme described as follows (Grafström, 2012).

Starting from the first population unit, for which a Bernoulli draw with probability π_1 is proposed, after the draw an indicator function is $I_1 = 1$ if the unit is sampled, and 0 otherwise. Then, at the general step $k = 2, \ldots, N$, the values for $I_1, \ldots, I_{k-1}$ are known and unit k is sampled with probability $\pi_k^{(k-1)}$, i.e. with an inclusion probability that was updated at the previous step, when sampling unit $k-1$. The inclusion probabilities for all remaining units $l = k+1, \ldots, N$ are updated:

$$\pi_l^{(k)} = \pi_l^{(k-1)} - (I_k - \pi_k^{(k-1)}) b_k^{(l)}. \quad (1)$$

This way, at each step k the inclusion probabilities of the visited units $1, \ldots, k$ leave the room to the corresponding indicator functions. At step N, the vector becomes $\pi_1^{(N)}, \ldots, \pi_N^{(N)} = I_1, \ldots, I_N$, which indeed sums to n.

Stressing on space, a geographical distance between population units is introduced for evaluating the component $b_k^{(l)}$, a factor that influences their selection in the sample. Negative correlation weights are attributed to units that are close in space in order to obtain the spatial spreading that is desirable for sampling. A "maximal weight strategy" has been proposed (Grafström, 2012), that produces samples of fixed size $n = \sum_{k \in U} \pi_k$ and is very efficient, provided that close units carry similar values.

We propose to enhance space as highlighted in (1) by building a new weighting system that exploits the theory of spatial entropy (Altieri et al., 2018a, 2019a,b). For a certain transformation Z of the variable under study X, for which the entropy

$$H(Z) = E[I(p_Z)] = \sum_{i=1}^{I} p(z_i) \log \left(\frac{1}{p(z_i)} \right). \tag{2}$$

can be constructed, a system of distances summarized by another variable W is also defined. This, for decomposing entropy $H(Z)$ as

$$H(Z) = SMI(Z, W) + H(Z)_W. \tag{3}$$

The first component of (3), called Spatial Mutual Information, is defined as

$$SMI(Z, W) = \sum_{m=1}^{M} p(w_m) SPI(Z|w_m) \tag{4}$$

where each mth component $SPI(Z|w_m)$, i.e. the Spatial Partial Information, describes the dependence of Z on a specific distance class w_m of W:

$$SPI(Z|w_m) = \sum_{r=1}^{R} p(z_r|w_m) \log \left(\frac{p(z_r|w_m)}{p(z_r)} \right). \tag{5}$$

.

In general, when $SMI(Z, W)$ is high, the value carried by a sampled unit gives us information about what to expect from its neighbouring units; the stronger the mutual information, the smaller our interest in sampling neighbouring units. A peculiar aspect of $SMI(Z, W)$ is that it can be decomposed into the partial terms $SPI(Z|w_m)$ of (5) at different distance ranges. Thanks to such decomposition, the distance ranges for the variable under study can be decided according to the problem and the corresponding $SPI(Z|w_m)$ terms chosen as contributions to weights $b_k^{(l)}$. This way, partial spatial information assumes the role of auxiliary variable for building a well founded weighting system for sampling. Each $SPI(Z|w_m)$ is always positive, and tunes sampling neighbouring units with a strength that depends on the spatial correlation of the study variable at the chosen distances.

Weights $b_k^{(l)}$ are assigned starting from one of the M SPI terms. In particular, if units k and l are in the mth distance range, then

$$b_k^{(l)} = \frac{SPI(Z|w_m)}{C} \qquad \text{for } d(k, l) \in w_m \tag{6}$$

where C is a normalizing constant so that $\sum_{l=k+1}^{N} b_k^{(l)} = 1$ for all k, i.e. the triangular weight matrix is row-standardized. The easiest solution is that the normalizing constant is just the sum of the unnormalized weights: $C = \sum_{l=k+1}^{N} \widetilde{b}_k^{(l)}$ with $\widetilde{b}_k^{(l)} = SPI(Z|w_m)$ for $d(k, l) \in w_m$.

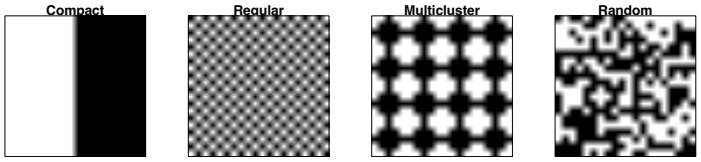


Figure 1: Basic different configurations for the same $\pi_P = 0.5$

Units are ordered according to some labelling in space. For instance, if spatial units are arranged over a grid, unit 1 can be the top-left unit, unit 2 can be at its right, or below, and so on. Different labelling orders return different updates in the inclusion probabilities of the remaining units; the method holds for any starting point and labelling criterion, as long as the distance between all pairs of units is well defined.

3 Simulation and results

In order to explore the potential improvement of spatially correlated Poisson sampling (SCPS) with the help of spatial partial information-based (SPI) weights, we run a comparative study. Simulated binary datasets are employed to estimate the variable mean (i.e. the proportion for binary data) and to evaluate the MSE of the estimator.

Consider $N = 400$ realizations of a binary variable X with half outcomes $x_0 = 0$ and half $x_1 = 1$. The true mean/proportion is $m(X) = \sum_k x_k/400 = 0.5$. Realizations are arranged over a square observation area gridded by 20×20 pixels; each pixel is assumed to be a 1×1 square, and are organized according to four different spatial configurations that produce different spatial entropy values, as in Altieri et al. (2019b). The first one is the most clustered spatial distribution, named "compact", obtained by assigning x_0 values to the pixels located at the left part of the window and x_1 values to pixels located at the right part. The second one is the most "regular" spatial distribution, corresponding to a chessboard, obtained by assigning x_0 values to pixels adjacent to x_1-valued pixels, and viceversa. The third one is a "multicluster" distribution with 16 clusters, whose centroids are regularly distributed over the area; then, x_0 values are assigned to pixels surrounding the centroids and x_1 values to the remaining pixels. The last one is a "random" pattern without spatial correlation whatsoever, obtained by assigning x_0 or x_1 values to pixels via simple random sampling without replacement from the generated sequence. The four datasets are displayed in Figure 1.

Four distance classes are chosen over the observation area: $w_1 = [0, 1]$, $w_2 =]1, 2]$, $w_3 =]2, 5]$, $w_4 =]5, 20\sqrt{2}]$, where $20\sqrt{2}$ is the maximum distance over the observation window. Reasons for choosing such classes are found in the fundamentals of spatial statistics, and are discussed in Altieri et al. (2018a, 2019b). Spatial partial information terms (5) are computed, following Altieri et al. (2019b), and the resulting values are shown in Table 1.

Afterwards, 100 samples of fixed size $n = 40$ are drawn from each dataset. The initial inclusion probabilities are constant: $\pi_k = n/N$ for all units k. Then, a sample is drawn from each dataset using the sequential approach of SCPS, where the weight assigned to each pair of units comes from the SPI value in Table 1, according to the distance between units.

Table 1: Spatial partial information at the four distance classes

	$[0,1]$	$]1,2]$	$]2,5]$	$]5,20\sqrt{2}]$
Compact	0.786	0.687	0.455	0.010
Regular	0.509	0.918	<0.001	<0.001
Multicluster	0.146	0.057	0.008	<0.001
Random	0.020	0.010	0.001	<0.001

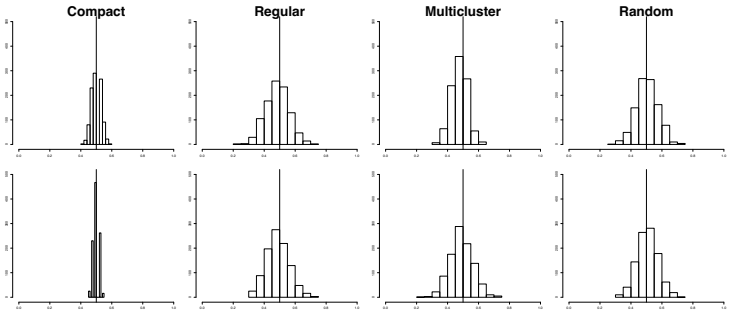


Figure 2: HT estimate for $m(X) = 0.5$; first line with SPI, second line with maximal weight

The SPI values are rescaled so that they sum to 1 for each population unit, and the constraint for positive weights is checked. The SCPS is implemented, updating the remaining units' inclusion probabilities sequentially, until all population units have been either sampled or rejected and a sample size of n is reached.

All results are compared to SCPS with maximal weights as implemented by the R package `BalancedSampling` (Grafström and Lisic, 2018). Note that the 100 samples produced with the maximal weights system are the same across spatial configurations, since such weights only consider the distance between units and not the strength of the spatial correlation.

We compare the two weighting systems, with the sample size constrained to be n for all samples, thanks to the sequential technique and to the sum-to-1 constraint for the weights. The HT estimate for the variable mean is displayed in Figure 2, where the thick vertical line marks the true mean $m(X) = 0.5$. The MSE has been computed over the simulated samples, with results displayed in Table 2. Results with the maximal weight technique are winning in the case of a compact spatial scheme. Since it is based only on distances, a strong positive correlation between the values of the variable is hidden in this case.

.

References

Altieri, L., D. Cocchi, and G. Roli (2018a). A new approach to spatial entropy measures. *Environmental and Ecological Statistics 25*(1), 95–110.

Table 2: Mean Squared Error of the HT estimates.

	Maximal weights	SPI weights
Compact	.0004	.0009
Regular	.0053	.0057
Multicluster	.0055	.0028
Random	.0049	.0052

Altieri, L., D. Cocchi, and G. Roli (2018b). *SpatEntropy: Spatial Entropy Measures*. R package version 0.1.0.

Altieri, L., D. Cocchi, and G. Roli (2019a). Measuring heterogeneity in urban expansion via spatial entropy. *Environmetrics 30*(2), e2548.

Altieri, L., D. Cocchi, and G. Roli (2019b). Advances in spatial entropy measures. *Stochastic Environmental Research and Risk Assessment*.

Bondesson, L. and D. Thorburn (2008). A list sequential sampling method suitable for real-time sampling. *Scandinavian Journal of Statistics 35*(3), 466–483.

Grafström, A. (2012). Spatially correlated poisson sampling. *Journal of Statistical Planning and Inference 142*(1), 139–147.

Grafström, A. and J. Lisic (2018). *BalancedSampling: Balanced and Spatially Balanced Sampling*. R package version 1.5.4.

Hajek, J. (1981). *Sampling from a finite population*. New York, New York: Marcel Dekker, Inc.

Ko, C.-W., J. Lee, and M. Queyranne (1995). An exact algorithm for maximum entropy sampling. *Operations Research 43*(4), 684–691.

Lee, J. (2006). Maximum entropy sampling. *Encyclopedia of Environmetrics 4*.

Shannon, C. (1948). A mathematical theory of communication. *Bell Dyditem Technical Journal 27*, 379–423, 623–656.

Shewry, M. C. and H. P. Wynn (1987). Maximum entropy sampling. *Journal of Applied Statistics 14*(2), 165–170.

Tillé, Y. (2006). *Sampling algorithms*. Springer.

Tillé, Y. and D. Haziza (2010). An interesting property of the entropy of some sampling designs. *Survey Methodology 36*(2), 229–231.

Tillé, Y. and M. Wilhelm (2017). Probability sampling designs: principles for choice of design and balancing. *Statistical Science 32*(2), 176–189.

Optimal adaptive estimation on $\mathbb{R}$ or $\mathbb{R}^+$ of the derivatives of a density

Fabienne Comte [1]& Céline Duval [1]& Ousmane Sacko [1]

[1] *MAP5 UMR 8145, Université Paris Descartes, 45 Rue des Saints-Pères, 75006 Paris, France. Email : ousmane.sacko@parisdescartes.com*

Résumé. Soient $X_1, \ldots, X_n$ des variables aléatoires (v.a.) i.i.d. de densité commune f. On s'intéresse à l'estimation de la dérivée d'ordre d de f, notée $f^{(d)}$. Nous construisons un estimateur par projection sur $\mathbb{R}$ ou $\mathbb{R}^+$ en utilisant les bases d'Hermite ou de Laguerre. Nous présentons une majoration du risque $\mathbb{L}^2$: si f est dans la boule de Sobolev-Laguerre ou Sobolev-Hermite de régularité s avec $s > d$, notre estimateur converge à la vitesse $n^{-2(s-d)/(2s+1)}$. Une minoration du risque garantit l'optimalité de cette vitesse au sens minimax. Enfin, nous décrivons une procédure adaptative pour choisir la dimension pertinente qui conduit à cette vitesse optimale.

Mots-clés. Estimation des dérivées d'une densité, base d'Hermite, base de Laguerre, sélection de modèle, estimateur par projection.

Abstract. Let $X_1, \ldots, X_n$ be i.i.d. random variables with common density f with respect to the Lesbegue measure. In this work, we are interested in the estimation of the d-th derivative of f, denoted $f^{(d)}$. We build a projection estimator on $\mathbb{R}$ or $\mathbb{R}^+$ using the Hermite or Laguerre bases. We present an upper bound on the $\mathbb{L}^2$-risk : if f belongs to the Sobolev-Hermite or Sobolev-Laguerre ball with regularity s, $s > d$, our estimator converges with the rate $n^{-2(s-d)/(2s+1)}$. We also establish a lower bound which ensures the optimality of the rate in the minimax sense. Finally, we describe an adaptive procedure to select the relevant dimension : the bias-variance compromise is automatically realized for this adaptive estimator.

Keywords. Estimation of derivatives of a density, Hermite basis, Laguerre basis, model selection, projection estimator.

1 Motivations

Les densités de probabilités ainsi que leurs dérivées jouent un rôle important dans plusieurs champs des statistiques. Des exemples d'applications sont donnés dans Singh (1977) : la courbe de régression $r(x) = \mathbb{E}(Y|X = x) = f^{(1)}(x)/f(x)$ pour une famille spécifique de distributions conditionnelles (voir aussi Park et Kang (2008)) ; estimation et test paramétrique pour la famille exponentielle (voir Gonese *et al* (2016)). Les dérivées d'une densité peuvent être utilisées pour la recherche de modes dans le cas du modèle de mélange et en analyse des données (voir *e.g.* Cheng (1995), Chacòn et Duong (2013)).

Enfin, les dérivées d'une densité donnent aussi de l'information sur la pente d'une courbe, les extrema locaux, les points selles....

Des méthodes à noyau et de projection ont été développées pour estimer les dérivées d'une densité : voir Bhattacharga (1967), Chacòn *et al* (2011), Chacòn et Duong (2013) pour les méthodes à noyau ; Efromovich (1958), Rao (1996), Schmisser (2013) (pour le cas dépendant), Giné et Nick (2016) pour les méthodes de projection. Dans ce papier, nous proposons un estimateur par projection en utilisant des propriétés spécifiques de la base de Laguerre et d'Hermite. L'ensemble des résultats énoncés ici proviennent de l'article Comte *et al* (2019).

2 Procédure d'estimation

2.1 Base de Laguerre et base d'Hermite

Base de Laguerre. C'est une base orthonormée sur $\mathbb{L}^2(\mathbb{R}^+)$ définie à partir du polynôme de Laguerre $(L_j)_{j\geqslant 0}$:

$$\ell_j(x) = \sqrt{2}L_j(2x)e^{-x}, \quad L_j(x) = \sum_{k=0}^{j} \binom{j}{k}(-1)^k \frac{x^k}{k!}, \quad x \geqslant 0, \quad j \geqslant 0.$$

Base d'Hermite. C'est une base orthonormée sur $\mathbb{L}^2(\mathbb{R})$ définie à partir du polynôme d'Hermite par $(H_j)_{j\geqslant 0}$:

$$h_j(x) = c_j H_j(x)e^{-x^2/2}, \ H_j(x) = (-1)^j e^{x^2}\frac{d^j}{dx^j}(e^{-x^2}), \ c_j = (2^j j!\sqrt{\pi})^{-1/2}, x \in \mathbb{R}, \ j \geqslant 0,$$

Dans la suite, on note $\varphi_j = \ell_j$ pour le cas Laguerre ou $\varphi_j = h_j$ pour le cas d'Hermite.

2.2 Estimateur par projection de la dérivée d'ordre d de f

Soient $X_1, \ldots, X_n$ des v.a i.i.d. de densité commune f par rapport à la mesure de Lebesgue. Soit $d, m \geqslant 1$, des entiers et $S_m = \text{Vect}\{\varphi_0, \ldots, \varphi_{m-1}\}$, l'espace engendré par $\varphi_0, \ldots, \varphi_{m-1}$. Considérons les hypothèses suivantes :

(**A**1) La densité f est d-fois dérivable et $f^{(d)}$ appartient à $\mathbb{L}^2(\mathbb{R}^+)$ pour le cas Laguerre ou $\mathbb{L}^2(\mathbb{R})$ pour le cas Hermite.

(**A**2) Pour tout r entier, $0 \leqslant r \leqslant d-1$, nous avons $\|f^{(r)}\|_\infty = \sup_{x\in\mathbb{R}}|f^{(r)}(x)| < +\infty$.

(**A**3) Pour tout r entier, $0 \leqslant r \leqslant d-1$, $\lim\limits_{x\to 0} f^{(r)}(x) = 0$ (spécifique au cas Laguerre).

Sous (**A**1), nous avons $f^{(d)} = \sum_{j\geqslant 0} a_j(f^{(d)})\varphi_j$ où $a_j(f^{(d)}) = \langle f^{(d)}, \varphi_j\rangle = \int f^{(d)}(x)\varphi_j(x)dx$ et sa projection orthogonale sur S_m est donnée par : $f_m^{(d)} = \sum_{j=0}^{m-1} a_j(f^{(d)})\varphi_j$. On réduit le problème d'estimation de $f^{(d)}$ à l'estimation des coefficients $a_j(f^{(d)})_{0\leqslant j\leqslant m-1}$.

Le Lemme suivant est la clé pour la construction de notre estimateur.

Lemme 2.1. *Sous* (**A**1) *et* (**A**2) *pour le cas Hermite ou* (**A**1), (**A**2) *et* (**A**3) *pour le cas Laguerre, nous avons* $a_j(f^{(d)}) = (-1)^d \mathbb{E}[\varphi_j^{(d)}(X_1)]$, $\forall j, d \geqslant 0$.

On déduit du Lemme 2.1 l'estimateur suivant pour $m \geqslant 1$:

$$\widehat{f}_{m,(d)} = \sum_{j=0}^{m-1} \widehat{a}_j^{(d)} \varphi_j, \quad \text{avec} \quad \widehat{a}_j^{(d)} = \frac{(-1)^d}{n} \sum_{i=1}^{n} \varphi_j^{(d)}(X_i). \tag{1}$$

Notons que pour $d = 0$ dans (1), on retrouve l'estimateur classique de la densité.

3 Vitesse de convergence de l'estimateur

3.1 Risque de l'estimateur

Sous des conditions de moment, nous avons la majoration suivante du risque de $\widehat{f}_{m,(d)}$.

Théorème 3.1. *Sous les hypothèses du Lemme 2.1 et si de plus*

$$\mathbb{E}[X_1^{-d-1/2}] < +\infty \text{ pour le cas Laguerre et } \mathbb{E}[|X_1|^{2/3}] < +\infty \text{ pour le cas Hermite.} \tag{2}$$

Alors, pour $m \geqslant d$ assez grand, nous avons que

$$\mathbb{E}\big[\|\widehat{f}_{m,(d)} - f^{(d)}\|^2\big] \leqslant \|f_m^{(d)} - f^{(d)}\|^2 + C\frac{m^{d+\frac{1}{2}}}{n} - \frac{\|f_m^{(d)}\|^2}{n}, \text{ où } \|t\|^2 = \int t^2(x)dx \tag{3}$$

où C est une constante qui ne dépend que des conditions de moment données en (2).

Les deux premiers termes de (3) ont un comportement antagoniste vis-à-vis de m : le premier terme est le terme de biais qui diminue lorsque m augmente et le deuxième est l'ordre du terme principal de variance qui est clairement un terme croissant de m. Le risque minimal s'obtient en faisant un compromis biais-variance.

La majoration du risque donnée en (3) est précise dans le sens où l'on peut établir la minoration suivante :

Proposition 3.1. *Sous les hypothèses du Théorème 3.1, il vient pour $c > 0$ que,*

$$\mathbb{E}\Big[\|\widehat{f}_{m,(d)} - f^{(d)}\|^2\Big] \geqslant \|f_m^{(d)} - f^{(d)}\|^2 + c\frac{m^{d+\frac{1}{2}}}{n} - \frac{\|f_m^{(d)}\|^2}{n}.$$

3.2 Optimisation de m et vitesse de convergence

Pour obtenir la vitesse de convergence, il nous faut quantifier l'ordre de grandeur du terme de biais. Pour ce faire, nous introduisons les espaces de régularité suivants (voir Bongioanni et Torrea (2009)) :

Définition 3.1. *(i) On définit la boule de Sobolev-Hermite de régularité $s > 0$ par :*

$$W_H^s(D) = \{\theta \in \mathbb{L}^2(\mathbb{R}), \sum_{k\geqslant 0} k^s a_k^2(\theta) \leqslant D\}, \quad a_k^2(\theta) = \langle \theta, h_k \rangle, \quad D > 0.$$

(ii) De la même manière, on définit la boule de Sobolev-Laguerre de régularité $s > 0$ par :

$$W_L^s(D) = \{\theta \in \mathbb{L}^2(\mathbb{R}^+), |\theta|_s^2 = \sum_{k\geqslant 0} k^s a_k^2(\theta) \leqslant D\},\ a_k(\theta) = \langle \theta, \ell_k \rangle, \quad D > 0.$$

Pour déduire la régularité de $f^{(d)}$ de celle de f pour le cas Laguerre, on définit le sous-ensemble suivant de $W_L^s(D)$ par :

$$\widetilde{W}_L^s(D) = \{\theta \in \mathbb{L}^2(\mathbb{R}^+),\, \theta^{(j)} \in C([0,\infty[),\, x \mapsto x^{k/2}\theta^{(j)}(x) \in \mathbb{L}^2(\mathbb{R}^+),\, 0 \leqslant j \leqslant k \leqslant s, |\theta|_s^2 \leqslant D\}.$$

Notons que $\widetilde{W}_L^s(D) \subset W_L^s(D)$ (voir Comte *et al* (2019), sect. 2.3.1, p. 6).

Lemme 3.1. *Soit $s > d$, $D > 0$, $f \in W_H^s(D)$ et (**A**1) est vérifiée pour le cas Hermite ou $f \in \widetilde{W}_L^s(D)$ pour le cas Laguerre, donc il existe une constante D_d tel que $f^{(d)}$ est dans $W_H^{s-d}(D_d)$ pour le cas Hermite ou $\widetilde{W}_L^{s-d}(D_d)$ pour le cas Laguerre.*

Donc pour $f \in W_H^s(D)$ ou $f \in \widetilde{W}_L^s(D)$, il vient du Lemme 3.1 que :

$$\|f_m^{(d)} - f^{(d)}\|^2 = \sum_{j\geqslant m} (a_j(f^{(d)}))^2 = \sum_{j\geqslant m} j^{s-d}(a_j(f^{(d)}))^2 j^{-(s-d)} \leqslant D_d m^{-(s-d)}.$$

En injectant ce résultat dans (3), on a $\mathbb{E}\big[\|\widehat{f}_{m,(d)} - f^{(d)}\|^2\big] \leqslant D_d m^{-(s-d)} + Cm^{d+\frac{1}{2}} n^{-1}$. En optimisant le terme à gauche de l'Inégalité précédente, on en déduit que

$$m_{opt} = \lfloor n^{2/(2s+1)} \rfloor \text{ et } \mathbb{E}\big[\|\widehat{f}_{m_{opt},(d)} - f^{(d)}\|^2\big] \leqslant C_{(s,d,D)} n^{-\frac{2(s-d)}{2s+1}}, \tag{4}$$

où $C_{(s,d,D)}$ dépend uniquement de s, d et D. Cette vitesse est la même que celle obtenue par Schmisser (2013) pour le cas dépendant et par Giné et Nickl (2016). Notons que les ordres de grandeur du biais et de la variance sont spécifiques à notre méthode : le rôle de la dimension est joué ici par $\sqrt{m}$. Notons également que notre hypothèse de régularité est naturellement faite sur f et non sur $f^{(d)}$ (contrairement à ce qui apparait dans certains articles). Cette vitesse est meilleure que celle obtenue par Rao (1996) dans le cas i.i.d. si on considère la même condition de régularité. Pour $d = 0$ dans (4), on retrouve la vitesse optimale pour l'estimation de f.

3.3 Borne inférieure et optimalité de la vitesse

En utilisant le schéma classique introduit par Tsybakov (2009), Chap. 2, on peut prouver le Théorème suivant : [1]

1. Tsybakov, A. B. (2009). Introduction to nonparametric estimation. Springer Series in Statistics. Springer, New York. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

Théorème 3.2. *Soit $s > d$ un entier et $\widetilde{f}_{n,d}$ un estimateur quelconque de $f^{(d)}$. Alors pour n assez grand, nous avons pour $c > 0$ une constante qui dépend de s et d uniquement*

$$\inf_{\widetilde{f}_{n,d}} \sup_{f \in W^s(D)} \mathbb{E}[\|\widetilde{f}_{n,d} - f^{(d)}\|^2] \geqslant c n^{-\frac{2(s-d)}{2s+1}},$$

où $W^s(D) = W_L^s(D)$ pour le cas Laguerre ou $W^s(D) = W_H^s(D)$ pour le cas Hermite.

Le Théorème 3.2 implique que la vitesse obtenue en (4) est minimax optimale.

4 Procédure d'estimation adaptative

Le choix de la dimension $m = m_{opt} = [n^{2/(2s+1)}]$ dépend de la régularité de f qui est inconnue, donc $\widehat{f}_{m_{opt},(d)}$ n'est pas calculable en pratique. Nous décrivons ici une procédure nous permettant de faire la sélection de la dimension m automatiquement à partir uniquement des données $X_1, \ldots, X_n$. Nous cherchons m tel que le risque

$$\mathbb{E}\big[\|\widehat{f}_{m,(d)} - f^{(d)}\|^2\big] = \|f_m^{(d)} - f^{(d)}\|^2 + \frac{1}{n}\sum_{j=0}^{m-1} \mathrm{Var}\big[\varphi_j^{(d)}(X_1)\big]$$

est minimal. On remarque que le biais est tel que : $\|f_m^{(d)} - f^{(d)}\|^2 = \|f^{(d)}\|^2 - \|f_m^{(d)}\|^2$. Comme $\|f^{(d)}\|^2$ est indépendant de m, en estimant $\|f_m^{(d)}\|^2$ par $-\|\widehat{f}_{m,(d)}\|^2$, on remplace le biais par $-\|\widehat{f}_{m,(d)}\|^2$. On estime la variance par : $\widehat{V}_{m,d} = \frac{1}{n}\sum_{i=1}^n \sum_{j=0}^{m-1} (\varphi_j^{(d)}(X_i))^2$. Ainsi, pour $\kappa > 0$ une constante à calibrer, on sélectionne m comme suit :

$$\widehat{m}_n := \arg\min_{m \in \mathcal{M}_{n,d}} \{-\|\widehat{f}_{m,(d)}\|^2 + \widehat{\mathrm{pen}}_d(m)\}, \quad \widehat{\mathrm{pen}}_d(m) = \kappa \frac{\widehat{V}_{m,d}}{n}, \tag{5}$$

où $\mathcal{M}_{n,d}$ est un sous-ensemble de $\mathbb{N}^*$. Dans la suite, on définit $\mathrm{pen}_d(m) := \mathbb{E}[\widehat{\mathrm{pen}}_d(m)] = \kappa \frac{V_{m,d}}{n}$, avec $V_{m,d} := \sum_{j=0}^{m-1} \mathbb{E}[(\varphi_j^{(d)}(X_1)^2]$. Notons qu'il est possible de piloter la procédure adaptative en remplaçant dans (5) $\widehat{V}_{m,d}$ par l'ordre de grandeur de $V_{m,d}$ (voir Théorème 3.1) *i.e.* par $cm^{d+1/2}/n$. Dans ce cas nous avons une pénalité déterministe qui dépend des conditions de moment données en (2) et des constantes non explicites qui apparaissent dans la Formule d'Askey et Wainger (1965), p. 699.

Théorème 4.1. *Soit $\mathcal{M}_{n,d} := \{d, \ldots, m_n(d)\}$, avec $m_n(d) \geqslant d$ entier. Supposons* (**A**1) *et* (**A**2)*, et* (**A**3) *pour le cas Laguerre, et supposons que $\|f\|_\infty < +\infty$.*

AL. Posons $m_n(d) = \lfloor (n/\log^3(n))^{\frac{2}{2d+1}} \rfloor$, et supposons que $\sup_{x \in \mathbb{R}^+} x^{-d} f(x) < +\infty$ pour le cas Laguerre,

AH. Posons $m_n(d) = \lfloor n^{\frac{2}{2d+1}} \rfloor$ pour le cas Hermite.

Alors, pour tout $\kappa \geqslant \kappa_0 := 32$*, nous avons que*

$$\mathbb{E}\Big[\|\widehat{f}_{\widehat{m}_n,(d)} - f^{(d)}\|^2\Big] \leqslant C \inf_{m\in\mathcal{M}_{n,d}} \big(\|f_m^{(d)} - f^{(d)}\|^2 + \mathrm{pen}_d(m)\big) + \frac{C'}{n}, \tag{6}$$

où $C > 0$ *est une constante* ($C = 3$ *convient) et* $C' > 0$ *une constante qui dépend de* $\sup_{x\in\mathbb{R}^+} x^{-d} f(x) < +\infty$ *et* $\mathbb{E}[X_1^{-d-1/2}] < +\infty$ *(cas Laguerre) ou* $\|f\|_\infty$ *(cas Hermite).*

La valeur théorique de $\kappa := 32$ est grande. En pratique cette valeur est calibrée à partir de simulations préliminaires : nous avons pris $\kappa = 4$ pour le cas Hermite ou $\kappa = 3.5$ pour le cas Laguerre (voir Comte *et al* (2019)). La contrainte sur la dimension maximale $m_n(d)$ n'influence pas la vitesse de convergence de $\widehat{f}_{\widehat{m}_n,(d)}$. L'Inégalité (6) implique que $\widehat{f}_{\widehat{m}_n,(d)}$ réalise automatiquement un compromis biais-variance et qu'il est aussi performant que le meilleur modèle dans la collection à une constante multiplicative près plus un terme résiduel C'/n qui est négligeable.

Bibliographie

Askey, R. and Wainger, S. (1965). Mean convergence of expansions in Laguerre and Hermite series, *Amer. J. Math.,* 87, pp. 695-708.

Bongioanni, B. and Torrea, J. L. (2009). What is a Sobolev space for the Laguerre function systems ?, *Studia Math.*, 192, pp. 147-172.

Bhattacharya, P. (1967). Estimation of a probability density function and its derivatives, *Sankhya : The Indian Journal of Statistics, Ser. A*, pp. 373-382.

Chacòn, J. E. and Duong, T. (2013). Data-driven density derivative estimation, with applications to nonparametric clustering and bump hunting, *Electron. J. Stat.*, 7, pp. 499-532.

Chacòn, J. E., Duong, T., and Wand, M. (2011). Asymptotics for general multivariate kernel density derivative estimators, *Statistica Sinica,* pp. 807-840.

Cheng, Y. (1995). Mean shift, mode seeking, and clustering, *IEEE transactions on pattern analysis and machine intelligence,* 17, pp. 790-799.

Comte F., Duval C., Sacko O. (2019). Optimal adaptive estimation on R or R+ of the derivatives of a density, *Preprint, hal-02296067.*

Efromovich, S. (1998). Simultaneous sharp estimation of functions and their derivatives, *Ann. Statist.,* 26, pp. 273-278.

Genovese, C. R., Perone-Pacifico, M., Verdinelli, I., and Wasserman, L. (2016). Non-parametric inference for density modes, *J. R. Stat. Soc. Ser. B. Stat. Methodol.,* 78, pp. 99-126.

Giné, E. and Nickl, R. (2016). Mathematical foundations of infinite-dimensional statistical models, volume 40, *Cambridge University Press.*

Park, C. and Kang, K. H. (2008). Sizer analysis for the comparison of regression curves, *Computational Statistics Data Analysis,* 52, pp. 3954-3970.

Rao, B. L. S. P. (1996). Nonparametric estimation of the derivatives of a density by the method of wavelets, *Bull. Inform. Cybernet.,* 28, pp. 91-100.

Schmisser, E. (2013). Nonparametric estimation of the derivatives of the stationary density for stationary processes, *ESAIM Probab. Stat.,* 17, pp. 33-69.

Singh, R. S. (1977). Applications of estimators of a density and its derivatives to certain statistical problems, *J. Roy. Statist. Soc. Ser. B,* 39, pp. 357-363.

Une méthodologie computationnelle pour faire de l'optimisation multi-objectifs en élevage de précision

Alexandre Conanec [1,2], Marie Chavent [2], Marie-Pierre Ellies-Oury [1] & Jérôme Saracco [2]

[1] *INRAE Biomarqueur team, Theix, 63122 Saint Genès Champanelle, France, Bordeaux Sciences Agro, Gradignan, France.*
alexandre.conanec@agro-bordeaux.fr ; mp.ellies@agro-bordeaux.fr
[2] *Inria BSO, CQFD team & IMB 5251 CNRS, Université de Bordeaux & Bordeaux INP, 33400 Talence, France.*
marie.chavent@math.u-bordeaux.fr ; jerome.saracco@inria.fr

Résumé. En élevage de précision, les problèmes d'optimisation sont multi-objectifs et stochastiques. En effet, les exigences des décideurs sont multiples et les fonctions objectifs ne peuvent être modélisées sous une forme analytique dû à la complexité inhérente des systèmes biologiques. La méthodologie proposée consiste à utiliser des quantiles conditionnels, estimés non paramétriquement, associés à différents niveaux de risque α choisis par le décideur, pour intégrer l'incertitude du modèle. Pour un risque α donné, l'algorithme génétique NSGAII permet ensuite de déterminer le Front de Pareto, qui porte l'ensemble des compromis possibles au sein duquel le décideur peut alors choisir. Le bon comportement numérique de l'approche développée est illustré sur une simulation numérique.

Mots-clés. Optimisation, Multi-objectifs, Incertitude, Élevage de précision, Quantiles conditionnels

Abstract. In precision rearing, optimization problems are multi-objective and stochastic because the requirements of decision-makers are multiple and because objective functions cannot be modeled in an analytical form due to the inherent complexity of biological systems. Our method consists in use a nonparametrically estimated quantile regression, associate with a α risk level decided by the decision maker, to deal with the model uncertainty. Then, the NSGAII genetic algorithm allows us to find the Pareto Front, associated with a α risk level, which carries the set of possible trade-offs within which the decision-maker can choose. The good numerical behavior of the proposed approach is illustrated on simulated data.

Keywords. Optimization, Multi-objectives, Uncertainty, Precision rearing, Conditionnal quantiles

1 Introduction

Dans le cadre de l'élevage de précision, l'élaboration de modèles d'optimisation pour piloter finement des processus biologiques soulève de nombreux défis mathématiques. En

collaboration avec INRAE, nous nous intéressons en particulier à la qualité des produits animaux, notamment celle de la viande bovine et aux facteurs permettant de la piloter. L'origine des défis qui sont posés provient en partie de la complexité inhérente aux systèmes biologiques. En effet, il est difficile de modéliser le pilotage de la qualité de la viande bovine de manière exacte. Ainsi, il est utopique de penser pouvoir le formuler de manière analytique tant les processus biologiques *ante* et *post-mortem* du muscle sont complexes et multi-factoriels. L'identification des principaux facteurs modulant la qualité permet néanmoins de construire des modèles statistiques du type

$$y = m(\mathbf{x}) + \epsilon, \tag{1}$$

où y est la variable réelle représentant une certaine qualité, $\mathbf{x}$ est la variable d-dimensionnelle désignant les facteurs permettant de piloter cette qualité (nous utiliserons par la suite le terme de variable de décision), m est la fonction de lien entre $\mathbf{x}$ et y, et ϵ est un terme d'erreur aléatoire. Ce modèle est naturellement imparfait puisque entaché d'une incertitude (via ϵ) qui doit être prise en compte lors de l'utilisation de m comme modèle de substitution (surrogate model) dans un processus d'optimisation, au risque de retenir une solution sous-optimale et/ou de prédire un optimum "trop optimiste". A cela, ajoutons que, dans la réalité, la fonction de lien m est inconnue et doit donc être estimée, notons par $\hat{m}$ l'estimateur de m. La qualité de l'estimation est dépendante du nombre n d'observations dont dispose le modélisateur. Or dans notre cas d'étude, l'acquisition de données est coûteuse, liée à une logistique importante dans les stations expérimentales et à la présence encore rare de capteurs de précision dans les élevages. Se rajoute donc à l'incertitude ϵ des modèles, l'erreur d'estimation de m qui peut être d'autant plus grande en estimation non paramétrique que la dimension d de la variable de décision est importante lorsque la taille n de l'échantillon est limitée (c'est la fameuse malédiction de la dimension).

L'autre défi majeur qui est posé dans notre étude est l'aspect multi-objectifs du problème d'optimisation. En effet, les objectifs à optimiser peuvent être contradictoires selon que l'on s'intéresse à la qualité nutritionnelle versus la qualité gustative du produit, ou à la performance économique versus la performance environnementale d'un animal.

L'application concrète d'optimisation multi-objectifs porte sur l'élaboration de cahiers des charges, avec obligation de résultats, prenant en compte différents objectifs de performance de l'animal et de qualité de sa viande.

Dans cette communication, on s'intéresse à la problématique multi-objectifs et à la gestion de l'incertitude liée à l'utilisation de modèles de substitution dans un problème d'optimisation sans contrainte. Nous proposons une méthodologie reposant sur l'estimation non paramétrique de quantiles conditionnels et sur l'estimation de front de Pareto par un algorithme génétique. Nous illustrons le bon comportement numérique de notre approche uniquement sur des données simulées, l'acquisition et le pré-traitement des données réelles étant toujours en cours.

2 État de l'art

On considère le problème multi-objectifs suivant :

$$\min_{\mathbf{x}\in\mathcal{X}} \quad \mathbf{m}(\mathbf{x}) = (m_1(\mathbf{x}), m_2(\mathbf{x}), \ldots, m_p(\mathbf{x})) \tag{2}$$

où $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^d$, est la variable de décision, ici considérée continue. La fonction $\mathbf{m} : \mathbb{R}^d \to \mathbb{R}^p$ est formée des p fonctions objectifs du problème d'optimisation. L'opérateur "min" a ici un sens ambigu puisqu'il n'existe pas d'ordre canonique dans $\mathbb{R}^p$. Dès lors, il existe deux approches pour traiter ce type de problème.

- Soit on transforme le problème multi-objectifs en problème mono-objectif par l'intégration d'un arbitrage $\mathcal{R}$ émanant du décideur. On retrouve le "goal programming", la méthode de "ϵ-constraint" ou d'autres reformulations variant dans leur manière d'intégrer $\mathcal{R}$ pour transformer le problème multi-objectifs en un problème mono-objectif (voir, par exemple, Collette et Siarry, 2003).

- Soit on introduit une relation de dominance, la plus utilisée étant la dominance de Pareto. On dira que la solution $\mathbf{x}'$ domine $\mathbf{x}''$, notée $\mathbf{x}' \succ \mathbf{x}''$ dans un contexte de minimisation, lorsque

  $$\forall\ j \in \{1, \ldots, p\},\ m_j(\mathbf{x}') \leq m_j(\mathbf{x}'') \quad \text{et} \quad \exists\ j \in \{1, \ldots, p\},\ m_j(\mathbf{x}') < m_j(\mathbf{x}''). \tag{3}$$

  À partir de cette définition de la relation de dominance, on peut introduire la notion d'optimalité de Pareto comme l'ensemble

  $$\mathcal{P}^* = \{\mathbf{x}^* \in \mathcal{X} \mid \nexists\ \mathbf{x} \in \mathcal{X} : \mathbf{x} \succ \mathbf{x}^*\}. \tag{4}$$

  On définit alors le front de Pareto par

  $$\mathcal{PF}^* = \{\mathbf{m}(\mathbf{x}^*) \mid \mathbf{x}^* \in \mathcal{P}^*\}. \tag{5}$$

  L'avantage de ce type de résolution est qu'elle permet d'obtenir l'ensemble des solutions non dominées, c'est-à-dire des compromis possibles au sein desquels le décideur doit choisir. Les algorithmes de recherche les plus utilisés pour résoudre ce type de problème sont les algorithmes dits "évolutionnaires". Ces algorithmes s'inspirent de processus biologiques. L'un des algorithmes les plus courant est le "Non-dominated Sorting Genetic Algorithm II" (NSGAII) (voir, par exemple, Deb et al., 2000).

Comme cela a été mentionné dans l'introduction, à la forme multi-objectifs du problème s'ajoute de l'incertitude liée à l'utilisation de modèles de substitution (1) pour établir les relations entre la variable de décision $\mathbf{x}$ et les p objectifs, notés $y^{(j)}, j = 1, \ldots, p$. Chapman et al. (2014) utilisent par exemple des modèles de régression linéaire

$$y^{(j)} = \mathbf{x}'\beta^{(j)} + \epsilon^{(j)} \quad \text{pour} \quad j \in \{1, \ldots, p\}, \tag{6}$$

sous l'hypothèse de normalité du terme d'erreur ($\epsilon^{(j)} \sim \mathcal{N}(0, \sigma_j^2)$), la notation $\mathbf{M}'$ désignant la transposé de $\mathbf{M}$. Le paramètre $\beta^{(j)}$ est estimé par $\hat{\beta}^{(j)}$ avec la méthode des moindres carrés. Pour intégrer l'incertitude du modèle, les auteurs simulent $S = 500$ scénarios où $\hat{\beta}^{(j)}$ est tiré aléatoirement selon la loi normale multivariée $\mathcal{MVN}(\beta^{(j)}, \sigma_j^2(\mathbf{X}'\mathbf{X})^{-1})$, où $\mathbf{X}$ est la matrice de dimension $n \times d$ de n observations de la variable de décision, $\beta^{(j)}$ est estimé par $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}^{(j)}$ avec $\mathbf{y}^{(j)}$ le vecteur de taille n de la $j$ème variable objectif, et σ_j^2 est estimé par l'erreur quadratique moyenne (MSE) du modèle correspondant.
La présence d'une variable aléatoire dans le modèle transforme le problème d'optimisation multi-objectifs (2) en problème d'optimisation multi-objectifs stochastique,

$$\min_{\mathbf{x}\in\mathcal{X}} (\mathbf{x}'\hat{\beta}^{(1)}, \ldots, \mathbf{x}'\hat{\beta}^{(p)}). \tag{7}$$

La stratégie classique pour résoudre ce type de problème est alors de passer à l'espérance mathématique. Pour le résoudre numériquement, on peut transformer le problème multi-objectifs en un problème mono-objectif stochastique comme le fait par exemple Stancu-Minasian (1984) avec une méthode de "goal programming" en considérant le problème :

$$\min_{\mathbf{x}\in\mathcal{X}} \mathbb{E}[\sum_{j=1}^{p} |T^{(j)} - \mathbf{x}\hat{\beta}^{(j)}|], \tag{8}$$

où $T^{(j)}$ désigne une valeur cible fixée par le décideur.
De leur côté, Turgut et Murat (2011) proposent de conserver la forme multi-objectifs du problème et de gérer l'incertitude par une pénalisation de l'espérance moyenne d'une solution par sa variance

$$\min_{\mathbf{x}\in\mathcal{X}} (\mathbf{w}_1^{(1)}\mathbb{E}[\mathbf{x}\hat{\beta}^{(1)}] + \mathbf{w}_2^{(1)}Var[\mathbf{x}\hat{\beta}^{(1)}]), \ldots, \mathbf{w}_1^{(p)}\mathbb{E}[\mathbf{x}\hat{\beta}^{(p)}] + \mathbf{w}_2^{(p)}Var[\mathbf{x}\hat{\beta}^{(p)}])) \tag{9}$$

où les pondérations sont telles que $\mathbf{w}_1^{(j)} \geq 0,\ \mathbf{w}_2^{(j)} \geq 0,\ \mathbf{w}_1^{(j)} + \mathbf{w}_2^{(j)} = 1,\ \forall j \in \{1, \ldots, p\}$. Cette approche permet de fait d'intégrer l'aversion au risque du décideur. Un décideur prudent ou "risk-averse" préférera une solution moins bonne (en moyenne) mais avec une probabilité raisonnablement faible que le minimum réel soit supérieur au minimum prédit ($\mathbf{w}_1 <<< \mathbf{w}_2$). A l'inverse un décideur ayant besoin de résoudre un grand nombre de fois le problème d'optimisation privilégiera le meilleur résultat "moyen" ($\mathbf{w}^1 >>> \mathbf{w}^2$). Cette distinction dans la prise de risque n'est pas présente dans (8), du fait de la seule utilisation de l'espérance. On notera toutefois que (9) repose sur une hypothèse de symétrie de la distribution de l'incertitude. En effet, la variance n'est pas nécessairement la statistique la plus adéquate pour gérer l'incertitude, puisque seule la variance "à la hausse" devrait être pénaliser, l'inverse étant plutôt une aubaine pour le décideur. Ainsi, il nous paraît plus raisonnable d'utiliser plutôt la "value at risk" (voir par exemple Shapiro et al., 2009), basée sur le quantile conditionnel la variable objectif sachant la variable de décision:

$$q_\alpha^{(j)}(\mathbf{x}) = \inf\{y | F^{(j)}(y|X = \mathbf{x}) \geq \alpha\} = (F^{(j)})^{-1}(y|X = \mathbf{x}), \tag{10}$$

où $F^{(j)}(.|X = \mathbf{x})$ désigne la fonction de répartition de la $j$ème variable objectif sachant que la variable de décision prend la valeur $\mathbf{x}$. Cette approche permet de choisir, de manière plus flexible, les niveaux de risque α que le décideur est prêt à prendre.

3 Méthodologie proposée

Dans le problème d'optimisation multi-objectifs considéré, nous avons utilisé les quantiles conditionnels introduits ci-dessus en prenant $r = 7$ valeurs différentes pour l'ordre α (de 5% à 95%), ceci afin de gérer l'incertitude liée aux modèles de substitution. Pour chaque valeur de α, on cherche alors le front de Pareto

$$\mathcal{PF}^*_\alpha = \{(q^{(1)}_\alpha(\mathbf{x}), \ldots, q^{(p)}_\alpha(\mathbf{x}))|\mathbf{x} \in \mathcal{P}^*_\alpha\} \quad \text{où} \quad \mathcal{P}^*_\alpha = \underset{\mathbf{x}\in\mathcal{X}}{\operatorname{argmin}}(q^{(1)}_\alpha(\mathbf{x}), \ldots, q^{(p)}_\alpha(\mathbf{x})). \tag{11}$$

Pour plus de flexibilité du modèle et afin d'éviter de dépendre d'une hypothèse sur la distribution de l'incertitude, les quantiles conditionnels $q^{(j)}_\alpha(\mathbf{x})$ sont estimés non paramétriquement avec un estimateur de type noyau (kernel) :

$$q^{(j)}_{\alpha,n}(\mathbf{x}) = \inf\{y|F^{(j)}_n(y|\mathbf{x}) \geq \alpha\}, \tag{12}$$

avec

$$F^{(j)}_n(y|\mathbf{x}) = \sum_{i=1}^{n} K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h^{(j)}_n}\right) \mathbb{1}_{\{y^{(j)}_i \leq y\}} \Bigg/ \sum_{i=1}^{n} K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h^{(j)}_n}\right) \tag{13}$$

où $y^{(j)}_i$ et $\mathbf{x}_i$ sont les valeurs pour la $i$ème observation respectivement pour la $j$ème variable objectif et pour la variable de décision, K est le noyau de lissage gaussien multidimensionnel

$$K(u) = (2\pi)^{-q/2} e^{-||u||^2/2} \tag{14}$$

avec $||.||^2$ la norme euclidienne, et $h^{(j)}_n$ est la largeur de fenêtre (bandwidth). Ce paramètre de lissage $h^{(j)}_n$ a une grande incidence sur la qualité de l'estimation de la fonction de répartition conditionnelle $F^{(j)}_n$ et doit donc être convenablement optimisé par une méthode de validation croisée "leave-one-out" :

$$h^{(j)}_{n,opt} = \underset{h>0}{\operatorname{argmin}} \sum_{k=1}^{n} \sum_{i=1, i\neq k}^{n} \{\mathbb{1}_{y^{(j)}_k < y^{(j)}_i} - F^{(j)}_{n,-k}(y^{(j)}_i|\mathbf{x}_k)\}^2, \tag{15}$$

où $F^{(j)}_{n,-k}$ désigne la fonction de répartition conditionnelle estimée à partir de l'échantillon privé de la $k$ème observation.

Dans un second temps, pour calculer les fronts de Pareto, l'algorithme génétique NSGAII a été utilisé. Les paramètres de cet algorithme n'ont pas fait l'objet d'un tunage approfondi : pour la population initiale $\mathbf{X}_0$, un tirage aléatoire de 50 observations de $\mathbf{X}$

a été réalisé, le nombre B d'itérations de l'algorithme a été fixé à 50, la fréquence de mutations a été fixée à 0.3 quelle que soit la composante de la variable de décision et le crossing over a été opéré à partir de la $[q/2]^e$ composante où $[a]$ désigne la partie entière de a.

4 Simulations numériques

Pour illustrer la méthodologie proposée, nous avons considéré $p = 2$ variables objectifs, une variable de décision de dimension $d = 3$ et un échantillon de taille $n = 300$. Soit $\mathbf{X}$ la matrice des covariables de dimension 300×3, chacune des d composantes ayant été générées aléatoirement selon la loi uniforme $\mathcal{U}([-2, 2])$. La matrice $\mathbf{Y}$ des objectifs de dimension 300×2 a été construite de la manière suivante : pour $i = 1, \ldots, n$,

$$\mathbf{y}_i' = \begin{pmatrix} y_i^{(1)} \\ y_i^{(2)} \end{pmatrix} = \begin{pmatrix} m_1(\mathbf{x}_i) + \epsilon^{(1)} \\ m_2(\mathbf{x}_i) + \epsilon^{(2)} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_i\beta^{(1)} + \epsilon^{(1)} \\ \mathbf{x}_i\beta^{(2)} + \epsilon^{(2)} \end{pmatrix}, \tag{16}$$

avec $\beta^{(1)} = (1, 2, 1)'$ et $\beta^{(2)} = (-1, -1, 3)'$. Pour les termes d'erreur $\epsilon^{(1)}$ et $\epsilon^{(2)}$, deux scénarii ont été considérés : pour $j = 1, 2$,

Scénario 1: cas homoscédastique

$$\epsilon^{(j)} \sim \mathcal{N}(0, 1), \tag{17}$$

Scénario 2: cas hétéroscédastique

$$\epsilon^{(j)}(\mathbf{x}) \sim \mathcal{N}(0, \sigma_{\mathbf{x}}^2) \quad \text{avec} \quad \sigma_{\mathbf{x}}^2 = m_j(\mathbf{x})^4. \tag{18}$$

Par soucis de comparaison, nous avons estimer les fronts de Pareto en utilisant les quantiles contionnels estimés non paramétriquement mais aussi les "vrais" quantiles conditionnels, c'est à dire en utilisant la connaissance de la distribution de l'incertitude que nous avons en simulation : $q_\alpha^{(j)}(\mathbf{x}) = \Phi^{-1}(\alpha|\mathbf{x})$ pour le scénario 1, et $q_\alpha^{(j)}(\mathbf{x}) = \Phi^{-1}(\alpha|\mathbf{x}/m_j(\mathbf{x})^4)$ pour le scénario 2. Enfin, un vecteur de poids $\mathbf{p} = (0.4, 0.6)$ a été utilisé pour choisir au sein de chaque front le meilleur compromis z^* en construisant une fonction d'utilité additive

$$z^* = \inf\{z | z = p_1 y^{(1)} + p_2 y^{(2)}, \text{ pour } \mathbf{y} = (y^{(1)}, y^{(2)}) \in \mathcal{PF}_\alpha^*\}. \tag{19}$$

Au vu des graphique de la Figure 4, on constate que l'algorithme NSGAII permet de raisonnablement bien trouver le front de Pareto. La prise en compte de l'incertitude semble cohérente, notamment visible sur les graphiques des "vrais" quantiles (à gauche) où peu de points se trouvent sous le front de Pareto le plus optimiste. On remarque aussi pour le scénario 2 (en bas) que la forte augmentation de la variance aux extrémités du support de $\mathbf{y}$ provoque un écart de plus en plus important entre la décision prudente et la décision risquée. Cela explique notamment que, pour des poids identiques au scénario 1,

un décideur voulant prendre des risques choisira les solutions extrêmes pour $\alpha = 5\%, 10\%$. Enfin, il faut souligner que l'estimation non paramétrique des quantiles conditionnels a un impact non négligeable sur la recherche du front de Pareto et du compromis. Cela s'explique naturellement par un nombre d'observations, $n = 300$, relativement modeste pour une estimation non paramétrique en dimensions $d = 3$. On fait face ici au fléau de la dimension.

5 Conclusion & Perspectives

La méthodologie proposée semble efficace pour gérer l'incertitude dans un contexte d'optimisation multi-objectifs. On notera toutefois que subsistent des sources d'erreur non prises en compte, il s'agit de celles liées à l'estimation des quantiles conditionnels et à la convergence stochastique de l'algorithme génétique.
En perspective de ce travail, il serait intéressant d'explorer des techniques de réduction dimensionnelle pour gérer la malédiction de la dimension pénalisant la qualité de l'estimation des quantiles conditionnels lorsque la dimension d est importante. Nous envisageons également de comparer notre approche à celle de Hughes (2001), Teich (2001) ou Coit et al. (2015) qui ont reformulé (3) en probabilité de dominance

$$P(\mathbf{x}' \succ \mathbf{x}'') = P(m_1(\mathbf{x}') + \epsilon' < m_1(\mathbf{x}'') + \epsilon'' \cap \cdots \cap m_p(\mathbf{x}') + \epsilon' < m_p(\mathbf{x}'') + \epsilon''), \quad (20)$$

permettant notamment de converger vers le front de Pareto "le plus probable". Mentionnons de plus qu'une généralisation de la méthodologie développée ici à un cadre (souvent rencontré dans la réalité) où des variables de décision peuvent être qualitatives est en cours d'élaboration. Enfin, une autre mesure du risque que la "value at risk" pourrait être considérée, comme l'"average value at risk" par exemple.

Bibliographie

Chapman JL., Lu L., Anderson-Cook C. (2014). Incorporating response variability and estimation uncertainty into Pareto front optimization, *Computers & Industrial Engineering*, 76, pp. 253-267.
Coit DW., Selcuklu S., Chatwattanasiri N., Wattanapongsakorn N. (2015). Stochastic multiple objective electric generation expansion planning, *Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, 12, pp. 1-6
Collette Y. et Siarry P. (2003). *Multiobjective optimization: principles and case studies*, Springer-Verlag, Berlin
Deb K., Agrawal S., Pratap A. et Meyarivan T. (2000). A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: NSGA-II, *parallel problem solving from nature - PPSN VI, Berlin*, pp.849-858

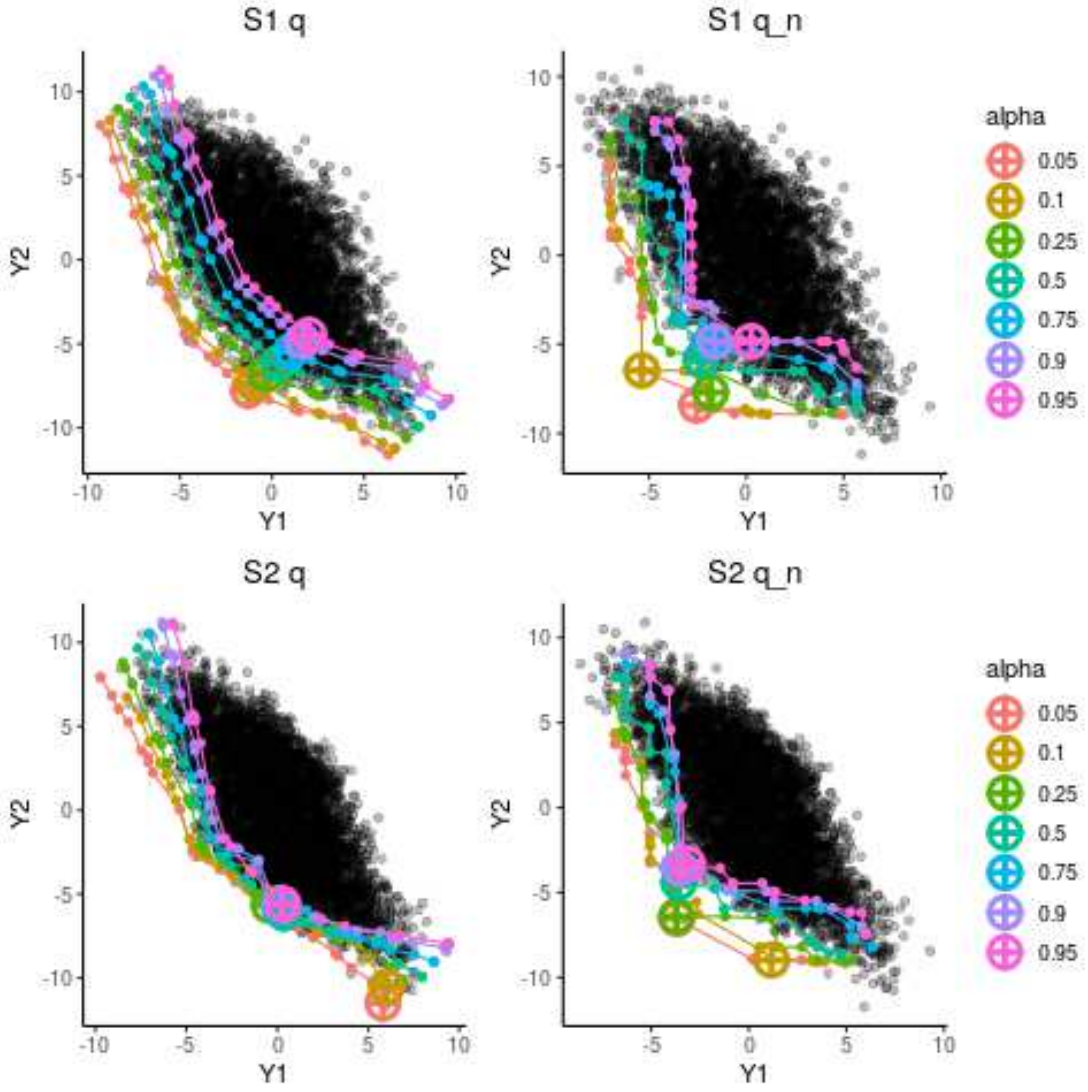


Figure 1: Représentation de l'espace de décision à partir de la génération de $\tilde{n} = 5000$ observations pour une meilleure visualisation. Les deux graphiques du haut résultent du scénario 1 et ceux du bas du scénario 2. On distingue également les simulations issues des "vrais" quantiles conditionnels (à gauche) de celles issues des estimations des quantiles conditionnels (à droite). Sur chacun des graphiques, $r = 7$ fronts de Pareto issus de la connaissance ou de l'estimation des quantiles conditionnels associés aux différentes valeurs de α ont été estimés sur un échantillon de taille $n = 300$. Les points de couleur positionnés sur les fronts sont les individus sélectionnés par l'algorithme génétique. La croix entourant l'un de ces points est la solution z^* choisie par le décideur pour chaque valeur α.

Hughes EJ. (2001), International Conference on Evolutionary Multi-Criterion Optimization, *Evolutionary Multi-Criterion Optimization*, 1993, pp. 329-343
Shapiro A., Dentcheva D. et Ruszczyński A. (2009). *Lectures on stochastic programming: modeling and theory*, SIAM, Philadelphia.
Stancu-Minasian IM. (1984). *Stochastic programming with multiple objective functions*, Reidel Publishing Company, Boston
Teich J. (2001). Pareto-Front Exploration with Uncertain Objectives. *Evolutionary Multi-Criterion Optimization*, 1993, pp. 314-328
Turgut O., Murat AE. (2011). Generating pareto surface for multi objective integer programming problems with stochastic objective coefficients, *Procedia Computer Science*, 6, pp. 46-51.

Caractérisation de zones critiques pour le dimensionnement en fatigue d'une pièce mécanique

Olivier Coudray [1,2] & Philippe Bristiel [1] & Gilles Celeux[2] & Miguel Dinis [1] & Christine Keribin [2] & Patrick Pamphile [2]

olivier.coudray@math.u-psud.fr

[1] *Groupe PSA - Centre d'Expertise Métiers et Régions, 2 à 10 Boulevard de l'Europe, 78300, Poissy, France.*
[2] *Université Paris-Saclay, CNRS, Inria, Laboratoire de mathématiques d'Orsay, 91405, Orsay, France.*
Collaboration OpenLab IA entre le Groupe PSA et Inria.

Résumé. Pour dimensionner une pièce en fatigue, les ingénieurs ont besoin d'identifier des zones critiques. Ils ont alors recours à des calculs numériques à l'aide de modèles par éléments finis et utilisent des critères de fatigue déterministes. Cependant, les essais sur des prototypes ne valident pas toujours les résultats numériques. L'objectif de ce travail est alors d'améliorer l'identification des zones critiques à l'aide de méthodes statistiques.
Mots-clés. dimensionnement en fatigue, critère de Dang Van, analyse multivariée.

Abstract. To design a mechanical part, engineers need to identify critical areas. They resort to numerical calculations based on finite element models and use deterministic fatigue criteria. However, tests on prototypes do not always validate numerical results. The objective of this work is to improve the identication of critical areas using statistical methods.
Keywords. fatigue design, Dang Van criterion, multivariate analysis.

Introduction

Sur une pièce mécanique soumise à des efforts faibles ou modérés, on note parfois l'apparition de fissures après une durée importante d'utilisation. Ce phénomène, connu sous le nom de fatigue mécanique, est dangereux parce qu'il peut amener une pièce à rompre de manière soudaine dans des conditions d'utilisation normales, sans sollicitation excessive. Le bureau d'étude chargé du dimensionnement des pièces lors de leur conception doit alors garantir la sûreté de la pièce lors de l'utilisation, tout en minimisant son coût de production. Formellement, on évalue la durée de vie d'une pièce comme le nombre de cycles de chargement auxquels elle peut résister avant de rompre. La durée de vie se décompose en une phase d'amorçage de la fissure suivie d'un phase de propagation jusqu'à la rupture de la pièce. Dans l'industrie automobile on s'intéresse à la partie amorçage en considérant que la phase de propagation est négligeable.

Le phénomène de fatigue mécanique est bien connu aujourd'hui (cf [4]) et pris en compte dans le dimensionnement d'une pièce mécanique. La phase de conception d'une pièce démarre par une modélisation par éléments finis d'une géométrie correspondant au cahier des charges. Cette modélisation sert à identifier des zones critiques de la pièce, c'est-à-dire pour lesquelles le risque d'amorçage est élevé sur le long terme. Cette phase est suivie d'essais effectués sur plusieurs prototypes afin de valider la résistance à la fatigue de la pièce dans diverses conditions de sollicitations opérationnelles. Si, lors des essais, on observe des fissures, alors on procède à une modification de la pièce donnant lieu à un nouveau calcul. Ces aller-retours impactent les coûts de développement et retardent la date de production.

Aussi, il est intéressant de disposer d'outils permettant d'estimer la tenue en fatigue à partir de la modélisation numérique de la pièce. L'objectif est de réduire la phase de conception en facilitant l'identification des zones critiques à l'aide de méthodes statistiques. Idéalement, une seule phase d'essai devrait ensuite permettre de valider la pièce sans modification ultérieure.

Dans un premier temps, nous présenterons les données à disposition. Nous verrons ensuite un exemple de critère de fatigue, communément utilisé pour identifier les zones critiques d'une pièce mécanique et nous en analyserons les limites. Enfin, nous présenterons une analyse en composantes principales (ACP) des données afin de tirer partie de l'ensemble des variables disponibles pour améliorer le critère de fatigue.

1 Données

1.1 Données issues de calculs par éléments finis

Les pièces étudiées ici sont des composants du châssis (berceaux et traverses), pièces importantes pour la sûreté d'un véhicule. La modélisation par éléments finis permet la résolution numérique des équations mathématiques d'un problème physique, ici celui de la réponse d'une pièce mécanique soumise à des contraintes données. La pièce est maillée (figure 1) et les caractéristiques des matériaux des éléments sont définies. Les résultats de calculs comportent des informations physiques (contraintes, invariants de contrainte, gradients de contrainte) en chaque élément du modèle. Pour chaque observation (élément du modèle), on dispose de huit variables descriptives relatives au maillage et de cinquante variables physiques. Pour chaque pièce, on compte plusieurs centaines de milliers d'observations.

1.2 Données issues de bancs d'essais

Les prototypes testés en conditions opérationnelles sur des bancs d'essais apportent des informations supplémentaires. On dispose de comptes rendus sur lesquels ont été notées et photographiées les fissures apparues sur les prototypes testés (figure 2). On

connaît également le nombre de cycles et le chargement auxquels a été soumise la pièce avant l'amorçage de chaque fissure. Pour chaque campagne d'essai, entre trois et sept prototypes identiques sont testés.

Les données d'essais sont incorporées aux données de calculs sous la forme de deux variables par prototype testé : l'une booléenne codant la présence de l'amorçage d'une fissure sur l'élément, l'autre quantitative mesurant l'effort appliqué à la pièce au moment d'apparition de la fissure correspondante ou par défaut, l'effort appliqué à la pièce à la fin de l'essai.

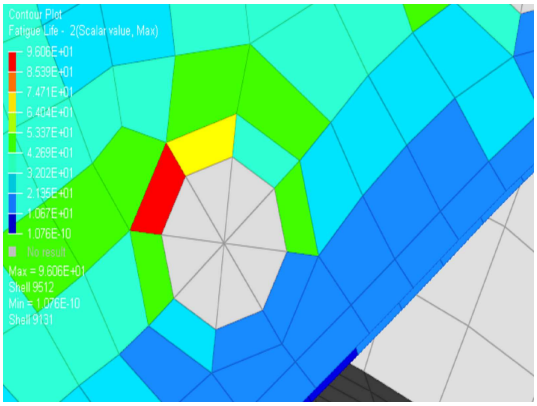


FIGURE 1 – Visualisation des résultats de calcul (taux de cisaillement critique) sur le modèle par éléments finis.

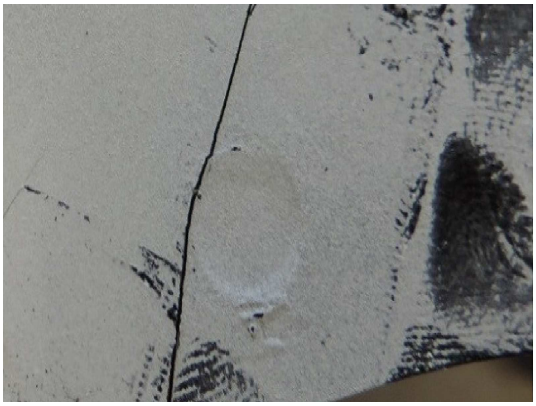


FIGURE 2 – Photographie d'une zone de la pièce (correspondant au modèle de la figure 1) après essai.

2 Le critère de fatigue de Dang Van

Classiquement, l'identification de zones critiques sur un modèle par éléments finis se fait à l'aide de calculs numériques et de l'utilisation de critères de fatigue. Un critère de fatigue est une évaluation, en chaque élément du modèle, du degré de criticité de cet élément à partir des contraintes calculées. Il est admis que l'amorçage de fissures de fatigue est principalement lié au cisaillement (contraintes tangentielles) et que les contraintes normales de traction accélèrent (extension) ou retardent (compression) leur apparition (cf [4], chap. 2). Le critère de Dang Van est utilisé comme critère de fatigue bien qu'il soit imparfait.

2.1 Définition

On note E l'ensemble des éléments du modèle et pour chaque élément e de E, on note respectivement Ph_e et τ_e la pression hydrostatique critique et le taux de cisaillement critique (invariants de contrainte calculés à partir du modèle par éléments finis). Le critère

de Dang Van est une frontière linéaire dans le plan formé par ces deux variables (cf [1]). Avec ce critère, un élément e de matériau m est critique s'il est situé au-dessus de la frontière, soit :

$$\tau_e + \alpha_m \cdot Ph_e > \tau_{0,m}$$

où $\alpha_m = 3 \cdot \left(\frac{t_m}{f_m} - \frac{1}{2}\right)$ et $\tau_{0,m} = t_m$.

Les constantes f_m et t_m représentent respectivement les limites de fatigue en flexion alternée et en torsion alternée. Elles dépendent du matériau m constituant l'élément et sont estimées à partir d'essais de fatigue uniaxiaux sur éprouvettes en flexion et torsion. On introduit alors le coefficient de danger $CD_e = \frac{\tau_e + \alpha Ph_e}{\tau_{0,m}} - 1$, degré de criticité de l'élément compte tenu de sa position à la frontière matériau.

2.2 Normalisation du critère

La frontière sur le plan de Dang Van dépend du matériau m. Or, une pièce peut compter plusieurs matériaux aux propriétés différentes. Aussi, pour pouvoir comparer ces points, nous proposons d'introduire une version normalisée du critère de Dang Van (cf [3]). Les grandeurs normalisées $Ph_e^{(n)}$ et $\tau_e^{(n)}$ sont définies de la façon suivante :

$$\left(Ph_e^{(n)},\ \tau_e^{(n)}\right) = \left(\alpha_m \frac{Ph_e}{\tau_{0,m}},\ \frac{\tau_e}{\tau_{0,m}}\right) .$$

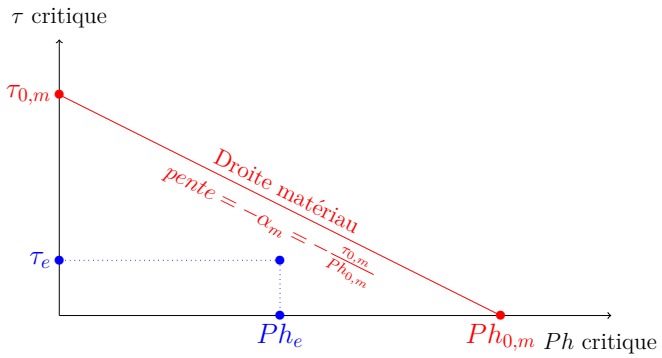


FIGURE 3 – Diagramme de Dang Van classique.

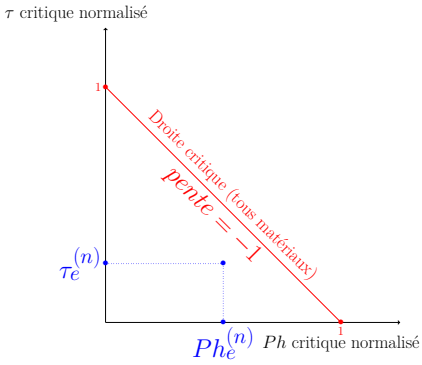


FIGURE 4 – Diagramme de Dang Van normalisé.

Dans ce nouveau plan, tous les matériaux partagent la même frontière (figures 3 et 4). On peut ainsi superposer les données de plusieurs pièces. De plus, le coefficient de danger se réécrit très simplement en fonction des grandeurs normalisées : $CD = \tau_e^{(n)} + Ph_e^{(n)} - 1$.

2.3 Limites du critère

On peut voir sur la figure 5 un premier exemple d'évaluation du critère de Dang Van sur un fichier (450000 observations).

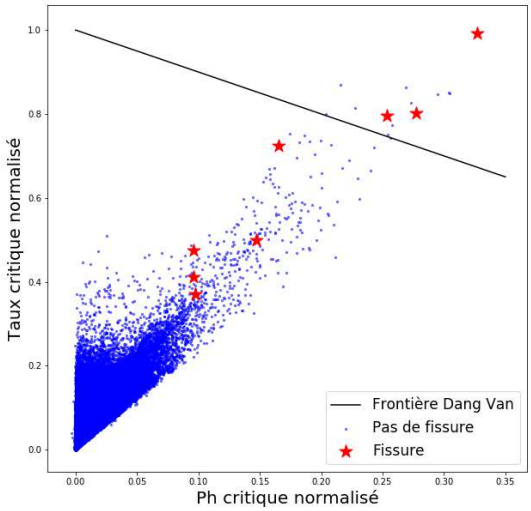


FIGURE 5 – Diagramme de Dang Van normalisé (points bleus issus des calculs numériques ; étoiles : éléments ayant fissurés lors des essais).

La figure 5 montre que des éléments sur lesquels une fissure a amorcé sont bien des points dans la partie supérieure droite du plan de Dang Van. Pourtant, certains de ces points se situent en-dessous de la droite critique. On voit donc clairement que le critère de Dang Van ne permet pas d'identifier tous les points d'amorçage de fissure. Cela peut s'expliquer par le caractère aléatoire de l'amorçage mais aussi par le fait que le critère ne tient pas compte des spécificités géométriques de certains points de la pièce qui peuvent pourtant favoriser grandement l'amorçage de fissure (extrémités de cordons de soudure, bords de tôle avec courbure importante...).

3 Étude multidimensionnelle des données de fatigue

Nous avons vu les limites de l'utilisation du critère de Dang Van. Celui-ci n'utilise que deux informations alors qu'on dispose de cinquante variables qui pourraient contribuer à définir un critère de fatigue plus efficace.

On effectue une analyse en composantes principales des données (ACP, cf [2] chap. 1) pour repérer d'éventuelles directions d'observation intéressantes et complémentaires à celles déjà considérées. Dès le premier plan principal, l'ACP permet non seulement de retrouver les informations du critère de Dang Van (corrélation avec les pression hydrostatique et taux de cisaillement critiques) mais aussi d'amener à une possible caractérisation des fissures (voir figure 6). En effet, le second axe apporte une information supplémentaire à la représentation de Dang Van puisqu'il permet d'identifier deux «types» de fissures : du point de vue des variables, c'est l'orientation des contraintes qui les différencie.

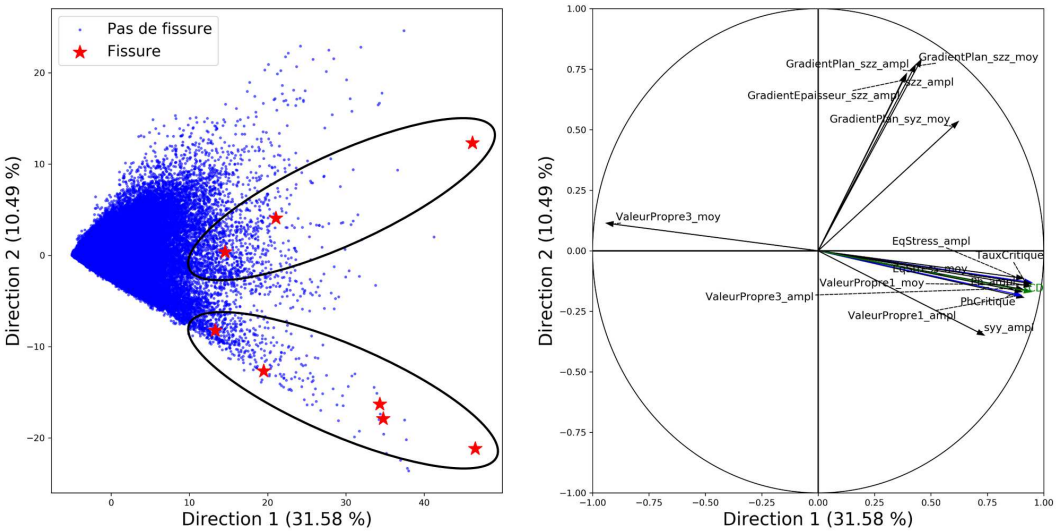


FIGURE 6 – Deux premières composantes principales (étoiles : points d'amorçage) et cercle des corrélations (représentation des 15 variables contribuant le plus) : identification de deux groupes de fissures.

Conclusion

Nous avons montré que l'analyse multivariée enrichit la caractérisation des zones critiques. En revanche, certains points restent relativement mal caractérisés, même en prenant en compte les différents axes de l'ACP. L'ajout d'un nouveau type de variable, prenant notamment en compte des informations sur la forme géométrique des zones considérées, est une piste intéressante. Dans une première approche, on envisage de définir des variables catégorielles indiquant la présence ou non de chaque type de singularité (cordon de soudure, bord de tôle) à proximité de l'élément. Cette étude va nous permettre de guider la construction d'un nouveau critère de fatigue probabilisé.

Références

[1] P. Ballard, K. Dang Van, A. Deperrois, and Y. V. Papadopoulos. High cycle fatigue and a finite element analysis. *Fatigue Fract. Eng. Mater. Struct.*, 18(3) :397–411, Mar 1995.

[2] Brigitte Escofier and Jérôme Pagès. *Analyses factorielles simples et multiples.* May 2016.

[3] S. Fouvry, P. Kapsa, and L. Vincent. A Global Methodology to Quantify Fretting Damages. *ASTM International*, Jan 2003.

[4] J. Schijve. *Fatigue of Structures and Materials.* Springer Netherlands, 2009.

Quels modèles pour le temps de stationnement des trains en Île de France ?

Rémi Coulaud [1,2] & Christine Keribin [1] & Gilles Stoltz [1]

[1] *Université Paris-Saclay, CNRS, Inria, Laboratoire de mathématiques d'Orsay, 91405, Orsay, France*

[2] *Transilien, SNCF Voyageurs, 10 rue Camille Moke, 93220, Saint-Denis, France*

Résumé. Nous modélisons le temps de stationnement des trains à l'arrêt à partir de données ferroviaires et de fréquentation issues des trains connectés de la ligne H. Le plan de transport structure l'exploitation ferroviaire ; nous modélisons le temps de stationnement différemment en fonction de l'avance ou du retard du train par rapport à l'heure d'arrivée théorique. Nous montrons que les variables pertinentes dépendent de la place de l'arrêt dans le jalonnement ainsi que de l'heure d'arrivée puis nous comparons les performances de nos modèles avec ceux obtenus par d'autres modèles de la littérature.

Mots-clés. Choix de modèles, données ferroviaires, étude de cas, forêts aléatoires

Abstract. We model train dwell times at train stops with railway data and passengers flow data coming from connected trains of the Île-de-France line H. The timetable constrains railway exploitation, which led us to set up different models of dwell time depending on arrival compared to planed hour. We show that variables selected depends on timetable accordance and train run then we compare our models results to the litterature best models.

Keywords. Models selection, railway data, case study, random forests

1 Introduction

Le temps de stationnement, différence entre l'heure de départ d et l'heure d'arrivée a, d'un train k à un arrêt s représente en Île-de-France en moyenne 30 % du temps de trajet entre une origine et une destination : la capacité d'une ligne ferroviaire et le temps de parcours des voyageurs sont ainsi directement impactés par le temps de stationnement. Sa prévision reste un problème largement ouvert comme en attestent les travaux de Kecman et Goverde (2015), Li et al. (2016), Cornet et al. (2019) et Palmqvist et al. (2019). L'étude des données de temps de stationnement ainsi que les intuitions de Kecman et Goverde (2015) et Pedersen et al. (2018) nous amènent à proposer trois régimes pour la modélisation : les trains en avance, à l'heure ou en retard par rapport à l'heure d'arrivée théorique. Cette segmentation permet d'isoler, pour chaque régime, les principaux facteurs influençant le temps de stationnement, pré-requis à la gestion du temps de stationnement en

opérationnel. Après la présentation des données, nous expliciterons le découpage en trois régimes. Nous sélectionnerons ensuite les variables importantes et nous finirons par la comparaison des performances des modèles induits avec certains de ceux proposés dans la littérature.

2 Présentation du jeu de données

L'étude porte sur les arrêts en zone dense des trains de la ligne H Transilien de septembre 2017 à septembre 2019. Cette ligne étant peu fréquentée, il convient de s'intéresser principalement à des périodes et à des branches où le nombre de trains et de passagers est conséquent : par exemple, les trains qui circulent de Paris-Nord à Pontoise, voir figure 1, en heures de pointe du soir (de 17h à 20h) pendant les jours de semaine hors vacances et jours fériés.

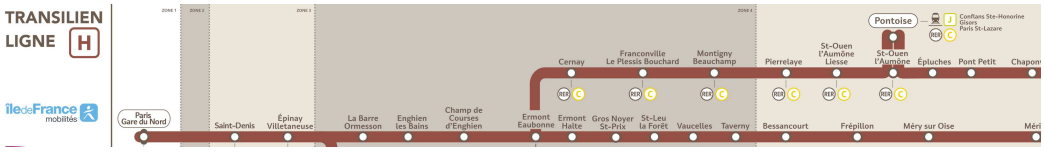


Figure 1 : Plan de la ligne Paris-Nord-Pontoise, passant par Ermont-Eaubonne.

Les données proviennent de deux systèmes embarqués des rames Z50000. Le premier système, ATESS[1], mesure notamment l'heure d'arrivée et de départ d'un train en gare. Le second système, CAVE[2], compte le nombre de montées et de descentes[3] à chaque arrêt à l'aide de caméras infrarouges au-dessus des portes. Ces fichiers donnent aussi des informations sur l'heure de création et d'envoi de l'événement de comptage au moment où le train repart, ces informations permettent de mesurer incidemment le temps de stationnement. Le jeu de données d'étude est constitué par le rapprochement de ces deux bases de données à partir du triplet (train k, gare s, jour t). Le taux de données manquantes (temps de stationnement ou comptage) est de l'ordre de 10 % pour ATESS, 20 % pour CAVE et de 25 % pour le jeu de données d'étude. En les mettant de côté, il reste un nombre total de 21 841 observations pour les onze gares de l'axe Paris-Nord-Pontoise. Nous nous focalisons sur les trois premiers arrêts : Saint-Denis, Epinay-Villetaneuse et Enghien-les-Bains. Le nombre d'observations pour chaque gare est 2 233, 2 243, 2 238. Ce sont trois gares importantes en terme de fréquentation pour la ligne H dont la distribution des montées et des descentes par heure est différente. Les deux premières gares sont des gares avec des double-flux (à chaque heure de la journée il y a autant de montées que de descentes) tandis que la gare d'Enghien-les-Bains est une gare pendulaire (à chaque heure

[1]Acquisition et Traitement des Enregistrements de Sécurité Statique.
[2]Comptage Automatique Voyageur Embarqué.
[3]Le nombre de voyageurs qui montent et qui descendent.

de la journée il y a soit principalement des montées, soit principalement des descentes). Ces différences ont été isolées à partir de méthodes de classification non supervisée.

2.1 Redressement des temps de stationnement

Comme souligné précédemment, nous avons accès à deux sources de mesures des heures d'arrivée et de départ d'un train à un arrêt. Les informations internes à la SNCF ainsi qu'une enquête terrain ont permis de montrer que le temps de stationnement issu d'ATESS était fiable mais il n'est disponible qu'à J+8. Les données issues du système CAVE sont accessibles quasiment en temps réel, mais présentent une différence de mesures des temps de stationnement, que l'on peut raisonnablement représenter par une relation affine, avec une erreur moyenne de l'ordre de 4,5 s. Ceci est acceptable pour une modélisation opérationnelle compte tenu de la précision des outils et des comportements des conducteurs. Nous utilisons, par la suite, les mesures du temps de stationnement d'ATESS.

2.2 Variables explicatives

Les variables susceptibles d'expliquer le temps de stationnement ($Y_{k,s,t}$) pour le train k, l'arrêt s, le jour t, se scindent en deux groupes, les variables de fréquentation utilisées notamment par Cornet et al. (2019) et Palmqvist et al. (2019) et les variables ferroviaires essentiellement utilisées par Kecman et Goverde (2015) et Li et al. (2016). Un des apports de notre travail est d'utiliser à la fois des variables ferroviaires et de fréquentation.

Logiquement, les variables de fréquentation à considérer sont : le nombre de montées ($M_{k,s,t}$), le nombre de descentes ($D_{k,s,t}$) et la charge à bord à l'arrivée ($C_{k,s,t}$).

Pour introduire les variables ferroviaires, nous distinguons les heures d'arrivée et de départ réalisées (a^{real}, d^{real}) des heures d'arrivée et de départ théoriques (a^{theo}, d^{theo}) prévues dans le plan de transport. Ainsi ces variables sont : l'écart à l'heure d'arrivée théorique ($a^{\text{theo}}_{k,s,t} - a^{\text{real}}_{k,s,t}$), l'espacement entre deux trains $d^{\text{real}}_{k-1,s,t} - a^{\text{real}}_{k,s,t}$ ($E_{k,s,t}$).

Suivant les travaux de Chandesris (2014), nous ajoutons le temps de stationnement aux arrêts précédents ($Y_{k,s-1,t}, \ldots, Y_{k,1,t}$) qui apporte une information spatio-temporelle ainsi que le temps de stationnement au même arrêt pour le train précédent ($Y_{k-1,s,t}$). Remarquons que pour l'arrêt Saint-Denis qui est le premier arrêt après l'origine (Paris-Nord) les variables de retard n'existent pas, voir tables 1 et 2.

3 Trois régimes du temps de stationnement

Un train est en avance si $a^{\text{real}} \leq a^{\text{theo}}$, en retard si $a^{\text{real}} \geq d^{\text{theo}}$ et est à l'heure sinon. Cette séparation se justifie car un conducteur d'un train en avance doit normalement attendre l'heure de départ théorique avant de repartir. Inversement, le conducteur d'un train en retard doit normalement repartir immédiatement après que l'échange voyageur est terminé afin de rattraper son retard. Ainsi, un train en avance aura généralement

tendance à stationner plus longtemps, comme constaté sur la figure 2. Pedersen et al. (2018) considèrent les temps de stationnement pour les trains en retard comme des temps de stationnement minimaux. Cependant, il faut être prudent avec cette notion de temps de stationnement minimal car, comme le notent Cornet et al. (2019), rien ne garantit en pratique que le train puisse partir dès que l'échange est terminé ni que le conducteur respecte scrupuleusement la règle en cas de retard. Cette segmentation prend tout son sens pour la gare d'Enghien-les-Bains, voir figure 2, où le temps de stationnement décroît linéairement avec l'écart à l'heure d'arrivée pour les trains en avance. Nous modélisons le temps de stationnement pour ces trois régimes sauf pour les trains en avance à l'arrêt Saint-Denis où il y a moins de 30 observations, d'où la première colonne grisée des tables 1 et 2.

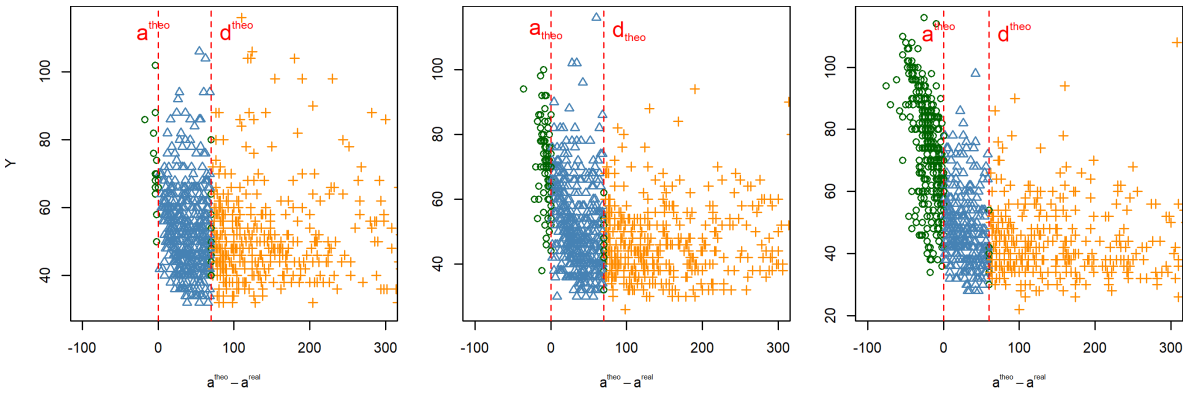


Figure 2 : Temps de stationnement (Y) en fonction de l'écart à l'heure d'arrivée théorique ($a^{\text{theo}} - a^{\text{real}}$) : une décomposition en trois régimes des arrêts en gare de Saint-Denis (à gauche), d'Epinay-Villetaneuse (au centre) et d'Enghien-les-Bains (à droite) ; nous distinguons les trains en avance (○), les trains à l'heure (△) et les trains en retard (+).

4 Modélisation du temps de stationnement

Nous nous intéressons à deux types de modèles, la régression linéaire, modèle paramétrique le plus simple, et les forêts aléatoires. Pour évaluer la qualité de la modélisation et sélectionner les variables intéressantes, nous divisons le jeu de données en : un jeu d'entraînement du 01/09/2017 - 31/08/2018 et un jeu de test du 01/09/2018 - 31/08/2019. La modélisation du temps de stationnement passe par la sélection des variables les plus pertinentes par régime et par gare parmi l'ensemble des variables présentées en section 2.2. Une fois ces variables identifiées par régime, nous comparons les performances de nos modèles à celles de méthodes élémentaires et de méthodes connues dans la littérature.

4.1 Choix de variables en régression linéaire et forêts aléatoires

Nous utilisons une régression linéaire et nous sélectionnons les variables par une recherche *backward* avec le critère BIC, voir Schwarz (1978). Nous utilisons aussi un modèle de forêts aléatoires, voir Breiman (2001), entraîné avec le package `randomForest` (le nombre d'arbres ainsi que le nombre de variables sélectionnées à chaque nœud sont laissés par défaut, 500 et 3 au plus, respectivement). Nous décidons de retenir les variables dont l'importance est supérieure à la moyenne de l'importance de toutes les variables.

	Arrêts en avance			Arrêts à l'heure			Arrêts en retard		
Variables \ Gares	G1	G2	G3	G1	G2	G3	G1	G2	G3
$M_{k,s,t}$				X			X	X	
$D_{k,s,t}$			X	X		X	X		X
$C_{k,s,t}$									
$a^{\text{theo}}_{k,s,t} - a^{\text{real}}_{k,s,t}$		X	X	X	X	X			
$E_{k,s,t}$									
$Y_{k-1,s,t}$			X						
$Y_{k,s-1,t}$			X		X				X
$Y_{k,s-2,t}$			X			X			X

Table 1: Résultats de la sélection de variables par la recherche *backward* avec le critère BIC avec la régression linéaire pour les trois régimes et les trois premières gares du parcours ; les gares de Saint-Denis (G1), Epinay-Villetaneuse (G2) et Enghien-les-Bains (G3).

	Arrêts en avance			Arrêts à l'heure			Arrêts en retard		
Variables \ Gares	G1	G2	G3	G1	G2	G3	G1	G2	G3
$M_{k,s,t}$				X			X	X	X
$D_{k,s,t}$				X		X	X	X	X
$C_{k,s,t}$				X		X	X	X	X
$a^{\text{theo}}_{k,s,t} - a^{\text{real}}_{k,s,t}$		X	X		X	X			
$E_{k,s,t}$		X				X	X	X	
$Y_{k-1,s,t}$									
$Y_{k,s-1,t}$									
$Y_{k,s-2,t}$									X

Table 2: Sélection de variables par l'importance des forêts aléatoires pour trois régimes et trois gares.

Nous remarquons pour les deux modèles que les variables de fréquentation, voir les tables 1 et 2, ne sont que très rarement sélectionnées pour modéliser le temps de sta-

tionnement des trains en avance tandis qu'elles le sont presque systématiquement pour celui des trains en retard. La variable écart à l'heure d'arrivée théorique ($a^{\text{theo}} - a^{\text{real}}$) est sélectionnée uniquement dans le cas où les trains ne sont pas en retard, ceci vient confirmer l'intuition des trois régimes. Nous constatons enfin que les variables de retards ne sont sélectionnées que dans le cas de la régression linéaire.

4.2 Performances des modèles pour les différents régimes

À notre connaissance aucun article ne compare sur un même jeu de données les performances des modèles statistiques existant. Les modèles mis en compétition par régime sont : la régression linéaire et la forêt aléatoire avec les variables sélectionnées dans la section 4.1 par régime ; des modèles de la littérature dont Kecman et al. (2015) qui utilisent des forêts aléatoires avec l'écart à l'heure d'arrivée théorique, Li et al. (2016) qui utilisent une régression linéaire avec $Y_{k-1,s,t}$, $Y_{k,s-1,t}$ et $Y_{k,s-2,t}$ et Palmqvist et al. (2019) qui utilisent une régression linéaire avec toutes les variables de fréquentation ; un modèle élémentaire (la moyenne du temps de stationnement passé pour un train) ; le temps de stationnement théorique prévu par le plan de transport de la SNCF (PdT SNCF). Les moyennes des différences absolues (MAE) et absolues relatives (MAPE) sont calculées pour les différents modèles sur le jeu de données test.

	Trains en avance		Trains à l'heure		Trains en retard	
	MAE	MAPE	MAE	MAPE	MAE	MAPE
Régression linéaire	**11.01**	**0.17**	9.09	**0.18**	**8.57**	**0.19**
Forêt aléatoire	11.42	0.18	**9.05**	**0.18**	8.86	0.20
Kecman et al.	11.42	0.18	9.44	0.19	9.74	0.22
Li et al.	13.73	0.21	9.21	**0.18**	9.15	0.20
Palmqvist et al.	15.36	0.22	9.56	0.19	8.72	**0.19**
Moyenne	31.38	0.38	9.94	0.19	9.35	0.21
PdT SNCF	19.78	0.25	14.35	0.33	17.71	0.46

Table 3: Performances des modèles sur le jeu de données test pour la gare d'Enghien-les-Bains.

Nos modèles font mieux ou aussi bien que les modèles de la littérature. Pour les trains en avance la différence de performances entre le modèle élémentaire et nos modèles est importante. Pour les trains en retard l'utilisation de variables explicatives ne permet pas une réelle amélioration des performances par rapport au modèle élémentaire.

4.3 Perspectives

Au vu des résultats précédents, il conviendra de prendre en compte de façon plus fine les dépendances spatiales et temporelles des temps de stationnement afin d'améliorer leur

estimation pour les trains en retard. Cette étape de modélisation focalisée sur le temps de stationnement nous permettra de passer plus facilement de la modélisation à la prévision.

Les trains connectés comptent non seulement le nombre de montées et de descentes à l'échelle du train mais aussi à l'échelle de la porte ce qui nous permettra de quantifier l'influence de la répartition à quai des échanges sur le temps de stationnement.

Enfin, la modélisation suppose connues les montées et descentes lors du temps de stationnement. C'est une première étape, qui devra s'enrichir par la prévision de ces informations.

Bibliographie

Breiman, L. (2001), *Random forests*, Machine learning, 45(1), 5-32.

Chandesris, M. (2014), *Approche non-paramétrique pour la prédiction d'heure d'arrivée dans les transports*, 46 ème journées de Statistique de la SFDS, Rennes.

Cornet, S., Buisson, C., Ramond, F., Bouvarel, P. et Rodriguez, J. (2019), *Methods for quantitative assessment of passenger flow influence on train dwell time in dense traffic areas*, Transportation Research Part C : Emerging Technologies, vol. 106, pp. 345-359

Kecman, P. et Goverde, R. M. P. (2015), *Predictive modelling of running and dwell times in railway traffic*, Public Transport 7, 295–319.

Li, D., Daamen, W. et Goverde, R. M. P. (2016), *Estimation of train dwell time at short stops based on track occupation event data: A study at a Dutch railway station*, Journal of Advanced Transportation 50, 877–896.

Palmqvist, C-W., Tomii, N., et Ochiai, Y. (2019), *Dwell Time Delays for Commuter Trains in Stockholm and Tokyo*, Paper presented at RailNorrköping, Norrköping, Suède.

Pedersen, T., Nygreen, T. et Lindfeldt, A. (2018), *Analysis of temporal factors influencing minimum dwell time distributions*, WIT Transactions on The Built Environment. 181, 447–458

Schwarz, G. (1978), *Estimating the dimension of a model*, The annals of statistics, 6(2), 461-464.

Une approche de classification croisée pour des séries temporelles fondée sur une approche dynamique

Christian Derquenne

Electricité de France - Recherche et Développement - 7, boulevard Gaspard Monge - 91120 Palaiseau - christian.derquenne@edf.fr

Résumé. La recherche de structures dans les données représente une aide essentielle pour comprendre les phénomènes à analyser. Les méthodes de classification croisée permettent de répondre à cette problématique lorsque l'on désire traiter conjointement les données sur les deux dimensions : lignes/colonnes (individus/variables). Dans certaines applications, la dimension temporelle est rajoutée (courbes de consommation, de températures, gènes, ...). La méthode proposée fournit des classes croisées de comportements des individus et des variables conjointement. Celle-ci se déroule en trois étapes : classification des variables, classification des individus et réconciliation des résultats des deux. Cette approche est appliquée sur des données simulées et fournit de très bons résultats.

Mots-clés. Classification croisée, analyse des correspondances multiples, apprentissage non supervisé.

Abstract. The search for structures in the data represents an essential help to understand the phenomena to be analyzed. Co-clustering methods make it possible to answer to this problem when we wish to process data on two dimensions jointly: rows / columns (individuals / variables). In certain applications, the temporal dimension is added (consumption curves, movies, microarray, ...). The proposed method provides co-clusters of behavior of individuals and variables jointly. This takes place in three stages: clustering of variables, clustering of individuals and reconciliation of the results of the two. This approach is applied to simulated data and provides very good results.

Keywords. Co-clustering, Multiple Correspondances Analysis, unsupervised learning.

1 Contexte - objectif

La recherche exploratoire de structures dans les données est essentielle dans de nombreuses applications (biologie, environnement, finance, management de l'énergie, ...) afin de comprendre les comportements des individus, les liens entre les variables, ... Les outils de visualisation, de réduction de dimension, de recherche de patterns permettent de répondre efficacement à ce type de problématiques. Nous nous plaçons dans le cadre de la classification non supervisée et plus particulièrement dans le domaine de la classification croisée. Celle-ci permet de regrouper conjointement des lignes et des colonnes d'un tableau de données. Une ligne peut être constituée d'autant de valeurs uniques qu'il y a de colonnes : un individu $\mathbf{x}_i$, $(i = 1, n)$ est renseigné par des mesures uniques issues de p variables X_j, $(j = 1, p)$. Ces mesures peuvent être de nature différente : numérique, catégorielle, comptage. De même, un individu peut posséder plusieurs valeurs pour une même variable, avec T points de mesures d'une série temporelle. Par exemple, l'objectif de l'étude peut être de classifier conjointement des courbes de consommations demi-horaires de clients (les individus) pour chaque jour de l'année, d'une part, et regrouper les 365 séries temporelles journalières (les variables) pour un même client, d'autre part (Bouveyron et al., 2018). La première partie de ce papier définit la classification croisée en général, puis nous proposons une méthode de classification croisée fondée notamment sur une approche dynamique de classification de variables (Derquenne, 2016, 2017). La troisième partie est consacrée à une application sur des données simulées. Enfin, nous concluons sur les améliorations à apporter, les applications potentielles et les voies futures.

2 La classification croisée

La première étape de toute construction de typologie consiste à se poser la question suivante : "quelle est la nature de mes données ?". En effet, il est souvent nécessaire de les transformer au préalable pour éviter toute incohérence dans les résultats de classification. La non prise en compte d'échelles différentes peut faire apparaitre une fausse structure dans les données ou au contraire en cacher une. Par exemple, une série temporelle peut être résumée sous la forme d'une fonction (polynomiale avec quelques coefficients "significatfs", en ondelettes), normalisée entre 0 et 1, centrée-réduite, discrétisée, en rangée en ordre croissant, ...

La deuxième étape de classification revient à construire un indicateur de ressemblance ou de dissemblance entre les objets (individus ou variables). Pour cela de nombreux indices de similarité, de dissimilarité, de distance sont disponibles et sont proposés au fil des années. Ils dépendent bien évidemment de la nature des données transformées ou non.

La troisième étape de classification non supervisée a pour objectif de rechercher des groupes d'objets en se reposant sur les étapes précédentes. Les méthodes proposées dans la littérature sont nombreuses ; elles s'appuient sur des approches de classification hiérarchique ou de partionnement. A ce stade, il est nécessaire d'introduire la notion de double-dimensionnalité ou de dualité des lignes et des colonnes (individus et variables, en général) d'un tableau de données. En effet, si l'on raisonne seulement sur l'une des deux dimensions, la classification identifiera des groupes d'individus en fonction des informations mesurées sur les variables ou des classes de variables à l'aide des valeurs observées sur les individus. Ce processus correspond à une approche locale de recherche de patterns. En complément, plutôt qu'à l'opposé, la classification croisée (Govaert et al., 2013, Madeira, 2011) identifie des groupes d'individus (lignes) ayant des comportements similaires au sein d'un sous-ensemble de variables (colonnes). Cela correspond à une approche globale dans laquelle chaque individu dans une classe croisée est sélectionné en utilisant seulement un sous-ensemble de variables. Et chaque variable dans un groupe croisé est retenue en utilisant un sous-ensemble d'individus. Par ailleurs, les classes croisées sont représentées par différentes structures : une seule classe détectée contenant un sous-ensemble de lignes et de colonnes ; plusieurs classes possédant exlusivement des lignes et des colonnes (il reste des objects non classés) ; des groupes ne se recouvrant pas et contenant l'ensemble des données ; des classes réparties sur toutes les lignes, resp. toutes les colonnes, mais ne se recouvrant pas et pouvant avoir des colonnes communes, resp. des lignes communes ; des classes structurées sous forme d'arbre sans chevauchement ; des groupes non recouvrants et non exclusifs (il peut rester des données non classées) ; des structures hiérarchiques et recouvrantes ; des classes se chevauchant arbitrairement. De plus, il faut distinguer trois approches de classification croisée : découvrir les groupes croisés, un seul à la fois ; découvrir un ensemble de classes croisées ; découvrir toutes les classes croisées en même temps (identification simultanée). Enfin, quatre principaux types d'algorithmes sont proposés dans la littérature. Combinaison itérative de classifications des lignes et des colonnes qui consiste à classifier celles-ci séparément, puis à utiliser une procédure itérative pour combiner les deux typologies. Diviser et conquérir : le problème est découpé en plusieurs sous-problèmes similaires sur les données originales, puis consiste à résoudre chacun d'eux de façon itérative et à combiner les solutions intermédiaires pour fournir une solution unique sur les données initiales. D'autres algorithmes sont fondés sur la recherche itérative "gourmande" (Cheng et al., 2000). La dernière approche revient à énumérer de façon exhaustive les classes croisées (SAMBA, Tanay et al., 2004). La méthode proposée repose sur une combinaison entremêlee des typologies des lignes/colonnes et des colonnes/lignes.

3 Une méthode de classification croisée fondée sur une approche dynamique

Soient $X_1, ..., X_q$, q variables numériques, avec $X_j \in \mathbb{R}^{n \times T}$, où n est le nombre d'individus et T la longueur de la série temporelle. $x_{ij(t)}$ représentera par exemple la consommation du client e_i, le jour j au pas de temps t (l'heure, la demi-heure, un point 10 minutes). Nous postulons que pour un même client, des ensembles de comportements de consommation journalière peuvent être similaires. Nous mesurons cette similarité à l'aide des corrélations entre ces variables-jours. D'autre part, pour un même jour des profils de consommations des n clients peuvent se ressembler. Ils seront mesurés à l'aide de la distance euclidienne entre courbes. La matrice X des données a la forme suivante :

$$X = \begin{pmatrix} x_{11(1)}, ..., x_{11(t)}, ..., x_{11(T)} & \cdots & x_{1j(1)}, ..., x_{1j(t)}, ..., x_{1j(T)} & \cdots & x_{1q(1)}, ..., x_{1q(t)}, ..., x_{1q(T)} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1(1)}, ..., x_{i1(t)}, ..., x_{i1(T)} & \cdots & x_{ij(1)}, ..., x_{ij(t)}, ..., x_{ij(T)} & \cdots & x_{iq(1)}, ..., x_{iq(t)}, ..., x_{iq(T)} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n1(1)}, ..., x_{n1(t)}, ..., x_{n1(T)} & \cdots & x_{nj(1)}, ..., x_{nj(t)}, ..., x_{nj(T)} & \cdots & x_{nq(1)}, ..., x_{nq(t)}, ..., x_{nq(T)} \end{pmatrix}$$

La démarche proposée contient trois étapes principales : la classification des colonnes (les variables-jours), la classification des lignes (les individus-clients) et la réconciliation des résultats des deux. Cependant, la typologie des colonnes intégrera aussi une classification des lignes afin d'enrichir sa structure et vice-versa pour la typologie des lignes comme nous le verrons dans les section 3.1 et 3.2. Détaillons ces trois étapes en utilisant le synoptique de la démarche fourni dans la figure 1.

3.1 Classification des colonnes-variables

Tout d'abord, la méthode de classification de variables (Derquenne, 2016, 2017) est appliquée sur chaque individu-client afin d'obtenir n partitions $(P_1, ..., P_i, ..., P_n)$ contenant respectivement $(q_1, ..., q_i, ..., q_n)$ groupes. Afin de lier immédiatement l'information obtenue à l'ensemble des individus, nous construisons une typologie de ceux-ci en établissant un tableau de dissimilarités entre les n partitions à l'aide du V de Cramer : $1 - d^2(P_i, P_l)/(q \times \min(q_i - 1, q_l - 1))$, où $d^2(P_i, P_l)$ est la statistique du χ^2 d'indépendance. Puis le critère d'agrégation de Ward est appliqué sur la matrice de dissimilarités grâce auquel, par coupure du dendrogramme, nous obtenons M classes $(C_1, ..., C_m, ...C_M)$ possédant respectivement $(n_1, ..., n_m, ..., n_M)$ individus. Afin de reconnecter les variables à chacune des M classes d'individus, nous calculons des dissimilarités entre variables avec la mesure suivante : $d(j, k) = 1 - u/n_m$, où u est le nombre de fois où X_j et X_k sont dans la même classe de variables pour un groupe C_m. Nous obtenons alors une nouvelle matrice de dissimilarités entre variables par classe de clients. Nous appliquons sur cette matrice le critère de Ward pour obtenir une nouvelle typologie de variables pour chacune des M partitions de clients $(G_1, ..., G_m, ..., G_M)$. Par conséquent, dans chacune de celle-ci, chaque couple individu/variable (e_i, X_j) est renseigné par un numéro de classe nommé g_{ml} issu de ces nouveaux résultats. Par exemple, pour l'individu e_i qui appartient à la partition de clients G_m, la variable X_j a été mise dans la classe V_l provenant de la typologie des demi-heures réalisée sur la partition G_m. Finalement, nous rassemblons ces résultats dans un tableau T_1 contenant $n \times q$ lignes et trois colonnes : les dénominations de l'individu e_i et de la variable X_j, et le numéro de classe g_{ml}.

3.2 Classification des lignes-individus

Cette deuxième étape est le dual de la précédente. En effet, nous appliquons tout d'abord le critère d'agrégation de Ward pour classifier les individus pour chaque variable-jour afin d'obtenir q parti-

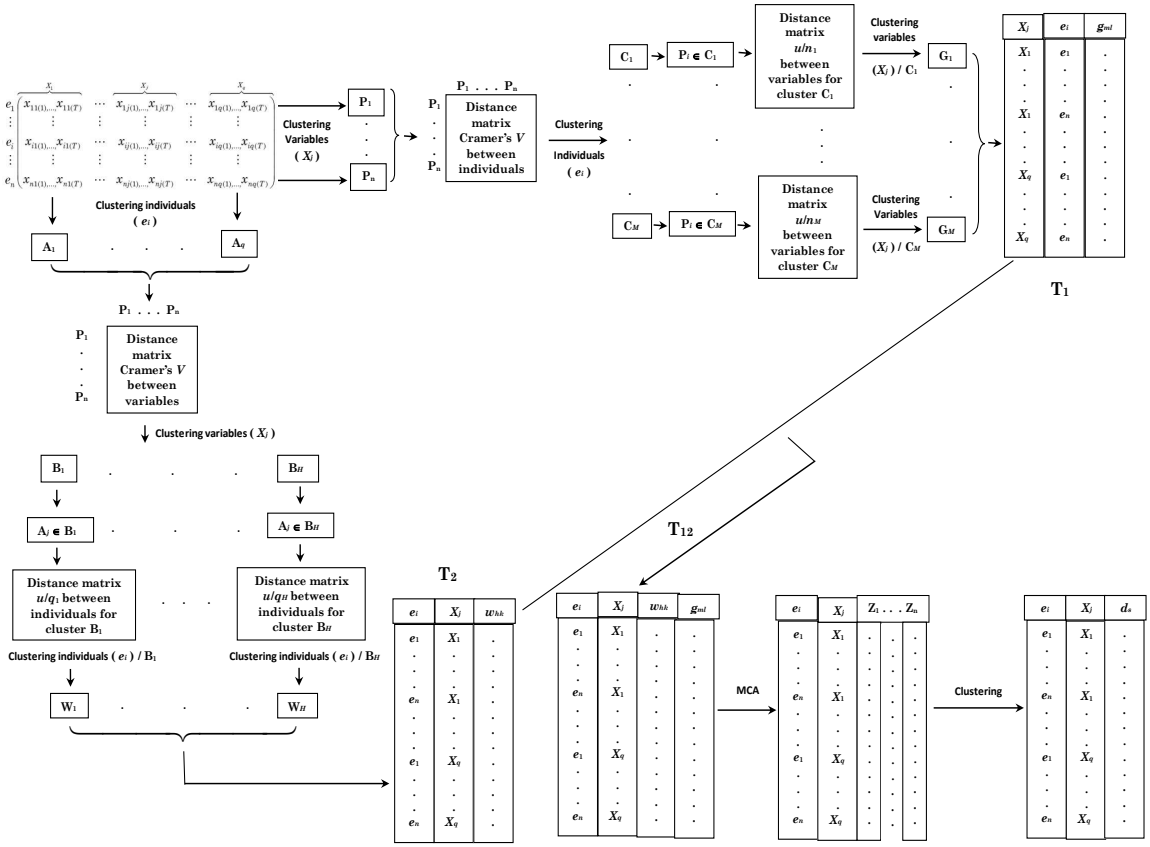


Figure 1: Synoptique du processus de la méthode

tions $(A_1, ..., A_j, ..., A_q)$ contenant respectivement $(n_1, ..., n_j, ..., n_q)$ groupes. Comme dans la première étape, nous lions les résultats issus de l'ensemble des variables, pour cela nous construisons une typologie de celles-ci en établissant un tableau de dissimilarités entre les q partitions à l'aide du V de Cramer : $1 - d^2(A_j, A_k)/(n \times \min(n_j - 1, n_k - 1))$. Le critère d'agrégation de Ward est à nouveau appliqué sur la matrice de dissimilarités grâce auquel, par coupure du dendrogramme, nous obtenons H groupes $(B_1, ..., B_h, ...B_H)$ possédant respectivement $(q_1, ..., q_h, ..., q_H)$ variables. Afin de reconnecter les individus à chacune des H classes de variables, nous calculons des dissimilarités entre individus à l'aide de la mesure suivante : $d(i, l) = 1 - u/q_h$, où u est le nombre de fois où e_i et e_l sont dans la même classe d'individus pour un groupe B_h. Nous obtenons alors une nouvelle matrice de dissimilarités entre individus par classe de variables. Nous appliquons sur cette matrice le critère de Ward pour obtenir une nouvelle typologie d'individus par partition de variables $(W_1, ..., W_h, ..., W_H)$. Par conséquent, dans chacune de celle-ci, chaque couple variable/individu (X_j, e_i) est renseigné par un numéro de classe nommé w_{hk} issu de ces nouveaux résultats. Par exemple, pour la variable X_j qui appartient à la partition de variables W_h, l'individu e_i a été placé dans la classe S_k provenant de la typologie des clients réalisée sur la partition W_h. Finalement, nous rassemblons ces résultats dans un tableau T_2 contenant $n \times q$ lignes et trois colonnes : la dénomination de l'individu e_i, celle de la variable X_j et le numéro de classe w_{hk} associé.

3.3 Réconciliation des deux typologies

L'objectif de cette dernière étape est d'obtenir une partition globale des données dans laquelle chaque couple (individu/variable) appartient à une classe croisée. Pour cela, nous fusionnons, tout d'abord les tableaux T_1 et T_2 issus des deux étapes précédentes. Ce nouveau tableau T_{12} contiendra donc quatre colonnes : la dénomination de l'individu e_i, celle de la variable X_j, le numéro de classe g_{ml} provenant de la classification des colonnes/variables et le numéro de classe w_{hk} provenant de la classification des lignes/individus. En fait, les deux premières colonnes forment simplement un individu renseigné par deux variables catégorielles correspondant aux numéros de classes g_{ml} et w_{hk}. Le tableau de contingence de ces deux variables fournit pour chaque croisement de modalités un nombre de couples (individu/variable) possédant les mêmes caractéristiques. En d'autres termes, des clients ayant le même profil de consommation d'électricité sur les mêmes jours. Afin de résumer l'ensemble de ces informations dans des classes croisées, nous appliquons une analyse des correspondances multiples (MCA) qui fournit des composantes principales associées aux observations (les couples individu/variable) et des composantes principales dédiées aux modalités des deux variables (les numéros de classes issus des deux typologies). Une classification par le critère de Ward est appliquée sur une sélection de composantes principales des observations. Celle-ci offre la typologie finale des données contenant S classes croisées $(d_1, ..., d_s, ..., d_S)$ pour les individus/variables.

4 Application de l'approche de classification croisée

Afin d'évaluer la qualité de l'approche proposée, nous avons simulé un jeu de données possédant 25 individus, 20 variables (séries temporelles de taille $T = 48$). Celui-ci est découpé en 8 classes croisées. Par exemple, la classe 1 (en vert clair sur la figure 2) contient 10 individus et quatre variables, telle que : $X_{j(t)} = X_{1(t)} + 2\epsilon_t$ pour $j = 16, 17, 18$ où $X_{1(t)} \rightsquigarrow \mathcal{N}(0, 1)$ et $t = 1, 48$.

Lors de la réconciliation des deux typologies individus/variables et variables/individus, nous avons sélectionné les trois premières composantes principales de l'ACM (75% de l'inertie totale). Nous trouvons 8 classes croisées estimées. L'évaluation de la qualité de celles-ci en regard des 8 classes croisées observées (cf. tableau de contigence, fig. 2) a été réalisée à l'aide des indices de Rand (Rand, 1971), de Jaccard et de γ, ainsi que la proportion de bien classés (bcl). Ces indices varient entre 0 et 1, plus la valeur obtenue est proche de l'unité, plus l'adéquation est bonne. Les valeurs encadrées en vert sur le tableau de contingence correspondent à des classes croisées parfaitement reconstituées ; celles en rouge sont relatives au mauvais classement. Nous obtenons : Rand=0,97 ; Jaccard=0,83 ; $\gamma = 0,89$ et bcl=0,95, ce qui offre une très bonne qualité de reconstitution sur notre exemple.

5 Apports, applications et voies futures

L'approche proposée permet de réaliser de la classification croisée sur des séries temporelles. Elle est originale dans le sens où elle mêle la recherche de classes de lignes durant la construction de groupes de colonnes et vice-versa. D'autres approches pour données non temporelles ont également été développées. L'application de cette méthode sur les données simulées présentées dans la section 4 a montré qu'elle reconstituait avec un très bon niveau de qualité la classication croisée observée. L'application sur données réelles de consommation d'électricité est en cours d'étude. D'autres applications sur données simulées et réelles ont fourni des résultats de même bonne qualité (prix spot avec commodités, demandes résiduelles pour classifier les zones géographiques, données statiques). Les améliorations et voies futures sont les suivantes : comparaison avec d'autres méthodes de classification croisée, plus de simulations afin d'évaluer l'incertitude, traitements sur des données plus complexes,

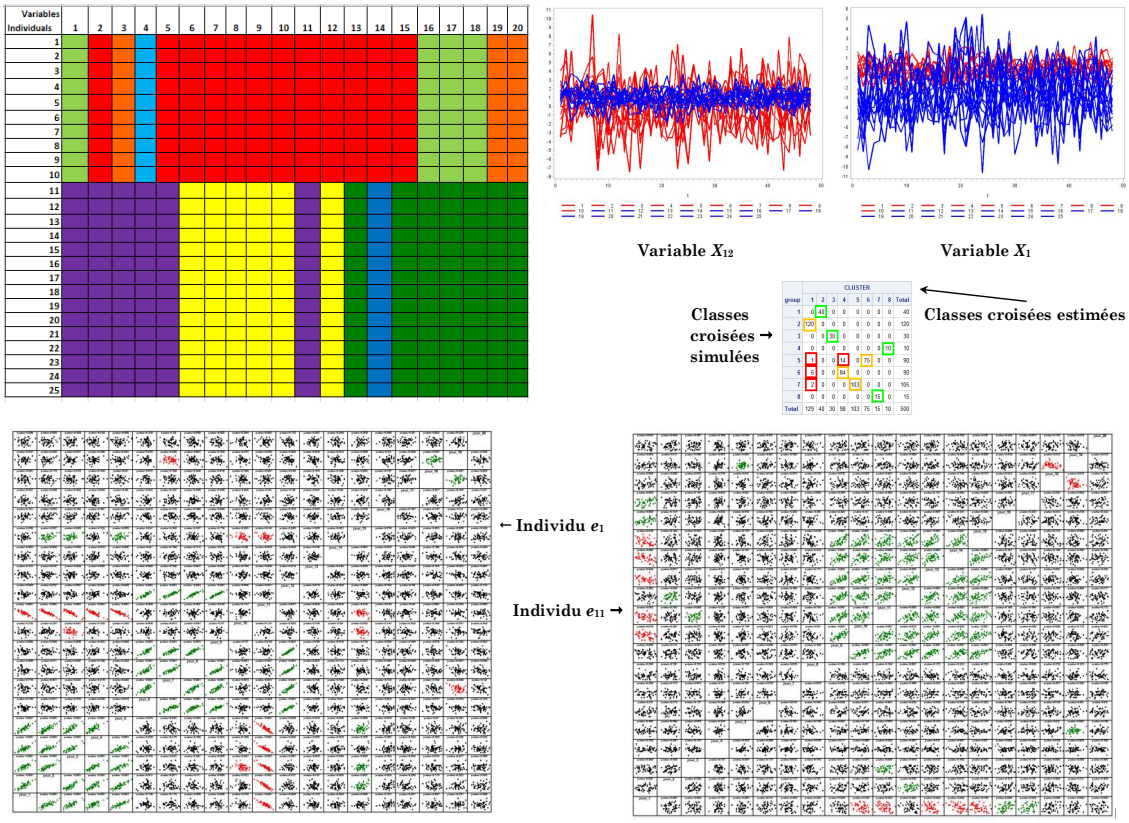


Figure 2: Visualisation des données

développement d'une méthode avec une approche fonctionnelle et extension de la méthode proposée à des données en grande dimension (Derquenne, 2018).

Bibliographie

Bouveyron C., Bozzi L., Jacques J., Jollois F-X., (2017): The functional Latent Block Model for the Co-Clustering of Electricity Consumption Curves, HAL Id: hal-01533438.
Cheng Y. and Church G., (2000): Biclustering of expression data, *Proc. ISMB'00.* AAAI Press.
Derquenne Ch., (2016): Classification de variables : une approche à double critères contrôlés dynamiques, *48ièmes Journées de Statistique*, Montpellier, France.
Derquenne Ch., (2017): Classification de variables avec des relations non linéaires, *49ièmes Journées de Statistique*, Avignon, France.
Derquenne Ch., (2018): A dynamic approach for clustering variables in high dimension, 23trd COMPSTAT, Iasi, Romania.
Govaert G. and M. Nadif (2013): Co-Clustering. Wiley-ISTE.
Madeira S., (2011): (Clustering and) Biclustering Gene Expression Data, VII International Course of Massive Data Analysis, MDA 2011.
Tanay A., Sharan R., Kupiec M. and Shamir R., (2004): SAMBA, *Proc. National Academy of Science*

Détection des change-points dans un modèle par la méthode expectile LASSO adaptative

Gabriela CIUPERCA [1]
gabriela.ciuperca@univ-lyon1.fr

Cedric DEFFO SIKOUNMO [2]
cedric.deffo-sikounmo@hdtechnology.fr

Nicolas DULAC [1 2 3]
dulac@math.univ-lyon1.fr

Résumé. Les modèles de change-points (ou à changement de régime) sont des modèles dont la forme change à des instants inconnus. Nous proposons une méthode pour les modèles de change-points linéaires qui estime la position des change-points, les coefficients de chacune des phases, et sélectionne les variables significatives simultanément grâce à une pénalité LASSO adaptative. De plus, la méthode choisie permet d'assouplir les hypothèses de départ concernant les erreurs du modèle, notamment la normalité, car elle ne requiert aucune connaissance préalable des deux premiers moments des erreurs.

Mots-clés. change-point, expectile (moindres carrés asymétriques), lasso adaptatif...

Abstract. Change-point models are models whose coefficients change at unknown times. We introduce a method for linear change-points (multiple regimes) models, which estimates the change-points location, the coefficients of each regime, and simultaneously performs feature selection thanks to an adaptive LASSO penalty. The method allows for weakened starting hypothesis, since it does not require any prior knowledge of the first two moments of the errors. Therefore it can be used when the errors don't come from a normal distribution.

Keywords. change-point, expectile (asymmetric least squares), adaptive lasso...

1 Introduction

Nous considérons ici un modèle linéaire à changement de régime, c'est-à-dire un modèle pour lequel au moins un coefficient varie à un(des) instant(s) inconnu(s). L'objectif de ce papier est d'estimer les change-points (points de rupture), ainsi que de sélectionner les variables significatives.

[1]Université de Lyon, Université Lyon 1, CNRS, UMR 5208, Institut Camille Jordan, Bat. Braconnier, 43, blvd du 11 novembre 1918, F - 69622 Villeurbanne Cedex, France
[2]HD Technology, Europarc du Chêne 8 Rue Pascal – BP 90 - 69672 BRON cedex
[3]autheur correspondant

La méthode des moindres carrés, qui modélise la moyenne conditionnelle d'un phénomène, est sans doute la plus répandue pour estimer les coefficients d'un modèle linéaire. Cependant, elle est sujette à de fortes conditions, notamment les erreurs qui sont supposées normales. Dans de nombreuses applications cette condition n'est pas vérifiée, et dès lors, modéliser la moyenne n'est plus suffisant pour expliquer la relation entre la variable dépendante et les variables indépendantes. La méthode de régression quantile de Koenker et Bassett (1978), utilisant une norme L_1 asymétrique, est une alternative qui permet d'étudier la variation des quantiles de la distribution conditionnelle en fonction des variables explicatives. Cette méthode est très utilisée en économie et en finance pour calculer des risques. Cependant, comme le soulignent Newey et Powell (1987), la méthode quantile présente quelques inconvénients : la non différentiabilité de la fonction objectif, l'inefficacité lorsque la distribution des erreurs est Gaussienne, et la difficulté à calculer la matrice de covariance des estimateurs quantiles. Newey et Powell ont donc proposé une modification du processus quantile qui utilise une norme L_2 asymétrique, et qui par conséquent possède les avantages de la méthode quantile sans les inconvénients. C'est cette méthode, appelée expectile, que nous utilisons pour estimer les coefficients du modèle.

Pour la sélection des variables nous utilisons le LASSO adaptatif de Zou (2006), qui corrige le biais des estimateurs obtenus par le LASSO de Tibshirani (1996). De plus les estimateurs obtenus par le LASSO adaptatif possèdent les propriétés oracles, c'est-à-dire que les composants des vrais paramètres qui ont pour valeur zéro sont estimé (réduits) à 0 avec une probabilité qui tend vers 1 (sparsité), et la vitesse de convergence pour les composants non nuls est optimale et asymptotiquement normale.

Nous proposons une méthode pour les modèles de change-points associant la pénalité LASSO adaptatif au processus expectile. Dans un premier temps, nous présentons succinctement les principaux résultats. Dans un second temps, nous illustrons ces résultats par des simulations.

2 Principaux résultats

2.1 Modèle et suppositions

Nous considérons un modèle avec K change-points (points de rupture), ou $(K+1)$ phases :

$$Y_i = \mathbf{X}_i^\top \boldsymbol{\beta}_1 \mathbb{1}_{1\leq i<l_1} + \cdots + \mathbf{X}_i^\top \boldsymbol{\beta}_{K+1} \mathbb{1}_{l_K\leq i\leq n} + \varepsilon_i, \qquad i = 1, \ldots, n. \tag{1}$$

Dans la suite du papier, nous supposons que le nombre K de change-points est connu et non dépendant de n, et que entre deux phases successives les paramètres sont différents : $\boldsymbol{\beta}_k \neq \boldsymbol{\beta}_{k+1}$, pour tout $k = 1, \cdots, K$. De plus nous supposons :

(A1) $\max_{1\leqslant i\leqslant n} \|\mathbf{X}_i\|_2 \leq C_0$, pour une constante $C_0 > 0$.

(A2) Il existe une matrice définie positive Ω telle que $\lim_{n\to\infty} n^{-1} \sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^\top = \Omega$.

Pour un $\tau \in (0,1)$ fixé, la fonction expetile d'ordre τ est définie par :

$$\rho_\tau(x) = |\tau - \mathbb{1}_{x<0}|x^2 \qquad \text{avec } x \in \mathbb{R}$$

Concernant les erreurs ε, nous considérons une supposition classique pour les modèles expectile, aussi considérée par Liao et al. (2019) et Ciuperca (2019) :

(A3) $(\varepsilon_i)_{1\leqslant i\leqslant n}$ i.i.d. t.q. $\mathbb{E}[\varepsilon_i^4] < \infty$ et $\mathbb{E}[g_\tau(\varepsilon)] = 0$ avec $g_\tau(x) \equiv \rho'_\tau(x-t)_{t=0} = 2\tau x \mathbb{1}_{x\geq 0} + 2(1-\tau)x\mathbb{1}_{x<0}$, la dérivée première de $\rho_\tau(x-t)$ en $t=0$.

Pour détecter automatiquement la sparsité, une pénalité de type LASSO adaptative est ajoutée au processus expectile. On considère alors le processus expectile avec pénalité LASSO adaptative :

$$\sum_{r=1}^{K+1}\left(\sum_{i=l_{r-1}^0+1}^{l_r^0}\rho_\tau(Y_i - \mathbf{X}_i^T\beta_r) + (l_r - l_{r-1})\lambda_{(l_{r-1},l_r)}\widehat{\omega}^T_{((l_{r-1},l_r),j)}|\beta_r|\right),$$

avec $\widehat{\omega}^T_{((l_{r-1},l_r),j)} \equiv |\widetilde{\beta}_{((l_{r-1},l_r),j)}|^{-\gamma}$, les poids appliqués aux coefficients dans la pénalité adaptative. Le paramètre γ est connu et $\gamma > 0$. La suite des paramètres de tuning $(\lambda_{(l_{r-1},l_r)})_{(l_{r-1},l_r)\in\mathbb{N}}$ est t.q. : $\lambda_(l_{r-1},l_r) = o((l_{r-1},l_r)^{-1/2})$ et $(l_{r-1},l_r)^{(1+\gamma)/2}\lambda_(l_{r-1},l_r) \underset{(l_{r-1},l_r)\to\infty}{\longrightarrow} \infty$ (voir Liao et al. (2019) et Ciuperca (2019)).

Pour définir les estimateurs des coefficients $(\boldsymbol{\beta}_1,\cdots,\boldsymbol{\beta}_{K+1})$ et des change-points $(l_1,\cdots,l_K)$, considérons d'abord la somme

$$S(l_1,\cdots,l_K) \equiv \inf_{(\boldsymbol{\beta}_1,\cdots,\boldsymbol{\beta}_{K+1})\in\mathbb{R}^{(K+1)p}}\left\{\sum_{r=1}^{K+1}\left(\sum_{i=l_{r-1}+1}^{l_r}\rho_\tau(Y_i - \mathbf{X}_i^\top\boldsymbol{\beta}_r) + (l_r - l_{r-1})\lambda_{(l_{r-1},l_r)}\widehat{\boldsymbol{\omega}}^\top_{((l_{r-1},l_r)}|\boldsymbol{\beta}_r|\right)\right\}.$$

Alors, nous proposons comme estimateurs pour les change-points, le K-vecteur aléatoire suivant :

$$\left(\widehat{l}_1,\cdots,\widehat{l}_K\right) \equiv \underset{(l_1,\cdots,l_K)\in\mathbb{R}^K}{\arg\min}\ S(l_1,\cdots,l_K).$$

Sur la base de ces estimateurs, nous considérons les estimateurs des coefficients du modèle (1) :

$$\left(\widehat{\boldsymbol{\beta}}_{(0,\widehat{l}_1)},\widehat{\boldsymbol{\beta}}_{(\widehat{l}_1,\widehat{l}_2)},\cdots,\widehat{\boldsymbol{\beta}}_{(\widehat{l}_K,n)}\right) \equiv$$
$$\underset{(\boldsymbol{\beta}_1,\cdots,\boldsymbol{\beta}_{K+1})\in\mathbb{R}^{(K+1)p}}{\arg\min}\left\{\sum_{r=1}^{K+1}\left(\sum_{i=\widehat{l}_{r-1}+1}^{\widehat{l}_r}\rho_\tau(Y_i - \mathbf{X}_i^\top\boldsymbol{\beta}_r) + (\widehat{l}_r - \widehat{l}_{r-1})\lambda_{(\widehat{l}_{r-1},\widehat{l}_r)}\widehat{\boldsymbol{\omega}}^\top_{((\widehat{l}_{r-1},\widehat{l}_r)}|\boldsymbol{\beta}_r|\right)\right\}.$$

Pour le modèle (1), soit $\left(\boldsymbol{\beta}_1^0,\cdots,\boldsymbol{\beta}_{K+1}^0\right)$ les vraies valeurs des coefficients dans les $(K+1)$ phases et $l_1^0,\cdots,l_K^0$ les vraies localisations des K change-points.

Pour étudier la sparsité des estimateurs expectile LASSO adaptatifs du modèle (1), on considère l'ensemble d'index pour la phase r :

$$\mathcal{A}_r^0 \equiv \big\{j \in \{1, \cdots, p\};\ \ \beta_{r,j}^0 \neq 0\big\}.$$

Pour les estimateurs des coefficients par la méthode expectile avec LASSO adaptatifs, soient :

$$\widehat{\mathcal{A}}_r^0 \equiv \big\{j \in \{1, \cdots, p\};\ \ \widehat{\beta}_{(l_{r-1}^0, l_r^0),j} \neq 0\big\}, \quad \text{et} \quad \widehat{\mathcal{A}}_r \equiv \big\{j \in \{1, \cdots, p\};\ \ \widehat{\beta}_{(\widehat{l}_{r-1}, \widehat{l}_r),j} \neq 0\big\}.$$

De plus on suppose :

(A4) $\gamma \in (0, 1]$.

(A5) $\mathbb{E}[|\varepsilon|^q] < \infty$, $\forall q \geq 2$.

(A6) $l_r - l_{r-1} \geq n^u$, avec $u \in [3/4, 1]$, pour tout $r = 1, \cdots, K+1$.

(A7) $\exists\ c > 0$ t.q. $\|\boldsymbol{\beta}_r - \boldsymbol{\beta}_{r-1}\|_2 > c$, $\forall\ r = 2, \cdots, K+1$.

(A4) permet de contrôler la taille de la pénalité par rapport à la fonction coût. (A6) a été considérée par Bai (1998) et Ciuperca (2014). (A7) est une supposition naturelle pour qu'il existe un changement après chaque change-point. Le fait que $u \geq 3/4$, permet d'obtenir des estimateurs consistants pour les coefficients de chaque phase, et donc d'obtenir des estimateurs pour les change-points pas trop loin (à une distance bornée) des vraies valeurs.

2.2 Comportement asymptotique

Ce premier théorème montre que si le nombre K de change-points est connu, alors chaque change-point possède un estimateur situé à une distance finie.

Theorem 2.1 *Si les suppositions (A1), (A3), (A4)-(A7), sont satisfaites et en plus dans chaque phase $r = 1, \cdots, K$, on a $\lim_{n\to\infty}(l_r^0 - l_{r-1}^0)^{-1} \sum_{i=l_{r-1}+1}^{n_r^0} \mathbf{X}_i \mathbf{X}_i^\top = \boldsymbol{\Omega}_r$, avec $\boldsymbol{\Omega}_r$ une matrice définie positive, la suite $(\lambda_{(l_{r-1},l_r)})$ satisfait les conditions : $\lambda_{(l_{r-1},l_r)} = o((l_r^0 - l_{r-1}^0)^{-1/2})$ et la suite $(c_{(l_{r-1},l_r)})$ est soit constante soit $c_{(l_{r-1},l_r)} \to 0$, $(l_r^0 - l_{r-1}^0)c_{(l_{r-1},l_r)}^2/(\log(l_r^0 - l_{r-1}^0)) \to \infty$ et $(l_r^0 - l_{r-1}^0)^{(\gamma-1)/2} c_{(l_{r-1},l_r)}^{-2} = O(1)$, alors $\widehat{l}_r - l_r^0 = O_{\mathbb{P}}(1) \quad \forall r = 1, \cdots, K$.*

Le théorème précédent implique le corollaire suivant, qui donne la vitesse de convergence des estimateur des coefficients de chacune des phases.

Corollary 2.1 *Sous les mêmes conditions considérées dans le Théorème 2.1, on a, $\big\|\widehat{\boldsymbol{\beta}}_{(\widehat{l}_{r-1},\widehat{l}_r)} - \boldsymbol{\beta}_r^0\big\|_2 = O_{\mathbb{P}}\big((l_r^0 - l_{r-1}^0)^{-1/2}\big)$.*

Nous montrons par le théorème suivant que, dans chaque phase estimée, les estimateurs expectile LASSO adaptatifs satisfont la propriété de sparsité, et la loi asymptotique des estimateurs des coefficients non nuls est gaussienne.

Theorem 2.2 *Sous les conditions du Theorem 2.1, si en plus* $\left(l_r^0 - l_{r-1}^0\right)^{(\gamma+1)/2} \lambda_{(l_{r-1}^0, l_r^0)} \xrightarrow[n\to\infty]{} \infty$, *alors nous avons :*
(i) $(\widehat{l}_r - \widehat{l}_{r-1})^{1/2} \left(\widehat{\boldsymbol{\beta}}_{(\widehat{l}_{r-1};\widehat{l}_r)} - \boldsymbol{\beta}_r^0\right)_{\mathcal{A}_r^0} = (l_r^0 - l_{r-1}^0)^{1/2} \left(\widehat{\boldsymbol{\beta}}_{(\widehat{l}_{r-1};\widehat{l}_r)} - \boldsymbol{\beta}_r^0\right)_{\mathcal{A}_r^0} (1 + o_{\mathbb{P}}(1)) \xrightarrow[n\to\infty]{\mathcal{L}}$
$\mathcal{N}(\boldsymbol{0}_{|\mathcal{A}_r^0|}, \sigma_{g_\tau}^2 \mu_{h_\tau}^{-2} \boldsymbol{\Omega}_{r,\mathcal{A}_r^0}^{-1})$.
(ii) Pour tout $r = 1, \cdots, K$, *nous avons* $\lim_{n\to\infty} \mathbb{P}\left[\widehat{\mathcal{A}}_r^0 = \widehat{\mathcal{A}}_r = \mathcal{A}_r^0\right] = 1$.

3 Simulations

Les simulations ont été réalisées pour une modèle avec un change-point. Nous considérons deux types d'erreurs : normales centrées $\mathcal{N}(0,1)$ et mixtes $0.2 * \mathcal{N}(0,1) + \chi_1^2$. Les paramètres sont estimés par trois méthodes différentes, moindres carrés (MC), expectile (EX), et quantile (QU), auxquelles on ajoute une pénalité LASSO adaptative pour la sélection des variables. 1000 itérations de Monte Carlo ont été réalisées. Pour étudier la sparsité, nous calculons $\hat{\mathcal{A}} \cap \mathcal{A}^0$ (les coefficients estimés non nuls dont la vraie valeur est non nulle) et $\hat{\mathcal{A}} \cap \mathcal{A}^c$ (les coefficients estimés non nuls dont la vraie valeur est nulle). Nous donnons aussi le temps d'exécution moyen pour chacune des méthodes, ainsi que $\|\frac{\hat{l}_k}{n} - \frac{l^0}{n}\|_1$, appelée L dans les tableaux. Nous donnons à titre indicatif $\|\hat{\beta} - \beta^0\|_2$ (la précision des estimateurs des coefficients) bien que cet indicateur revêt moins d'importance que la sparsité. Concernant la pénalité LASSO adaptative, nous choisissons $\lambda_{(\hat{l},\hat{k})} = (\hat{k} - \hat{l})^{-2/5}$ pour les méthodes moindres carrés et expectile, et $\lambda_{(\hat{l},\hat{k})} = (\hat{k} - \hat{l})^{2/5}$ pour la méthode quantile comme Ciuperca (2019). De même pour le paramètre γ, nous utilisons les valeurs recommandées par Ciuperca (2019), à savoir $\gamma = 1$ pour les méthodes des moindres carrés et expectile, et $\gamma = 1.225$ pour la méthode quantile. Les échantillons ont été générés à partir du modèle suivant :

$$Y_i = \mathbf{X}_i^t \beta_1^0 \mathbb{1}_{1 \leqslant i \leqslant l_1} + \mathbf{X}_i^t \beta_2^0 \mathbb{1}_{l_1 < i \leqslant n} + \varepsilon_i$$

avec $\beta_1^0 = (0, 0, -1, 0, 5, -6, 0, 3, 0, 0)$ et $\beta_2^0 = (-2, 0, 1, 4, -3, 0, 0, 0, 1, 0)$, $\mathbf{X} = (X_1, \ldots, X_10)$, $X_3 \sim \mathcal{N}(2,1)$, $X_5 \sim \mathcal{N}(4,1)$, $X_6 \sim \mathcal{N}(-2,1)$, $X_9 \sim \mathcal{N}(1,1)$, et $X_j \sim \mathcal{N}(0,1)$ pour $j \in 1, 2, 4, 7, 8, 10$. Les simulations ont été effectuées pour des tailles d'échantillon différentes, à savoir 50, 100, et 200, avec le change-point à l'observation $\lfloor \frac{n}{3} \rfloor$.

Nous déduisons des tableaux 1 et 2 que les trois méthodes utilisées donnent des résultats satisfaisants concernant l'estimation du change-point. Deux avantages principaux de la méthode proposée sur la méthode quantile se dégagent. Le premier étant la vitesse de convergence. Pour les deux types d'erreur (symétrique et asymétrique), la sparsité est satisfaite plus rapidement par la méthode expectile (EX) que par la méthode quantile

n	méthode	$\left\|\hat{\beta}-\beta^0\right\|_2$	$\sum_{r=1}^2 \hat{\mathcal{A}}_r \cap \mathcal{A}_r^0$	$\sum_{r=1}^2 \hat{\mathcal{A}}_r \cap \mathcal{A}_r^c$	L	Temps (s)
50	MC	1.19	96.9%	5.25%	0	0.00302
	EX	1.19	96.9%	5.23%	0	0.00305
	QU	1.22	97.2%	8.87%	0	0.0047
100	MC	0.694	99.8%	0.436%	0	0.00302
	EX	0.695	99.8%	0.4%	0	0.00304
	QU	0.692	99.8%	2.44%	0	0.00528
200	MC	0.476	100%	0.0545%	0	0.00315
	EX	0.476	100%	0.0455%	0	0.00313
	QU	0.463	100%	0.755%	0	0.0066

TABLE 1 – Résultats des simulations quand les erreurs sont $\varepsilon \sim \mathcal{N}(0,1)$

n	méthode	$\left\|\hat{\beta}-\beta^0\right\|_2$	$\sum_{r=1}^2 \hat{\mathcal{A}}_r \cap \mathcal{A}_r^0$	$\sum_{r=1}^2 \hat{\mathcal{A}}_r \cap \mathcal{A}_r^c$	L	Temps (s)
50	MC	1.36	96.1%	11.8%	0	0.00294
	EX	1.56	93.3%	2.11%	0	0.00327
	QU	1.4	92.9%	3.05%	2e-04	0.00515
100	MC	0.816	99.6%	4.99%	5e-05	0.00292
	EX	0.814	99.3%	0.0818%	0	0.00311
	QU	0.367	99.8%	0.127%	0	0.00532
200	MC	0.589	99.9%	1.55%	0	0.00306
	EX	0.587	100%	0%	0	0.00335
	QU	0.201	100%	0%	0	0.00724

TABLE 2 – Résultats des simulations quand les erreurs sont $\varepsilon \sim 0.2 * \mathcal{N}(0,1) + \chi_1^2$

(QU). Elle est donc plus adaptée lorsque le nombre d'observations est limité. Le deuxième avantage notable est la vitesse d'exécution. La méthode expectile est plus rapide que la méthode quantile. C'est une information à prendre en compte lors d'applications sur des données volumineuses. Il faut aussi noter que la méthode quantile peut rencontrer des problèmes de résolution numérique, ce qui n'est pas le cas de la méthode expectile.

4 Conclusion

Le processus expectile, associé à une pénalité LASSO adaptative, donne des résultats très satisfaisants pour la détection de la sparsité et l'estimation de la position des change-points dans un modèle linéaire multi-régime. Si la condition (A3) sur les erreurs est respectée, la méthode expectile fonctionne aussi bien lorsque les erreurs proviennent d'une distribution symétrique, que lorsque les erreurs proviennent d'une distribution asymétrique, contrairement à la méthode quantile. Sa rapidité d'exécution et sa facilité de résolution numérique, en font une alternative à la méthode quantile lorsqu'on n'a que très peu de connaissance concernant les erreurs du modèle.

Bibliographie

Bai, J. (1998), *Estimation of multiple-regime regressions with least absolute deviation*, Journal of Statistical Planning Inference, 74, 103-134.
Ciuperca, G. (2014), *Model selection by LASSO methods in a change-point model*, Statistical Papers, 55, 349-374.
Ciuperca, G. (2016), *Adaptive LASSO model selection in a multiphase quantile regression*, Statistics, 50 :5, 1100-1131.
Ciuperca, G. (2019), *Variable selection in high-dimensional linear model with possibly asymmetric errors*.
Koenker, R. et Bassett, G. (1978), *Regression quantiles*, Econometrica 46 (1), 33–50.
Liao, L. et al. (2019), *Penalized expectile regression : an alternative to penalized quantile regression*, Ann. Inst. Statist. Math. 71 (2), 409-438.
Newey, W. et Powell, J. (1987), *Asymmetric least squares estimation and testing*, Econometrica 55 (4), 819–847.
Tibshirani, R. (1996), *Regression shrinkage and selection via the LASSO*, J R Stat Soc B 58 :267-288.
Zou, H. (2006), *The Adaptive Lasso and Its Oracle Properties*, J. Amer. Statist. Assoc. 101 (476), 1418–1429.

PRÉDICTION DYNAMIQUE INDIVIDUELLE D'ÉVÈNEMENT DE SANTÉ À PARTIR DE MULTIPLES DONNÉES LONGITUDINALES

Anthony Devaux [⋆,1], Robin Genuer [⋆,†,2] & Cécile Proust-Lima [⋆,3]

[⋆] *INSERM U1219, 146 rue Léo Saignat, 33076 Bordeaux Cedex, FRANCE*
[†] *INRIA Bordeaux Sud-Ouest, 200 avenue de la vieille tour, 33405 Talence Cedex, FRANCE*

[1] *anthony.devaux@u-bordeaux.fr* [2] *robin.genuer@u-bordeaux.fr*
[3] *cecile.proust-lima@u-bordeaux.fr*

Résumé. Les données individuelles collectées tout au long du suivi de patients sont des informations essentielles pour évaluer la probabilité de survenue d'un évènement de santé, et in fine adapter par exemple une stratégie thérapeutique. Des méthodes statistiques, basées sur la modélisation conjointe ou l'approche landmark, ont été proposées pour incorporer les données répétées d'un ou deux marqueurs dans des outils de prédiction dynamique, qui peuvent ensuite être mis à jour à chaque nouvelle information disponible. Ces méthodologies ne sont néanmoins pas adaptées pour un grand nombre de marqueurs. Nous proposons donc une méthodologie pour la prédiction dynamique individuelle prenant possiblement en compte un très grand nombre de marqueurs répétés dans le temps. Nous combinons pour cela l'approche de landmark dynamique aux techniques d'apprentissage statistique adaptées aux données de survie. Nous illustrons l'approche dans le contexte de cirrhose biliaire primitive, et comparons les performances de plusieurs modèles à travers une étude de simulations.

Mots-clés. Prédiction, Landmark, Modèles mixtes, Apprentissage automatique, Données longitudinales, Données de survie

Abstract. The individual data collected throughout patient follow-up is crucial information for assessing the risk of a health event, and for finally adapting a therapeutic strategy, for example. Statistical methods based on joint models or landmark approach have been proposed to include repeated measures from one or two markers in dynamic prediction models, which can then be updated when new information becomes available. However, these methodologies can't handle a large number of repeated markers. We thus propose a methodology for individual dynamic prediction that may take into account a very large number of repeated markers over time. We combine the dynamic landmark approach with machine learning techniques adapted to survival data. We illustrate the approach in the context of primary biliary cirrhosis, and compare the performance of the different models through a simulation study.

Keywords. Prediction, Landmark, Mixed models, Machine learning, Longitudinal data, Survival data

1 Contexte

Prédire de façon individuelle le risque de survenue d'un évènement clinique est devenu une question centrale en santé. Au cours du suivi d'un patient, cela permet notamment d'adapter la stratégie thérapeutique ou revoir la fréquence de suivi. Avec l'informatisation des établissements de santé, de nombreuses données sont désormais collectées à chaque venue du patient et peuvent être exploitées pour fournir des prédictions individuelles les plus précises possibles et pouvant être mises à jour dynamiquement lorsque de nouvelles mesures sont disponibles.

Les méthodes statistiques permettant de fournir des prédictions dynamiques individuelles sont fondées sur les modèles conjoints et les modèles landmark (Ferrer et al., 2019) pour données répétées et temps d'événement. Cependant, les méthodes actuelles sont limitées à l'analyse d'un voire deux marqueurs répétés alors que dans de nombreux contextes, beaucoup plus de marqueurs répétés peuvent être disponibles et informatifs.

Dans ce travail, nous proposons donc une méthodologie pour la prédiction dynamique individuelle prenant possiblement en compte un très grand nombre de marqueurs répétés dans le temps. Nous combinons pour cela l'approche de landmark dynamique aux techniques d'apprentissage statistique adaptées aux données de survie. Nous illustrons l'approche dans le contexte de cirrhose biliaire primitive, et comparons les performances de plusieurs modèles candidats à travers une étude de simulations.

2 Modèle landmark dynamique

Le modèle landmark dynamique consiste à prédire le risque de survenue d'un événement à partir d'un temps de landmark noté t_{LM} en utilisant l'histoire d'information collectée jusqu'en t_{LM} pour les patients n'ayant pas encore subi l'évènement avant t_{LM}. Pour incorporer dans cette histoire d'information un grand nombre de biomarqueurs mesurés de façon intermittente avec erreur, nous adoptons une approche en deux étapes (comme résumé en figure 1) : (1) nous utilisons des modèles mixtes pour prédire des résumés de l'histoire des marqueurs ; (2) ces résumés sont inclus en variables explicatives de modèles de survie pour le temps d'événement adaptés à la grande dimension et aux variables corrélées.

2.1 Première étape : modèles mixtes

Soit Y_{ijk}, la mesure d'un marqueur $k \in \{1, \ldots, K\}$ pour un individu $i \in \{1, \ldots, N\}$ au temps t_{ijk} où $j \in \{1, \ldots, n_i\}$. A l'aide des modèles mixtes généralisés (Laird et Ware, 1982), nous modélisons $E(Y_{ijk}|b_{ik})$ par :

$$\begin{aligned} g(E(Y_{ijk}|b_{ik})) &= Y^*_{ik}(t_{ijk}) \\ &= X^\top_{ik}(t_{ijk})\beta_k + Z^\top_{ik}(t_{ijk})b_{ik} \end{aligned} \tag{1}$$

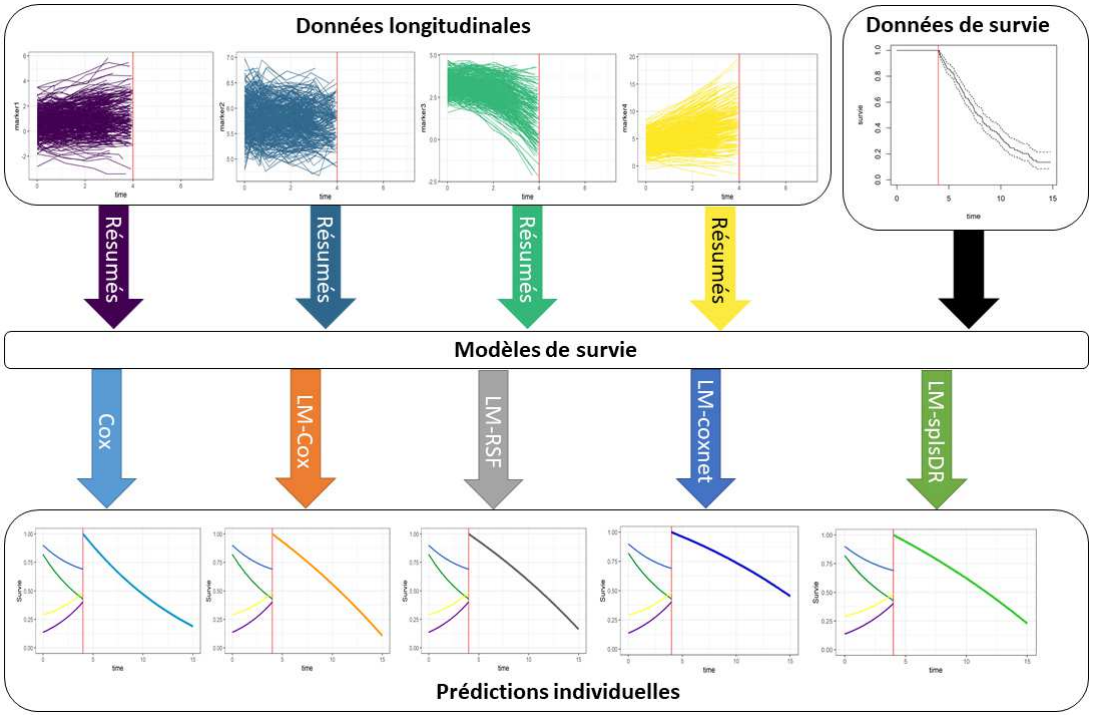


Figure 1: Principe du modèle landmark dynamique pour un grand nombre de marqueurs répétés. Le modèle final retenu (parmi Cox, LM-Cox, LM-RSF, LM-coxnet, LM-splsDR, ...) est le modèle possédant le meilleur pouvoir prédictif.

où $X_{ik}^{\top}(t_{ijk})$ est le p_k-vecteur des variables explicatives et β_k le p_k-vecteur des coefficients associées. $Z_{ik}(t_{ijk})$ est le q_k-sous-vecteur de $X_{ik}(t_{ijk})$ avec $q_k \leq p_k$ et b_{ik} le q_k-vecteur des effets aléatoires individuels associés avec $b_{ik} \sim \mathcal{N}(0, B_k)$. $g(.)$ représente la fonction de lien entre le prédicteur linéaire $Y_{ik}^*(t_{ijk})$ et $E(Y_{ijk}|b_{ik})$. La fonction de lien est choisie en fonction de la nature de Y_{ijk}, par exemple la fonction identité pour les marqueurs continues Gaussien ou la fonction logit pour les marqueurs binaires.

A partir du modèle mixte de l'équation 1, nous pouvons calculer différents résumés individuels au temps t_{LM} pour caractériser au mieux le comportement du marqueur k jusqu'en t_{LM}, comme par exemple :

- Effets aléatoires: $\widehat{b}_{ik}$; estimés comme suit :
  - Si marqueur continu : $\widehat{b}_{ik} = \widehat{B}_k Z_{ik}^{\top} \widehat{V}_{ik}^{-1}(Y_{ik} - X_{ik}\widehat{\beta}_k)$ où $\widehat{V}_{ik} = Z_{ik}\widehat{B}_k Z_{ik}^{\top} + \widehat{\sigma}_{\epsilon k} I_{n_i}$
  - Sinon : $\widehat{b}_{ik} = \underset{b_{ik}}{\text{argmax}}\ f(b_{ik}|Y_{ik}^*) = \underset{b_{ik}}{\text{argmax}}\ f(Y_{ik}^*|b_{ik}) f(b_{ik})$;
- Niveau courant sous-jacent: $\widehat{Y}_{ik}^*(t_{LM}) = X_{ik}^{\top}(t_{LM})\widehat{\beta}_k + Z_{ik}^{\top}(t_{LM})\widehat{b}_{ik}$;

- Pente courante sous-jacente : $\widehat{Y}_{ik}^{*\prime}(t_{LM}) = \frac{\partial \widehat{Y}_{ik}^{*}(t_{LM})}{\partial t_{LM}}$;
- Niveau cumulé du niveau sous-jacent: $\widehat{h}_{ik}(t_{LM}) = \int_0^{t_{LM}} \widehat{Y}_{ik}^{*}(u)du$.

Dans la suite, nous appellerons $\Gamma_i(t_{LM})$ le vecteur des résumés calculés au temps t_{LM} des K marqueurs pour l'individu i. Avec les exemples ci-dessus, le vecteur des résumés $\Gamma_i(t_{LM}) = (\widehat{Y}_i^{*}(t_{LM}), \widehat{Y}_i^{*\prime}(t_{LM}), \widehat{h}_i(t_{LM}), \widehat{b}_i)^\top$ est de taille $\sum_{k=1}^{K}(q_k + 3)$. Il a donc la particularité de posséder un très grand nombre de caractéristiques, potentiellement très corrélées entre elles. Suivant la maladie, des résumés spécifiques peuvent être définis en plus, comme par exemple le temps passé avec un marqueur au-dessus d'un seuil.

2.2 Étape deux : modèle de survie adapté à la grande dimension

Dans l'étape 2, nous cherchons à évaluer le risque d'un évènement à partir de t_{LM} en fonction des caractéristiques Γ_i et des covariables X_i. Soit $T_i^{*}(t_{LM}) = \min\left(T_i(t_{LM}), C_i(t_{LM})\right)$, avec $T_i^{*}(t_{LM})$ le temps d'évènement observé, $T_i(t_{LM})$ le temps d'évènement et $C_i(t_{LM})$ le temps de censure, tous définis à partir de t_{LM}.

2.2.1 Modèle à risques proportionnels

Tout d'abord, nous proposons d'utiliser deux modèles à risques proportionnels classiques (Cox, 1972). avec (1) les covariables ainsi que les marqueurs mesurés à l'inclusion (noté Cox) ; (2) les covariables et les résumés de marqueurs calculés au temps t_{LM} (noté LM-Cox). Une sélection automatique des variables peut être réalisée (de façon ascendante et descendante) en se basant sur le critère d'information d'Akaike.

2.2.2 Forêts aléatoires en survie

Les forêts aléatoires sont un outil de classification supervisé non paramétrique, pouvant prendre en compte un très grand nombre de données avec des relations possiblement complexes. Les forêts aléatoires, pour données de survie censurées à droite (Ishwaran et al., 2008), sont construites à partir d'un critère basé sur le test du log-rank pour découper chaque noeud de l'arbre en deux noeuds fils. De plus, le critère d'arrêt est basé sur un nombre minimum d'évènements dans chaque feuille de l'arbre.

LM-RSF est la forêt aléatoire optimisée, selon l'erreur *Out-Of-Bag* qui estime l'erreur de prédiction, à partir des paramètres *mtry* représentant le nombre de variables tirées aléatoirement à chaque noeud de l'arbre, et *nodesize* le nombre minimum d'évènement à chaque feuille. Une sélection des variables les plus importantes pourra également être réalisée.

2.2.3 Régression régularisée

LM-coxnet est une méthode de régression régularisée utilisant un modèle de Cox pénalisé par *Elastic Net* (Simon et al., 2011), combinaison des normes ℓ_1 (Lasso) et ℓ_2 (Ridge), pour réduire le nombre de variables tout en prenant en compte la possible corrélation entre elles. Deux paramètres doivent être optimisés par validation croisée selon l'*Area Under the Curve* (AUC), λ l'intensité de la pénalité et α la pondération entre les normes ℓ_1 et ℓ_2.

2.2.4 Réduction de dimension

LM-splsDR est une extension de la *sparse-Partial Least Square* (sPLS), méthode de réduction de dimension, aux données de survie (Bastien et al., 2015). Cette méthode utilise la déviance, obtenue par un modèle de Cox sans variables, comme variable réponse pour la construction des variables latentes issues des variables explicatives. De plus, une sélection des variables explicatives est faite dans chaque composante par la pénalité ℓ_1. L'optimisation du nombre de composantes *ncomp* et du paramètre de contrôle de la pénalité η est faite par l'AUC.

2.3 Étape trois : Calcul de la prédiction individuelle

Pour un nouveau sujet $\star$, les caractéristiques $\widehat{\Gamma}_\star$ sont prédites à partir des observations des marqueurs (partie 2.1). La probabilité prédite de survenue d'un évènement $\pi_\star(t_{LM}, t_{Hor})$ est calculée pour un temps landmark t_{LM} et un temps d'horizon t_{Hor} par :

$$\pi_\star(t_{LM}, t_{Hor}) = P(T_\star(t_{LM}) \leq t_{Hor} \mid \Gamma_\star(t_{LM}), X_\star) \tag{2}$$

où $X_\star$ le vecteur des covariables. Dans les modèles de Cox, LM-Cox, LM-coxnet et LM-splsDR, cette probabilité est estimée par :

$$\widehat{\pi}_\star(t_{LM}, t_{Hor}) = 1 - \exp\left(-\widehat{\Lambda}_0(t_{Hor}) \exp\left(U_\star^\top(t_{LM})\widehat{\gamma}\right)\right) \tag{3}$$

où $\widehat{\Lambda}_0(.)$ est l'estimation de la fonction de risque de base. $U_\star(t_{LM})$ est un sous-vecteur des variables explicatives ($U_\star(t_{LM}) \subseteq (\widehat{\Gamma}_\star(t_{LM}), X_\star)$) dans les modèles Cox, LM-Cox et LM-coxnet, ou des composantes créées dans le modèle LM-splsDR. $\widehat{\gamma}$ est le vecteur des coefficients estimés, potentiellement pénalisés, associés à $U_\star$.

Dans le modèle LM-RSF, la probabilité est estimée par :

$$\widehat{\pi}_\star(t_{LM}, t_{Hor}) = 1 - \exp\left(-\frac{1}{B}\sum_{b=1}^{B} \widehat{\Lambda}_\star^b(t_{Hor})\right) \tag{4}$$

où $\widehat{\Lambda}_\star^b(t_{Hor})$ est l'estimateur de Nelson-Aalen dans la feuille contenant l'individu $\star$ de l'arbre $b \in \{1, \ldots, B\}$.

3 Illustration : prédiction du décès chez les patients atteints de cirrhose biliaire primitive

Nous illustrons les méthodes proposées pour prédire la probabilité de décès chez les patients atteints de la cirrhose biliaire primitive (CBP), maladie chronique du foie. Pour cela, nous utilisons le jeu de données public *pbc2* (Murtaugh et al., 1994), où 312 patients atteints de CBP ont été recrutés lors d'un essai thérapeutique à la Mayo Clinic aux États-Unis entre 1974 et 1984. Au cours du suivi, le décès et les données individuelles ont été recueillies parmi les données démographiques et biologiques, en particulier un ensemble de marqueurs spécifiques à la maladie tels que la bilirubine, le cholestérol ou encore l'albumine.

4 Simulations

A travers une étude de simulations, nous montrons l'utilité des méthodes pour la grande dimension, par rapport à des méthodes plus classiques, pour la prédiction dynamique individuelle en présence de plusieurs marqueurs répétés. En particulier, nous observons les limites des méthodes proposées en fonction de scénarios se différenciant par le nombre de résumés associés à l'évènement, ou la forme de cette relation. L'évaluation des méthodes est réalisée au travers de l'erreur de prédiction, obtenue par l'écart moyen entre la probabilité théorique et celle prédite par chaque modèle.

Bibliographie

Bastien, P., Bertrand, P., Meyer, N. et Maumy-Bertrand, M. (2015). Deviance residuals-based sparse PLS and sparse kernel PLS regression for censored data, *Bioinformatics*, 31(3), 397–404.

Cox, D. R. (1972). Regression models and life-tables, *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2), 187–202.

Ferrer, L., Putter, H., Proust-Lima, C. (2019). Individual dynamic predictions using landmarking and joint modelling: Validation of estimators and robustness assessment, *Statistical Methods in Medical Research*, 28(12), 3649–3666

Ishwaran, H., Kogalur U. B., Blackstone, E. H. et Lauer, M. S. (2008). Random survival forests, *The Annals of Applied Statistics*, 2(3), 841–860.

Laird, N. M. et Ware, J. H. (1982). Random-Effects Models for Longitudinal Data, *Biometrics*, 38(4), 963–974.

Murtaugh, P. A., Dickson, E. R., Van Dam, G. M., Malinchoc, M., Grambsch, P. M., Langworthy, A. L. et Gips, C. G. (1994). Primary biliary cirrhosis: Prediction of short-term survival based on repeated patient visits, *Hepatology*, 20(1), 126–134.

Simon, N., Fiedman, J., Hastie, T. et Tibshirani, R. (2011). Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent, *Journal of Statistical Software*, 39(5), 1–13.

Algorithme d'ensembles actifs par fenetre glissante pour l'estimation parcimonieuse de modèle convolutionel

Laurent Dragoni [1] & Karim Lounici [2] & Rémi Flamary [3] & Patricia Reynaud-Bouret [1]

[1] *Université Côte d'Azur, CNRS, Laboratoire J.A. Dieudonné, 06108 Nice, France*
[2] *Ecole Polytechnique, Centre de Mathématiques Appliquées, 91128 Palaiseau*
[3] *Université Côte d'Azur, CNRS, Laboratoire Lagrange, OCA, 06108 Nice, France*

Résumé. Nous présentons un algorithme de résolution rapide du Lasso dans le cadre de modèles convolutionnels en grande dimension. Des simulations numériques illustrent l'efficacité de notre approche. De plus, nous démontrons théoriquement que la complexité temporelle de cet algorithme croît linéairement avec la taille du signal enregistré.

Mots-clés. Parcimonie, Lasso, Optimisation, Neurosciences, Tri de potentiels d'action

Abstract. We present a fast algorithm for the resolution of the Lasso for convolutional models in high dimension. Numerical simulations illustrate the efficiency of our approach. Moreover we show theoretically that the temporal complexity of this algorithm growths linearly w.r.t the size of the recorded signal.

Keywords. Sparsity, Lasso, Optimization, Neurosciences, Spike sorting

1 Introduction

Nous proposons un nouvel algorithme de résolution du Lasso pour l'étude de modèles convolutionnels. Sous une hypothèse de parcimonie du vecteur à estimer, nous affinons la stratégie dite d'ensembles actifs ou *active set* en un algorithme en ligne performant en grande dimension. Nous montrons de plus que la complexité temporelle théorique de cet algorithme croît linéairement avec la taille du signal enregistré. Cet algorithme générique peut s'appliquer à divers domaines, comme notamment au problème du tri de potentiels d'action en neurosciences dont la problématique porte sur l'estimation des formes des potentiels d'action et des instants d'activation des différents neurones à partir de l'enregistrement de l'activité neuronale.

Modèle convolutionnel Durant une expérience, d électrodes enregistrent l'activité de q neurones. Chaque électrode enregistre un signal de taille n (nombre de pas de temps). Nous proposons de de mettre en relation l'activité des neurones et les signaux enregistrés

en utilisant un modèle convolutionnel, introduit par Ekanadham et al. (2011). Ce modèle s'écrit sous la forme

$$\boldsymbol{S} = \sum_{j=1}^{q} \boldsymbol{W}_j * \boldsymbol{a}_j + \boldsymbol{N}, \tag{1}$$

où $\boldsymbol{S} \in \mathbb{R}^{d\times n}$ est la matrice des observations, contenant les d signaux enregistrés de taille n. La matrice $\boldsymbol{W}_j \in \mathbb{R}^{d\times \ell}$ contient les formes des potentiels d'action du neurone j sur toutes les électrodes. Notons que toute forme de potentiel d'action est décrite par ℓ points. Le vecteur $\boldsymbol{a}_j \in \mathbb{R}^n$ est appelé le vecteur d'activation du neurone j. Les entrées non nulles de $\boldsymbol{a}_j$ correspondent aux instants d'activation de ce neurone. Étant donné que la fréquence de décharge des neurones est très petite devant la fréquence d'acquisition du signal, notons que $\boldsymbol{a}_j$ est un vecteur sparse. L'opérateur de convolution par rapport au temps est noté $*$ (d'où le nom de modèle *convolutionnel*). Finalement, $\boldsymbol{N} \in \mathbb{R}^{d\times n}$ est une matrice de bruit.

La convolution étant un opérateur linéaire, le modèle convolutionnel (1) se reformule en un modèle linéaire, après une étape de vectorisation. On obtient alors le problème suivant

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{a} + \boldsymbol{\sigma}, \tag{2}$$

où $\boldsymbol{H}$ est une matrice bloc-Toeplitz de taille $dn \times qn$ codant la convolution entre les formes des potentiels d'action et le vecteur d'activation $\mathbf{a}$. Les potentiels d'action étant de longueur ℓ très petite devant n, la matrice $\boldsymbol{H}$ est en pratique extrêmement sparse.

Lasso et conditions d'optimalité Estimer $\boldsymbol{a}$ quand le nombre de neurones est supérieur au nombre d'électrodes serait impossible sans hypothèse structurelle supplémentaire. Comme annoncé en section 1, une propriété importe du problème est que les neurones ont tendance à peu décharger. Par conséquent, le nombre de coefficients non nuls dans $\boldsymbol{a}$ est petit devant qn. Ceci nous invite à considérer un estimateur promouvant la parcimonie, comme le Lasso proposé par Tibshirani (1996).

$$\hat{\boldsymbol{a}} = \underset{\boldsymbol{a}\in\mathbb{R}^{qn}}{\operatorname{argmin}} \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{a}\|_2^2 + \lambda \|\boldsymbol{a}\|_1 , \tag{3}$$

où $\lambda > 0$ est l'unique paramètre de la méthode et qui dépend du ratio signal sur bruit. Une propriété cruciale du Lasso est la condition d'optimalité suivante : soit $\hat{a}$ une solution de (3) et écrivons $\boldsymbol{H}_j$ la j-ème colonne de $\boldsymbol{H}$, où $1 \leq j \leq qn$.

$$\forall j, \quad \text{si} \quad |\boldsymbol{H}_j^T(\boldsymbol{y} - \boldsymbol{H}\hat{\boldsymbol{a}})| < \lambda, \quad \text{alors} \quad \hat{\boldsymbol{a}}_j = 0. \tag{4}$$

Algorithme d'ensembles actifs (active set) Le calcul d'un estimateur Lasso, c'est-à-dire d'une solution du problème (3), peut être très coûteux en grande dimension. Nous exploitons alors la stratégie dite de l'*active set* nous permettant de calculer de manière plus efficace le Lasso. Notons que cette stratégie assez générique a déjà été exploitée

dans d'autres contextes, cf. Lee et al. (2007), Szafranski et al. (2008) et Boisbunon et al. (2014). Ici l'idée principale de l'*active set* repose sur l'exploitation des conditions d'optimalité (4). Initialisant l'estimateur Lasso au vecteur nul, l'objectif est d'activer de manière itérative ses coordonnées j ne vérifiant pas (4), tout en mettant à jour l'estimateur Lasso à chaque activation de coordonnées. On appelle active set et on note J l'ensemble de ces coordonnées actives. Le vecteur $\boldsymbol{a}$ étant très sparse, on s'attend ainsi à résoudre des problèmes Lasso de taille $|J|$ très petite devant qn. Dans le pire des cas, l'algorithme de l'active set active toute les coordonnées possibles, ce qui conduit à calculer une solution Lasso sur l'espace tout entier. Ainsi, l'algorithme de l'active set termine toujours en temps fini.

2 Active set par fenêtre glissante adaptative

En pratique, le calcul des conditions d'optimalité et la mise à jour du Lasso sur J sont les étapes les plus coûteuses de l'active set. Au vu des dimensions du problème, ceci ne permet pas encore de calculer rapidement une solution en temps raisonnable. Nous proposons alors l'idée de l'active set par fenêtre glissante, qui exploitera davantage la structure du problème. Une notion cruciale est celle de recouvrement temporel entre activations : les activations étant rares et leurs effets sur le signal étant très localisés, deux activations suffisament distantes n'ont aucune influence réciproque. Plus précisément, rassemblons les temps d'activation de tous les neurones dans un vecteur noté $\boldsymbol{a}^{times}$ et considérons alors $\boldsymbol{a}_i^{times}$ et $\boldsymbol{a}_j^{times}$ deux activations successives de a^{times}. Si on a $|\boldsymbol{a}_i^{times} - \boldsymbol{a}_j^{times}| \leq \ell$, on dira que ces deux activations se recouvrent. On appelle alors recouvrement de l'activation $\boldsymbol{a}_i^{times}$ sa composante connexe pour la relation qui précède. On partionne ainsi l'ensemble des temps d'activations en des recouvrements disjoints. L'idée essentielle de l'active set par fenêtre glissante est de retrouver ces recouvrements en parcourant le domaine temporel à l'aide de fenêtres dont la taille peut évoluer et donc d'exploiter cette séparation des activations afin de résoudre des problèmes indépendants et de taille plus petite.

Algorithm 1 Structure de l'active set par fenêtre glissante

```
Initialisation de la fenetre ω = [[i, j]] = [[1, η]]
repeat
    a_ω = solution du Lasso sur ω via l'active set
    if Le signal reconstruit est contenu dans ω then
        Stockage de la solution a_ω, décalage de la fenêtre ω = [[j + 1, j + η]]
    else
        Extension de la fenêtre actuelle ω = [[i, j + ℓ]]
    end if
until j = n // Fin du signal
```

Retrouver les recouvrements des activations exige une évolution précise des fenêtres de

l'algorithme. Une fois les conditions d'optimalité satisfaites sur ω (ligne 3 de algorithm 1), si le signal reconstruit est entièrement contenu dans ω, autrement dit si tous les temps d'activation sont à distance supérieure à ℓ du bord j, alors le recouvrement est entièrement contenu dans ω. Or on sait que la résolution du Lasso pour le recouvrement suivant est indépendante du recouvrement actuel. Dans ce cas, on a donc trouvé la solution Lasso pour la fenetre actuelle peut donc travailler sur une nouvelle fenetre immédiatement après. Sinon, on étend la fenêtre actuelle et on résout à nouveau le Lasso sur celle-ci en mettant à jours de manière itérative avec l'*active set*. Le vecteur précédemment calculé complété par des zéros constitue alors une initialisation raisonnable pour la résolution du nouveau Lasso.

Dans l'*active set* standard, le coût du calcul des conditions d'optimalité est de l'ordre de $O(nqd\ell)$ à chaque étape. Dans l'*active set* par fenêtre glissante, ce coût se réduit donc à $O(|\omega|qd\ell)$, avec $|\omega| << n$.

3 Expérimentations numériques

Afin d'illustrer les performances de l'algorithme de l'active set par fenêtre glissante, nous présentons ici une comparaison des temps d'exécution pour trois approches différentes : l'approche frontale qui consiste à résoudre le Lasso (3) globalement, l'active set générique et l'active set par fenêtre glissante. Quelle que soit l'approche, nous résolvons les Lasso en utilisant l'algorithme du gradient proximal accéléré, implémenté dans FISTA par Beck et Teboulle (2009). Nous avons simulé notre jeu de données de manière réaliste en utilisant notamment le modèle classique de Hodgkin et Huxley (1952) pour la description des formes de potentiels d'action et implémenté par Pouzat (2016). Afin de favoriser l'étude l'influence de n, nous nous sommes limité à des valeurs raisonnables pour le nombre de neurones ($q = 5$) et d'électrodes ($d = 4$).

FISTA global et l'active set générique nécessitent une grande utilisation de la mémoire pour le stockage de la matrice $\boldsymbol{H}$ dont la taille augmente en $O(n^2)$. Ainsi les simulations pour ces deux méthodes deviennent rapidement prohibitives. Nous constatons néanmoins en figure 1 que l'active set par fenêtre glissante est visiblement plus rapide que ces deux méthodes. En effet, celui-ci semble croître linéairement en n. De plus, il nécessite une occupation de la mémoire nettement plus raisonnable.

4 Étude mathématique de l'algorithme et complexité

Étant donné la structure de l'algorithme par fenêtre glissante, une question naturelle apparaît : peut-on s'assurer que la solution $\hat{\boldsymbol{a}}$ calculée est bien une solution du problème initial ? Par construction, $\hat{\boldsymbol{a}}$ est obtenu par recollements successifs de solutions sur des fenêtres disjointes. Nous pouvons répondre par l'affirmative.

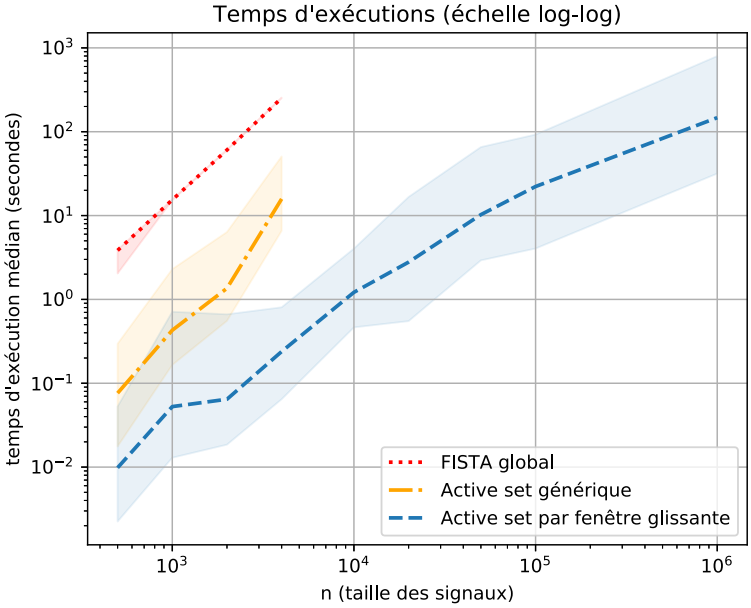


Figure 1: Temps d'exécution des algorithmes pour différentes valeurs de n.

Theorem 1. *La solution* $\hat{\mathbf{a}}$ *calculée par l'active set par fenêtre glissante est une solution Lasso du problème initial, c'est-à-dire de* (3).

Dans la suite, on s'intéresse à l'estimation de la complexité algorithmique de l'active set par fenêtre glissante. On appelle fenêtre *maximale* toute fenêtre qui a été quittée à l'issue de l'étape (5) de l'algorithme 1.

Lemma 1. *On note respectivement* I^* *et* $\hat{I}$ *l'ensemble des coordonnées non nulles des vecteurs* a *et* $\hat{a}$*. Alors sur l'événement* $I^* = \hat{I}$*, toute fenêtre maximale contient au plus un recouvrement. De plus, si une fenêtre maximale contient un recouvrement de taille* k*, alors cette fenêtre est de longueur au plus* $k + \eta + \ell$*.*

Utilisant des résultats théoriques sur le Lasso, comme Bunea (2008), nous pouvons montrer que la condition $I^* = \hat{I}$ est satisfaite avec grande probabilité. Le lemme précédent nous informe donc bien sur le lien entre la complexité de l'algorithme (la taille des fenêtres) et la taille des recouvrements. Pour des raisons techniques, nous étendons ici la distance de détection des recouvrements à 3ℓ.

Lemma 2. *Supposons que le vecteur des temps d'activation* $\boldsymbol{a}^{times}$ *soit tiré selon un processus de Bernoulli de probabilité* p*. Alors la taille moyenne d'un recouvrement est inférieure à* $3\ell(1-p)^{-3\ell}$*.*

Les résultats obtenues en section 3 semblaient indiquer que la complexité temporelle de l'active set par fenêtre glissante croît en $O(n)$. Nous sommes ici en mesure de le démontrer mathématiquement.

Theorem 2. *La complexité temporelle moyenne de l'algorithme de l'active set par fenêtre glissante est de l'ordre*

$$O(n\bar{\omega}^4 q^2 (d+q)^2), \tag{5}$$

où $\bar{\omega}$ est la taille moyenne d'une fenêtre maximale.

Les lemmes précédents nous assurent alors que $\bar{\omega}$ est indépendant de n. Ceci est à notre connaissance le premier résultat théorique sur la **résolution d'un Lasso structuré avec une complexité** $O(n)$. Notons que le résultat du théorème 2 est pessimiste. Dans des travaux futurs, nous nous attacherons à prendre en compte la dimension spatiale du problème, ce qui permettrait de résoudre des sous-problèmes associés à de petits sous-ensembles de neurones. Auquel cas les termes d et q apparaissant dans (5) seraient remplacés par des valeurs bien inférieures.

Bibliographie

Beck, A., & Teboulle, M. (2009). *A fast iterative shrinkage-thresholding algorithm for linear inverse problems.* SIAM journal on imaging sciences, 2(1), 183-202.

Boisbunon, A., Flamary, R., Rakotomamonjy, A., Giros, A., & Zerubia, J. (2014, September). *Large Scale Sparse Optimization for Object Detection in High Resolution Images.*

Bunea, F. (2008). *Consistent selection via the Lasso for high dimensional approximating regression models.* In Pushing the limits of contemporary statistics: contributions in honor of Jayanta K. Ghosh (pp. 122-137). Institute of Mathematical Statistics.

Ekanadham, C., Tranchina, D., & Simoncelli, E. P. (2011). *Recovery of sparse translation-invariant signals with continuous basis pursuit.* IEEE transactions on signal processing, 59(10), 4735-4744.

Hodgkin, A. L., & Huxley, A. F. (1952). *A quantitative description of membrane current and its application to conduction and excitation in nerve.* The Journal of physiology, 117(4), 500-544.

Lee, H., Battle, A., Raina, R., & Ng, A. Y. (2007). *Efficient sparse coding algorithms.* In Advances in neural information processing systems (pp. 801-808).

Pouzat, C. (2016). *Origin of the high frequency extra-cellular signal.* http://christophe-pouzat.github.io/LASCON2016/OriginOfTheHighFrequencyExtraCellularSignal.html.

Szafranski, M., Grandvalet, Y., & Morizet-Mahoudeaux, P. (2008). *Hierarchical penalization.* In Advances in neural information processing systems (pp. 1457-1464).

Tibshirani, R. (1996). *Regression shrinkage and selection via the lasso.* Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267-288.

Modèle de détection d'anomalies pour données longitudinales : application aux arrêts maladie

Tom DUCHEMIN [1], Angela NOUFAILLY[2], Mounia N. HOCINE [3]

[1] *Laboratoire MESuRS, Le Cnam, 292 Rue Saint-Martin, 75003 Paris; Malakoff Humanis, 21 Rue Laffitte, 75009 Paris, tom.duchemin@cnam.fr*
[2] *Warwick Medical School, A.Noufaily@warwick.ac.uk*
[3] *Laboratoire MESuRS, Le Cnam, 292 Rue Saint-Martin, 75003 Paris, mounia.hocine@cnam.fr*

Résumé. Le lieu de travail est un important lieu de diffusion des maladies infectieuses mais aussi des troubles liés aux conditions de travail comme le burnout ou les troubles musculo-squelettiques. Il serait ainsi bon de détecter quand les salariés sont trop fortement exposés à ces troubles afin de proposer aux entreprises les plans d'action adaptés. Un moyen possible est de détecter les prévalences trop élevées d'arrêt maladie, qui est une conséquence de ces nombreux troubles. C'est dans ce cadre que nous proposons un modèle statistique de détection d'anomalies pour données longitudinales. Le modèle présenté est une adaptation de l'algorithme de Farrington, un modèle basé sur une régression de Quasi-Poisson s'ajustant sur la saisonnalité et sur les alertes passées. Notre adaptation intègre au modèle de nouvelles covariables et surtout un effet aléatoire pour s'ajuster aux données longitudinales d'arrêts maladie. Le modèle est validé par un ensemble de 32 simulations et est testé sur des données d'arrêts maladie regroupant environ 1700 entreprises suivies entre 2010 et 2017.

Mots-clés. détection d'anomalies, arrêts maladie, modèle mixte, modèle linéaire généralisé, surveillance, Quasi-Poisson.

Abstract. The workplace is a major vector for the spread of infectious diseases but also for disorders related to working conditions such as burnout or musculoskeletal disorders. It would therefore be useful to detect when employees are too highly exposed to these disorders in order to propose appropriate action plans to companies. One possible way is to detect excessively high prevalences of sick leave, which is a consequence of these disorders. To meet this objective, we propose a statistical model for detecting anomalies for longitudinal data adapted to sick leave data. The model presented is an adaptation of Farrington algorithm, a model based on Quasi-Poisson regression adjusting for seasonality and past alerts. Our adaptation integrates new covariates into the model and also a random effect to fit longitudinal sick leave data. The model is validated by a set of 32 simulations and is tested on sick leave data from approximately 1700 companies monitored between 2010 and 2017.

Keywords. outbreak detection, sick leave, mixed model, generalized linear model, surveillance, Quasi-Poisson.

1 Modèle de détection d'anomalies pour données longitudinales : application aux arrêts maladie

Le lieu de travail est un important lieu de diffusion des maladies infectieuses comme la grippe. Il s'agit aussi d'un lieu où se développent des troubles liés aux conditions de travail et à l'environnement comme le burnout ou les troubles musculo-squelettiques. Ces différents phénomènes peuvent avoir pour conséquence une forte prévalence d'arrêt maladie (Labriola *et al.* 2006 [1], O'Reilly *et al.* 2002 [2]): la détection de ces "épidémies" d'absence pourraient ainsi permettre de détecter à temps les entreprises à risque pour prévoir des plans d'action afin d'améliorer le bien-être des salariés et la performance des entreprises.

C'est dans ce cadre que nous proposons un modèle statistique de détection d'anomalie pour données longitudinales. De nombreux modèles de détection ont été développés, notamment dans le domaine des maladies infectieuses ou de l'épidémiologie, les domaines qui se rapprochent le plus de nos données d'étude (Unkel *et al.* 2012 [3]). Nous proposons ici une adaptation de l'algorithme de Farrington, un algorithme utilisé en routine dans plusieurs pays, notamment au Royaume-Uni, pour la détection d'épidémie de maladies infectieuses. Cet algorithme repose sur une régression de Quasi-Poisson s'ajustant sur la saisonnalité et la tendance, et prenant en compte les alertes du passé en les sous-pondérant lors de l'estimation des paramètres. Nous proposons d'intégrer un effet aléatoire afin de s'ajuster à la dimension longitudinale des données ainsi que d'intégrer des covariables pour capter les déterminants des arrêts maladies exogènes aux phénomènes infectieux et aux conditions de travail. Le modèle est validé par un ensemble de simulation puis est appliqué à des données d'arrêts maladie.

1.1 Méthode

L'algorithme de Farrington est un algorithme utilisé en routine au Royaume Uni par la *Health Protection Agency* pour la surveillance des données épidémiologiques. L'algorithme est présenté en détails dans plusieurs articles : (Noufaily *et al.* 2013 [4], Farrington *et al.* 1996 [5]). Nous présentons ici les concepts principaux qui sont utilisés dans notre adaptation.

L'implémentation du modèle fonctionne en quatre étapes :

1. on ajuste un modèle de Quasi-Poisson sur une tendance et sur les variations saisonnières construites sous forme de facteurs.Le modèle log-linéaire de l'algorithme est donc:

$$log(\mu_i) = \beta_0 + \eta t + \delta_{j(t_i)}$$

avec μ_i l'espérance de la variable de comptage Y_i au temps t_i, β_0 l'intercept, η le paramètre associé à la tendance et $j(t_i)$ le niveau du facteur saisonnier correspondant à la semaine t_i.

2. ensuite, on pondère les observations pour donner moins de poids aux alertes du passé. Les poids sont définis comme suit :

$$w_i = \begin{cases} \gamma s_i^{-2}, & \text{if } s_i > S. \\ \gamma, & \text{otherwise.} \end{cases}$$

où γ est une constante telle que $\sum_{i=1}^{n} w_i = n$ avec n le nombre de sites et les s_i sont les résidus standardisés d'Anscombe. S est une constante qui est définit afin de limiter le taux de Faux Positifs.

3. La troisième étape est l'entraînement du dernier modèle de Quasi-Poisson repondéré avec une moyenne μ_i et une variance $\phi\mu_i$ avec ϕ le paramètre de dispersion défini tel que:

$$\hat{\phi} = max\left\{1, \frac{1}{n-2}\sum_{i=1}^{1} w_i \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}\right\}$$

4. La quatrième étape est le calcul de la borne supérieure de l'intervalle de prédiction de notre variable de comptage. La borne supérieure de l'algorithme de Farrington est une approximation du quantile $100(1-\alpha)\%$ pour Y et est défini comme suit :

$$U = \hat{\mu}_0 \left\{1 + \frac{2}{3} z_\alpha \hat{\mu}_0^{-1} (\hat{\phi}\hat{\mu}_0 + var(\hat{\mu}_0))^{1/2}\right\} \quad (1)$$

Une transformation à la puissance 2/3 est utilisée pour accentuer la symétrie du modèle et avoir des résultats plus gaussiens.

Notre adaptation du modèle commence tout d'abord par une nouvelle équation de régression :

$$\forall i \in 1, 2, ..., n \text{ and } \forall t \in 1, 2, ..., T, log(\mu_{it}) = \beta_0 + \eta t + \delta_{j(t_i)} + \mathbf{X_{it}}\beta_{\mathbf{X}} + u_i \quad (2)$$

avec n le nombre de sites, T le nombre de semaines, $\mathbf{X_{it}}$ un vecteur de covariables et $u_i \sim N(0, \sigma^2)$ un effet aléatoire.

L'intégration de nouvelles covariables n'a pas d'impact majeur sur la définition du modèle ; en revanche, l'intégration d'effet aléatoire peut poser problème puisqu'elle introduit une nouvelle source d'incertitude dans notre intervalle de prédiction. Comme la distribution de la prédiction est plus complexe, la définition d'une forme close pour l'intervalle de prédiction est impossible et une autre solution est nécessaire. L'estimation de l'intervalle de prédiction se fera ainsi par simulation par l'algorithme de Metropolis-Hastings.

Toutes les analyses ont été effectuées sur R, notamment grâce au package *GLMMadaptive* qui a été adapté aux besoins de l'étude.

1.2 Simulation

Afin de valider le modèle, un ensemble de 32 scénarii est testé sur des données simulées comme suit.

Une variable de comptage hebdomadaire est simulée selon la loi binomiale négative suivante:

$$Y_{i,t} \sim NB(\mu_{i,t}, \theta)$$

$Y_{i,t}$ est le compte hebdomadaire (d'arrêt maladie par exemple) sur le site $i > 0$ pendant la semaine $t > 0$. $\mu_{i,t}$ est l'espérance de la loi Binomiale négative et θ est le paramètre de surdispersion du modèle. $\mu_{i,t}$ est défini par les variables incluant la tendance, la saisonnalité définie par des termes de Fourier et par deux covariables X et Z ainsi que par un effet aléatoire $u_i \sim \mathcal{N}(0, \sigma^2)$ comme suit

$$\mu_{i,t} = exp\left\{\beta_0 + \beta_X X_{i,t} + \beta_Z Z_{i,t} + \eta t + \sum_{s=1}^{2}\left\{\gamma cos\left(\frac{2\pi st}{52}\right) + sin\left(\frac{2\pi st}{52}\right)\right\} + u_i\right\}$$

X et Z sont simulés selon les lois suivantes :

$$\begin{aligned} X_{i,t} \sim & \mathcal{N}(m_i, 1) \text{ avec } m_1 \sim \mathcal{U}(30, 50), \\ Z_{i,t} \sim & Bernoulli(p_i) \text{ avec } p_i \sim \mathcal{U}(0, 1). \end{aligned}$$

En pratique, nous pouvons faire face à différentes structures de données et nous avons donc choisi 32 scenarii possibles prenant en compte les différentes valeurs possibles pour 4 paramètres : la tendance et des covariables (données par β_X, β_X et η), le volume de base (donné par β_0), l'effet aléatoire (donné par σ^2) et la dispersion (donné par θ). Pour chaque simulation, 10 réplications de 50 entreprises suivies 6 ans sont effectuées: les dernières 52 semaines sont utilisés pour évaluer le modèle et les 270 premières sont utilisées pour l'entraîner.

Pour évaluer le modèle, de fausses alertes sont ajoutées aux données et la capacité du modèle à les détecter est évaluée par deux critères : le taux de faux positifs et la probabilité de détection.

1.3 Application

Une application à des données d'entreprise sera proposée. Le modèle sera entraînée sur les données d'arrêt maladie de 1785 entreprises de plus de 50 employés suivies au moins 6 ans entre 2010 et 2017. Ces données proviennent des données administratives des entreprises ayant un contrat de prévoyance avec l'institut de prévoyance Malakoff Médéric pendant cette période (maintenant appelé Malakoff Humanis). Le modèle sera ajusté sur la structure démographique des entreprises : sexe des salariés, catégorie socio-professionnelle, âge des salariés, proportion de CDD/CDI, *etc.* Pour illustration, un exemple de série temporelle pour une entreprise est présentée en Figure 1.

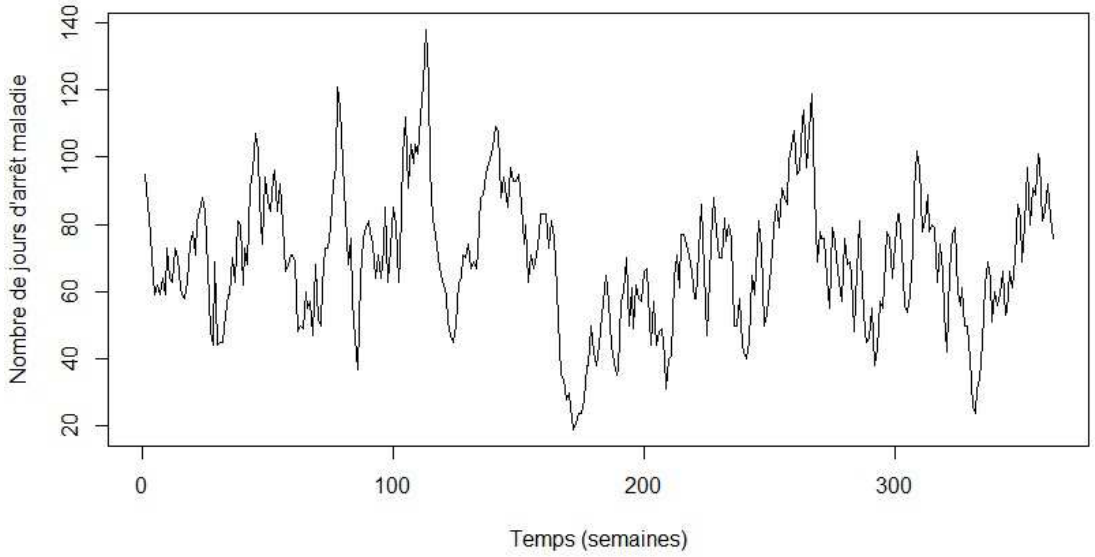


Figure 1: Exemple d'un graphique de suivi du nombre de jours d'absence hebdomadaire pour une entreprise

1.4 Discussion

Le modèle présenté permet d'adapter le modèle de Farrington au cadre des données longitudinales. La construction de la borne supérieure par algorithme de Metropolis-Hastings permet d'éviter l'approximation de la borne avec une transformation à la puissance 2/3 et permet donc de lever cette hypothèse d'intervalle de confiance gaussien.

En pratique, le modèle permet de détecter les entreprises avec des niveaux anormaux d'arrêt maladie et pourrait être d'un véritable intérêt pour les accompagner vers une meilleure gestion des maladies infectieuses et/ou des maladies liées au contexte professionnel. Le modèle ne peut en revanche pas expliquer pourquoi le volume d'absence de l'entreprise est anormalement haut : la détection d'une entreprise en dérive d'arrêt maladie nécessite donc une enquête plus approfondie pour caractériser l'alerte levée.

References

[1] Merete Labriola, Thomas Lund, and Hermann Burr. Prospective study of physical and psychosocial risk factors for sickness absence. *Occupational Medicine*, 56(7):469–474, October 2006.

[2] F. W. O'Reilly and A. B. Stevens. Sickness absence due to influenza. *Occupational Medicine*, 52(5):265–269, August 2002.

[3] Steffen Unkel, C. Paddy Farrington, Paul H. Garthwaite, Chris Robertson, and Nick Andrews. Statistical methods for the prospective detection of infectious disease outbreaks: a review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 175(1):49–82, 2012.

[4] Angela Noufaily, Doyo G. Enki, Paddy Farrington, Paul Garthwaite, Nick Andrews, and André Charlett. An improved algorithm for outbreak detection in multiple surveillance systems. *Statistics in Medicine*, 32(7):1206–1222, March 2013.

[5] C. P. Farrington, N. J. Andrews, A. D. Beale, and M. A. Catchpole. A Statistical Algorithm for the Early Detection of Outbreaks of Infectious Disease. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 159(3):547–563, 1996.

Déconvolution sur $\mathbb{R}_+^d$ par projection sur la base de Laguerre

Florian Dussap [1]

[1] *Université de Paris, CNRS, MAP5 UMR 8145, F-75006 Paris, France*
florian.dussap@parisdescartes.fr

Abstract. We investigate adaptive density estimation in the additive model $\mathbf{Z} = \mathbf{X} + \mathbf{Y}$, where $\mathbf{X}$ and $\mathbf{Y}$ are independent d-dimensional random vectors with non-negative coordinates. Our goal is to recover the density of $\mathbf{X}$ from independent observations of $\mathbf{Z}$, assuming the density of $\mathbf{Y}$ is known. In the $d = 1$ case, an estimation procedure using projection on the Laguerre basis have already been studied. We generalize this procedure in the multivariate case : we establish non-asymptotic upper bounds on the mean integrated squared error of the estimator and we derive convergence rates on anisotropic functional spaces. Moreover, we provide a data-driven strategy for selecting the right projection space. We illustrate these procedures on simulated data.

Keywords. anisotropic multivariate projection estimator, laguerre basis, model selection, nonparametric density estimation

1 Présentation du modèle

On considère un modèle additif dans lequel on observe la somme de la variable d'intérêt $\mathbf{X}$ avec une variable de nuisance $\mathbf{Y}$, indépendante de $\mathbf{X}$, de loi supposée connue ; et on souhaite estimer la loi de $\mathbf{X}$. C'est un problème classique en statistiques non-paramétriques, l'estimateur le plus populaire étant un estimateur à noyau introduit par Stefanski et Carroll (1990).

Nous nous intéressons ici au cas particulier où les variables $\mathbf{X}$ et $\mathbf{Y}$ sont des vecteurs de $\mathbb{R}^d$, à coordonnées positives. Dans le cas unidimensionnel, ce modèle a déjà été étudié par Mabon (2017). Nous généralisons son travail au cas multidimensionnel.

Plus précisément, le modèle considéré s'écrit :

$$\mathbf{Z}_i = \mathbf{X}_i + \mathbf{Y}_i, \quad i = 1, \dots, n \tag{1}$$

où les $\mathbf{X}_i$ (resp. les $\mathbf{Y}_i$) sont des variables aléatoires à valeurs dans $\mathbb{R}_+^d$, i.i.d. de densité f (resp. g), et où les $\mathbf{X}_i$ sont indépendants des $\mathbf{Y}_i$.

Notre but est d'estimer la densité f à partir des observations $\mathbf{Z}_i$, en supposant que la densité g est connue.

2 Estimation

Les variables $\mathbf{Z}_i$ sont i.i.d. et admettent une densité h sur $\mathbb{R}_+^d$, donnée par le produit de convolution de f et g. On suppose que les densités f, g et h appartiennent à $\mathrm{L}^2(\mathbb{R}_+^d)$, et on les décompose sur une base.

On choisit pour base de décomposition la base de Laguerre multivariée. Rappelons que les fonctions de Laguerre (univariées) sont définies sur $\mathbb{R}_+$ par :

$$\forall k \in \mathbb{N}, \quad \varphi_k(x) \coloneqq \sqrt{2}\, L_k(2x)\, \mathrm{e}^{-x}, \text{ avec } L_k(x) \coloneqq \sum_{j=0}^{k} \binom{k}{j} \frac{(-x)^j}{j!}. \tag{2}$$

Pour $\mathbf{k} = (k_1, \ldots, k_d) \in \mathbb{N}^d$ un multi-indice, la fonction de Laguerre multivariée $\varphi_{\mathbf{k}}$ est alors définie sur $\mathbb{R}_+^d$ comme le produit tensoriel des fonctions de Laguerre unidimensionnelle $\varphi_{\mathbf{k}} \coloneqq \varphi_{k_1} \otimes \cdots \otimes \varphi_{k_d}$.

L'intérêt des fonctions de Laguerre pour ce problème réside dans la relation suivante : si $k, j \in \mathbb{N}$, alors le produit de convolution des fonctions de Laguerre s'exprime $\varphi_k * \varphi_j = 2^{-1/2}(\varphi_{k+j} - \varphi_{k+j+1})$. En utilisant cette relation et le fait que $h = f * g$, on montre qu'il existe une relation linéaire inversible entre les coefficients de Laguerre de h et ceux de f.

Cette relation linéaire s'exprime à l'aide d'hypermatrices (tableau de nombre multidimensionnel). Notons $a_{\mathbf{k}}$ (resp. $c_{\mathbf{k}}$) le $\mathbf{k}$-ème coefficient de Laguerre de f (resp. de h). Pour $\mathbf{m} = (m_1, \ldots, m_d) \in (\mathbb{N}^*)^d$, on note $\mathbf{a}_{\mathbf{m}}$ l'hypermatrice des $a_{\mathbf{k}}$ pour $\mathbf{k}$ vérifiant $k_j < m_j$ (de même pour $\mathbf{c}_{\mathbf{m}}$). Il existe alors une hypermatrice $\mathbf{G}_{\mathbf{m}}$ dont les entrées dépendent des coefficients de g telle que :

$$\mathbf{c}_{\mathbf{m}} = \mathbf{G}_{\mathbf{m}} \times_d \mathbf{a}_{\mathbf{m}} \text{ et } \mathbf{a}_{\mathbf{m}} = \mathbf{G}_{\mathbf{m}}^{-1} \times_d \mathbf{c}_{\mathbf{m}} \tag{3}$$

où "$\times_d$" désigne le d–produit contracté d'hypermatrices [1].

Soit $S_{\mathbf{m}}$ le sous-espace de $\mathrm{L}^2(\mathbb{R}_+^d)$ engendré par les $\varphi_{\mathbf{k}}$ pour $\mathbf{k}$ vérifiant $k_j < m_j$. Ce sous-espace est de dimension $D_{\mathbf{m}} \coloneqq m_1 \times \cdots \times m_d$. On approche f par sa projection sur $S_{\mathbf{m}}$, notée $f_{\mathbf{m}}$, et on estime cette dernière. Estimer $f_{\mathbf{m}}$ revient à estimer $\mathbf{a}_{\mathbf{m}}$, ce que l'on peut faire en estimant $\mathbf{c}_{\mathbf{m}}$ d'après la relation (3). Puisque $c_{\mathbf{k}} = \mathbb{E}_f[\varphi_{\mathbf{k}}(\mathbf{Z})]$, on estime cette quantité par la moyenne empirique. Notre estimateur est donc :

$$\hat{f}_{\mathbf{m}} \coloneqq \sum_{\substack{\mathbf{k} \in \mathbb{N}^d \\ \forall j,\, k_j < m_j}} \hat{a}_{\mathbf{k}}\, \varphi_{\mathbf{k}}, \qquad \hat{\mathbf{a}}_{\mathbf{m}} \coloneqq \mathbf{G}_{\mathbf{m}}^{-1} \times_d \hat{\mathbf{c}}_{\mathbf{m}}, \qquad \hat{c}_{\mathbf{k}} \coloneqq \frac{1}{n} \sum_{i=1}^{n} \varphi_{\mathbf{k}}(\mathbf{Z}_i). \tag{4}$$

Il reste à choisir le paramètre $\mathbf{m}$ de l'espace de projection.

1. les notions relatives aux hypermatrices seront rappelées pendant la présentation.

3 Majoration de la MISE et vitesses de convergence

On quantifie la qualité de notre estimateur par sa *Mean Integrated Squarred Error* :

$$\mathrm{MISE}(\hat{f}_{\mathbf{m}}, f) \coloneqq \mathbb{E}_f \|f - \hat{f}_{\mathbf{m}}\|^2_{\mathrm{L}^2}. \tag{5}$$

La MISE se décompose comme la somme d'un terme de biais et d'un terme de variance :

$$\mathrm{MISE}(\hat{f}_{\mathbf{m}}, f) = \|f - f_{\mathbf{m}}\|^2_{\mathrm{L}^2} + \mathbb{E}_f \|f_{\mathbf{m}} - \hat{f}_{\mathbf{m}}\|^2_{\mathrm{L}^2}. \tag{6}$$

Hypothèses On suppose que la loi des $\mathbf{Y}_i$ vérifie les hypothèses suivantes.

(H1) La densité g est bornée.

(H2) Pour toute partie non vide J de $\{1, \dots, d\}$, les moments :

$$M_J(g) \coloneqq \mathbb{E}\left[\prod_{j \in J} \frac{1}{\sqrt{Y^j}}\right] \tag{7}$$

sont finis, où Y^j est la j-ème coordonnée de $\mathbf{Y}$.

Proposition 1. *On suppose que g vérifie **(H1)** et **(H2)**. Alors le terme de variance dans (6) est majoré :*

$$\mathbb{E}_f \|f_{\mathbf{m}} - \hat{f}_{\mathbf{m}}\|^2_{\mathrm{L}^2} \leqslant \frac{c_d(g)\sqrt{D_{\mathbf{m}}}\,\rho^2(\mathbf{G}_{\mathbf{m}}^{-1})}{n} \wedge \frac{\|g\|_\infty \|\mathbf{G}_{\mathbf{m}}^{-1}\|_F^2}{n} \eqqcolon V_g(\mathbf{m}) \tag{8}$$

où $c_d(g)$ est une constante qui dépend des $M_J(g)$, ρ est la norme d'opérateur 2 et $\|\cdot\|_F$ est la norme de Frobenius.

Sous certaines hypothèses techniques de régularité sur la transformée de Laplace de g, on peut montrer que le majorant du terme de variance dans (8) a une croissance au plus polynomiale :

$$V_g(\mathbf{m}) \leqslant \mathbf{m}^{\boldsymbol{\alpha}} \coloneqq m_1^{\alpha_1} \cdots m_d^{\alpha_d} \tag{9}$$

où $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_d) \in (\mathbb{N}^*)^d$ est un multi-indice lié à la régularité de la transformée de Laplace de g.

Pour étudier le terme de biais dans (6), on introduit les espaces de Sobolev–Laguerre multidimensionnels. Dans le cas unidimensionnel, ces espaces ont été introduits initialement par Bongioanni et Torrea (2009) pour étudier l'opérateur de Laguerre. Le lien avec les coefficients de Laguerre d'une fonction a été établi par Comte et Genon-Catalot (2015).

Pour $\mathbf{s} \in (\mathbb{N}^*)^d$ et $L > 0$, on définit la boule Sobolev–Laguerre de paramètres $\mathbf{s}$ et L comme :

$$\mathrm{W}^{\mathbf{s}}(\mathbb{R}^d_+, L) \coloneqq \left\{ f \in \mathrm{L}^2(\mathbb{R}^d_+) \;\middle|\; \sum_{\mathbf{k} \in \mathbb{N}^d} a^2_{\mathbf{k}}(f)\, \mathbf{k}^{\mathbf{s}} \leqslant L \right\} \tag{10}$$

où $a_{\mathbf{k}}(f)$ est le $\mathbf{k}$-ème coefficient de Laguerre de f. On montre que lorsque $f \in \mathrm{W}^{\mathbf{s}}(\mathbb{R}^d_+, L)$, le terme de biais est majoré par $L\,(m_1^{-s_1} + \cdots + m_d^{-s_d})$.

Théorème 1 (Vitesse d'estimation)**.** *Soient* $\mathbf{s} \in (\mathbb{N}^*)^d$ *et* $L > 0$*. Supposons que* g *soit telle que* $V_{\mathbf{m}}(g)$ *vérifie la majoration* (9) *avec* $\boldsymbol{\alpha} \in (\mathbb{N}^*)^d$*. Alors pour* $\mathbf{m}_{\mathrm{opt}} \in (\mathbb{N}^*)^d$ *donné par :*

$$m_{\mathrm{opt},j} \propto n^{1/\left(s_j + s_j \sum_{i=1}^{d} \frac{2\alpha_i}{s_i}\right)}, \quad j = 1, \dots, d, \tag{11}$$

il existe une constante $C > 0$ *qui ne dépend que de* $\mathbf{s}$*,* L *et* g *telle que :*

$$\sup_{f \in \mathrm{W}^{\mathbf{s}}(\mathbb{R}_+^d, L)} \mathrm{MISE}(\hat{f}_{\mathbf{m}_{\mathrm{opt}}}, f) \leqslant C\, n^{-1/\left(1 + \sum_{i=1}^{d} \frac{2\alpha_i}{s_i}\right)}. \tag{12}$$

Remarque 1. Ces vitesses d'estimation sur les boules Sobolev–Laguerre sont similaires à celles trouvées par Comte et Lacour (2013) sur les boules Sobolev anisotropes, pour le problème de déconvolution sur $\mathbb{R}^d$ avec un estimateur à noyau, et pour un bruit *ordinary smooth*.

4 Sélection de modèles

En pratique, la régularité sous-jacente de f n'est pas connue, on ne peut donc pas calculer le modèle optimal du théorème 1. On veut une procédure *data-driven* qui réalise le compromis biais–variance, sans plus d'hypothèses sur f. Plus précisément, on veut choisir $\hat{\mathbf{m}}$ à partir des observations tel que la MISE de $\hat{f}_{\hat{\mathbf{m}}}$ soit proche de la MISE oracle :

$$\inf_{\mathbf{m}} \mathbb{E}_f \|f - \hat{f}_{\mathbf{m}}\|_{\mathrm{L}^2}^2.$$

Soit $\mathbf{m}^* \in (\mathbb{N}^*)^d$ fixé tel que $D_{\mathbf{m}^*} \leqslant n$. On n'explorera que les sous-modèles de $\mathbf{m}^*$ qui ne sont pas "trop grands". Plus précisément, on se restreint à la collection de modèles :

$$\mathcal{M}_n \coloneqq \left\{ \mathbf{m} \in (\mathbb{N}^*)^d \,\middle|\, \forall j,\, m_j \leqslant m_j^* \text{ et } D_{\mathbf{m}}\, \rho^2(\mathbf{G}_{\mathbf{m}}^{-1}) \leqslant \frac{n}{\log n} \right\}. \tag{13}$$

On fait également une hypothèse supplémentaire sur la loi des Y_i.

(H3) Pour tout $b > 0$, $\sum_{m \in \mathcal{M}_n} \rho^2(\mathbf{G}_m^{-1})\, \mathrm{e}^{-b\sqrt{D_m}} \leqslant K(b)$, avec $K(b)$ une constante positive qui ne dépend pas de n.

On utilise une procédure inspirée par Goldenshluger et Lepski (2011), qui a été introduite en sélection de modèles par Chagny (2013) pour l'estimation d'une densité conditionnelle. On montre que cette procédure s'applique à notre problème de déconvolution, dans un cadre multidimensionnel.

On choisit $\hat{\mathbf{m}}$ dans la collection de modèles $\mathcal{M}_n$ qui minimise :

$$\hat{\mathbf{m}} \coloneqq \operatorname*{arg\,min}_{\mathbf{m} \in \mathcal{M}_n} \left[A_g(\mathbf{m}) + \kappa_2 \tilde{V}_g(\mathbf{m}) \right] \tag{14}$$

où $A_g(\mathbf{m})$ et $\tilde{V}_g(\mathbf{m})$ sont définis par :

$$\tilde{V}_g(\mathbf{m}) \coloneqq \frac{c_d(g)\sqrt{D_{\mathbf{m}}}\,\rho^2(\mathbf{G}_{\mathbf{m}}^{-1})}{n} \wedge \frac{(\|g\|_\infty \vee 1)\|\mathbf{G}_{\mathbf{m}}^{-1}\|_F^2 \log(n)}{n} \tag{15}$$

$$A_g(\mathbf{m}) \coloneqq \max_{\mathbf{m}' \in \mathcal{M}_n} \left(\left\| \hat{f}_{\mathbf{m}'} - \hat{f}_{\mathbf{m}\wedge\mathbf{m}'} \right\|_{\mathrm{L}^2}^2 - \kappa_1 \tilde{V}_g(\mathbf{m}') \right)_+ \tag{16}$$

et où κ_1, κ_2 sont deux constantes numériques à ajuster.

Théorème 2 (Inégalité oracle)**.** *Sous les hypothèse **(H1)** à **(H3)**, il existe une constante numérique $\kappa_0(d) > 0$ qui dépend de la dimension d telle que pour tout choix de κ_1, κ_2 vérifiant $\kappa_0(d) < \kappa_1 \leqslant \kappa_2$, l'estimateur $\hat{f}_{\hat{\mathbf{m}}}$ vérifie :*

$$\forall f \in \mathrm{L}^2(\mathbb{R}_+^d), \quad \mathrm{MISE}(\hat{f}_{\hat{\mathbf{m}}}, f) \leqslant C \inf_{\mathbf{m}\in\mathcal{M}_n} \left(\|f - f_{\mathbf{m}}\|_{\mathrm{L}^2}^2 + \tilde{V}(\mathbf{m}) \right) + \frac{C'}{n}, \tag{17}$$

avec C une constante positive qui dépend de κ_1 et κ_2, et C' une constante positive qui dépend de g, d et κ_1.

5 Illustrations

Illustrons notre procédure d'estimation sur un exemple en dimension $d = 2$ sur des données simulées. Pour cela, on génère $\mathbf{X} = (X^1, X^2)$ de la façon suivante. On génère $\mathbf{W} = (W^1, W^2) \in \mathbb{R}_+^2$ avec W^1 indépendant de W^2, puis on multiplie $\mathbf{W}$ par une matrice :

$$\begin{bmatrix} X^1 \\ X^2 \end{bmatrix} = \begin{bmatrix} 1 & 0.1 \\ 0.2 & 1 \end{bmatrix} \begin{bmatrix} W^1 \\ W^2 \end{bmatrix}. \tag{18}$$

On obtient ainsi un vecteur aléatoire $\mathbf{X}$ de $\mathbb{R}_+^2$ dont les composantes sont corrélées. Pour cet exemple, on simule W^1 de loi $\Gamma(3,1)$ et W^2 de loi Bêta $\mathrm{B}(4,5)$ renormalisée pour être de variance égale à 1.

Pour la variable de nuisance, on choisit de simuler $\mathbf{Y}$ de loi $\Gamma(2,\sqrt{20}) \otimes \Gamma(2,\sqrt{20})$. En effet, les coefficients de Laguerre des lois Gamma se calculent explicitement, ce qui permet de calculer l'hypermatrice $\mathbf{G}_{\mathbf{m}}$ facilement.

Les constantes κ_1 et κ_2 ont été calibrées sur plusieurs exemples au préalable : on choisit $\kappa_1 = \kappa_2 = 10^{-5}$. De plus, on choisit comme modèle maximal $\mathbf{m}^* = (12, 12)$. On remarque qu'en pratique, la collection de modèles $\mathcal{M}_n$ définie par (14) est trop petite. C'est pourquoi on utilisera plutôt la collection :

$$\mathcal{M}_n' \coloneqq \left\{ \mathbf{m} \in (\mathbb{N}^*)^2 \,\middle|\, \forall j,\, m_j \leqslant m_j^* \text{ et } D_{\mathbf{m}}\,\rho^2(\mathbf{G}_{\mathbf{m}}^{-1}) \leqslant K\frac{n}{\log n} \right\}, \tag{19}$$

avec $K = 10^3$ ou 10^4.

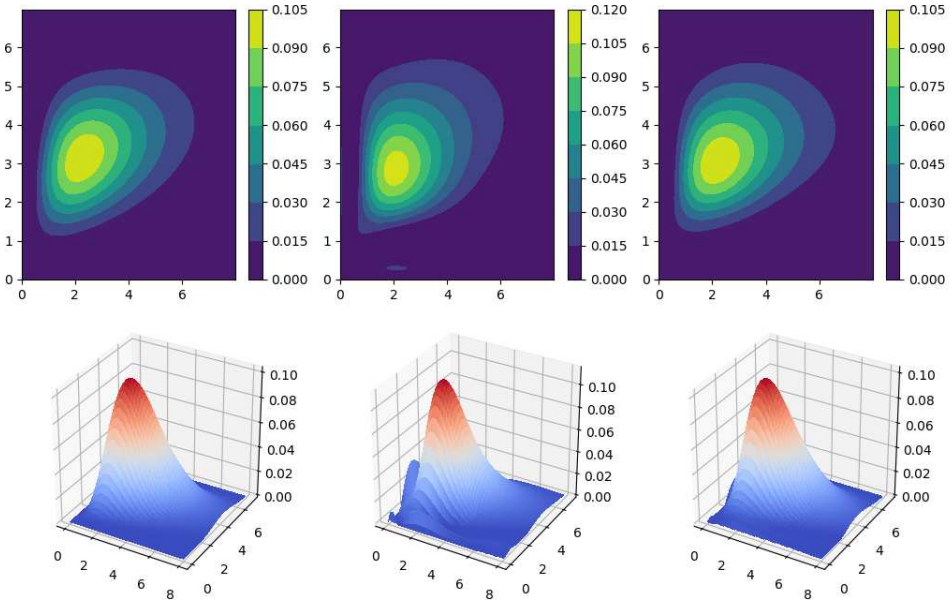


FIGURE 1 – Gauche : vraie densité. Centre : estimation pour $n = 500$. Droite : estimation pour $n = 5000$.

La figure 1 montre le résultat d'une estimation sur un échantillon de taille $n = 500$ et sur un échantillon de taille $n = 5000$. Dans le premier cas, le modèle sélectionné est $\hat{\mathbf{m}} = (5, 4)$ et dans le second cas, $\hat{\mathbf{m}} = (4, 9)$. On remarque dans les deux cas que le modèle sélectionné est anisotrope et de dimension bien inférieure à celle du modèle maximal ($D_{\mathbf{m}^*} = 144$). L'estimateur $\hat{f}_{\hat{\mathbf{m}}}$ s'adapte bien à la régularité de f.

Références

Bruno BONGIOANNI et José L. TORREA : What is a Sobolev space for the Laguerre function systems ? *Studia Mathematica*, 192(2):147–172, 2009.

Gaëlle CHAGNY : Warped bases for conditional density estimation. *Mathematical Methods of Statistics*, 22(4):253–282, octobre 2013.

Fabienne COMTE et Valentine GENON-CATALOT : Adaptive Laguerre density estimation for mixed Poisson models. *Electronic Journal of Statistics*, 9(1):1113–1149, 2015.

Fabienne COMTE et Claire LACOUR : Anisotropic adaptive kernel deconvolution. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 49(2):569–609, mai 2013.

Alexander Goldenshluger et Oleg Lepski : Bandwidth selection in kernel density estimation : Oracle inequalities and adaptive minimax optimality. *The Annals of Statistics*, 39(3):1608–1632, juin 2011.

Gwennaëlle Mabon : Adaptive Deconvolution on the Non-negative Real Line. *Scandinavian Journal of Statistics*, 44(3):707–740, septembre 2017.

Leonard A. Stefanski et Raymond J. Carroll : Deconvolving kernel density estimators. *Statistics*, 21(2):169–184, janvier 1990.

Asymptotic distribution of the test for change-points detection based on two-sample U-Statistics when the observations are absolutely regular

Echarif Elharfaoui [1] & Michel Harel [2] & Joseph Ngatchou-Wandji [3]

[1] *Université Chouaib Doukkali, Faculté des Sciences, 24000 El Jadida, Maroc*
elharfaoui.e@ucd.ac.ma
[2] *Université de Limoges, France et* [3] *Université de Lorraine, France*

Abstract. In this paper, we study change-points detection test using two-sample tests based on U-statistics for weak dependence (essentially absolutely regular) observations. The asymptotic distribution of this test has been studied by Dehling, Fried, Garcia and Wendler (2015) under restrictive technical conditions and was only applied to particular kind of kernels of the U-statistics. We extend their results for more general models and tests.

Keywords. U-statistics, nonparametric change-point tests, weighted approximation, weak invariance, Wiener process, geometrical absolute regularity.

1 Introduction

Change-point tests address the question whether a stochastic process is stationary during the entire observation period or not. In the case of independent data, there is a well-developed theory; see the book of Csörgő and Horváth (1997) for a complete survey. When the data are dependent, the CUSUM statistic has been intensively studied again by Csörgő and Horváth (1997). The CUSUM statistic for detection of jumps in mean is known to be sensitive to outliers. Change-point analysis has been massively analyzed and developed, and have various application in all kinds of fields such as industry quality control, financial market, and medical diagnostics.

Let $X_1, X_2, ..., X_n$ be real dependent random variables with distribution functions $F_1, F_2, ..., F_n$. In this paper, we will assume that the sequence $\{X_i\}_{i\in\mathbb{N}}$ is absolutely regular with the rate

$$\beta(n) = \mathcal{O}(\tau^n),\ 0 < \tau < 1 \tag{1}$$

where

$$\beta(k) = \sup_{n\in\mathbb{N}} \max_{1\le j\le n-k} E\Big\{ \sup_{A\in\mathcal{A}_{j+k}^{\infty}} \big|P(A \mid \mathcal{A}_0^j) - P(A)\big| \Big\}$$

with $\mathcal{A}_i^j$ the σ-algebra generated by $X_i, \ldots, X_j,\ i, j \in \mathbb{N} \cup \{\infty\}$.

One of the main topics in nonparametric statistic is the detection of a possible difference in the distributions of $X_1, X_2, \ldots, X_n$. A model with one abrupt change is very often used and it is formulated way: we wish to test the no change null hypothesis

$\mathcal{H}_0 : F_1(x) = F_2(x) = \ldots = F_n(x)$ for all $x \in \mathbb{R}$
against the alternative hypothesis that there is a change-point in the sequence $X_1, X_2, ..., X_n$, namely that we have

$\mathcal{H}_1$: there is a $\lambda \in (0,1)$ such that $F_1(x) = F_2(x) = \ldots = F_{[n\lambda]}(x) = F(x)$,
$F_{[n\lambda]+1}(x) = \ldots = F_n(x) = G(x), \forall x \in \mathbb{R}$ and $F(x_0) \neq G(x_0)$ for some $x_0 \in \mathbb{R}$.
Under the null hypothesis $\mathcal{H}_0$ we have $F = G$.

In order to derive the asymptotic distribution of the test, we study the stochastic process

$$Z_n^*(\lambda) = n^{-3/2} \sum_{i=1}^{[n\lambda]} \sum_{j=[n\lambda]+1}^{n} h(X_i, X_j),\ 0 \leq \lambda \leq 1,$$

where $h : \mathbb{R}^2 \to \mathbb{R}$ is a kernel function. In the case of independent data, the asymptotic distribution of this process has been studied by Csörgő and Horváth (1988), then in the dependent case by Dehling, Fried, Garcia and Wendler (2015) under restrictive conditions such as the notions of one-side functional and 1-continuity. We prove their results in the dependent case without their restrictive conditions and we can apply it for more general processes. Moreover Dehling, Fried, Garcia and Wendler (2015) study the weak convergence of the estimator $Z_n^*(\lambda)$ only for bounded kernels whereas we also establish the weak convergence of the estimator for unbounded kernels and also under the alternative. More, we generalize our results for the case of multi-change-points.

We study the asymptotic behavior of $Z_n^*(\lambda)$ under the null hypothesis in Section 2 and under the alternative hypothesis in Section 3. Csörgő and Horváth (1997) proved it in probability for i.i.d. random variables and in this paper we generalize their results for the almost sure convergence when the random variables are absolutely regular.

2 Asymptotic under $\mathcal{H}_0$

Now, we assume that h is symmetric, i.e $h(x,y) = h(y,x)$ for all $x, y \in \mathbb{R}$.
Denote

$$Z_n(\lambda) = n^{-3/2} \sum_{i=1}^{[n\lambda]} \sum_{j=[n\lambda]+1}^{n} \big\{h(X_i, X_j) - \theta(F)\big\},\ 0 \leq \lambda \leq 1$$

where

$$\theta(F) = \int \int h(x,y) dF(x) dF(y).$$

Define the U-statistic U_n with h as kernel by

$$U_n = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} h(X_i, X_j).$$

Let

$$h_1^{(1)}(x) = \int h(x,y) dF(y) - \theta(F), \quad h_2^{(1)}(y) = \int h(x,y) dF(x) - \theta(F)$$

and

$$g(x,y) = h(x,y) - h_1^{(1)}(x) - h_2^{(1)}(y) + \theta(F).$$

Consider the Hoeffding's decomposition of the U-statistics U_n under $\mathcal{H}_0$

$$U_n = \theta(F) + U_{n,1}^{(1)} + U_{n,2}^{(1)} + U_n^{(2)} \tag{2}$$

where

$$U_{n,1}^{(1)} = n^{-1} \sum_{i=1}^{n} h_1^{(1)}(X_i), \quad U_{n,2}^{(1)} = n^{-1} \sum_{i=1}^{n} h_2^{(1)}(X_i)$$

and

$$U_n^{(2)} = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} \left\{ h(X_i, X_j) - h_1^{(1)}(X_i) - h_2^{(1)}(X_j) \right\} + \theta(F).$$

Let

$$\sigma_{kl} = E(h_k^{(1)}(X_1) h_l^{(1)}(X_1)) + 2 \sum_{j=1}^{\infty} Cov(h_k^{(1)}(X_1), h_l^{(1)}(X_{1+j})), \ k,l = 1,2.$$

Theorem 1. *Under $\mathcal{H}_0$, if $E|h(X_i,X_j)|^{2+\delta} < \infty$, $\int\int_{\mathbb{R}^2} |h(x,y)|^{2+\delta} dF(x) dF(y)$ for some $\delta > 0$ and the condition of absolute regularity (1) is satisfied, then $\sigma_{kl} < \infty$. If $\sigma_{kl} > 0$, $k,l = 1,2$ then the distribution of $\{Z_n(\lambda); 0 \le \lambda \le 1\}_{n \in \mathbb{N}}$ converges in distribution towards a mean-zero Gaussian process with representation*

$$Z(\lambda) = (1-\lambda) W_1(\lambda) + \lambda (W_2(1) - W_2(\lambda)), \ 0 \le \lambda \le 1$$

where $\{W_1(\lambda), W_2(\lambda)\}_{0 \le \lambda \le 1}$ is a two-dimensional Brownian motion with mean zero and covariance function $cov(W_k(s), W_l(t)) = min(s,t)\sigma_{kl}$, $k,l = 1,2$.

Sketch of the proof From the Hoeffding decomposition (2) of the U-statistics, from Lemma 1 of Yoshihara (1976) and from Propositions 1, 2 and 3 in the Appendix. □

3 Asymptotic under $\mathcal{H}_1$

First we introduce some notation.
Let

$$\mu(G) = \int\int_{\mathbb{R}^2} h(x,y) dG(x) dG(y), \quad \tau(F,G) = \int\int_{\mathbb{R}^2} h(x,y) dF(x) dG(y),$$

$$h_1^{(2)}(x) = \int h(x,y)dG(y) - \tau(F,G) \quad \text{and} \quad h_2^{(2)}(y) = \int h(x,y)dG(x) - \tau(F,G).$$

Let also $(Y_i)_{1\leq i\leq n}$ a sequence of absolute regular random variables with the same rate as the sequence $(X_i)_{1\leq i\leq n}$, under $\mathcal{H}_0$ and the distribution function of Y_i is G. For any $i, j \in \mathbb{N}$, the absolute regular dependence between Y_i and Y_j is the same as the dependence between X_i and X_j.
Now, we use the results of the preceding sections to test the null hypothesis $\mathcal{H}_0$

$\mathcal{H}_0 : F_1(x) = F_2(x) = \ldots = F_n(x)$ for all $x \in \mathbb{R}$
against the local alternative $\mathcal{H}_1$ and the local alternatives defined respectively by

$\mathcal{H}_1$: There is a $\lambda_0 \in (0,1)$ such as $F_1(x) = F_2(x) = \ldots = F_{[n\lambda_0]}(x) = F(x)$, $F_{[n\lambda_0]+1}(x) = \ldots = F_n(x) = G(x)$ for all $x \in \mathbb{R}$, it exists some $x_0 \in \mathbb{R}$ such as $F(x_0) \neq G(x_0)$ and $\theta(F) \neq \tau(F,G)$.

$\mathcal{H}_{1,n}$: There is a $\lambda_0 \in (0,1)$ such as $F_1(x) = F_2(x) = \ldots = F_{[n\lambda_0]}(x) = F(x)$, $F_{[n\lambda_0]+1}(x) = \ldots = F_n(x) = G(x)$ for all $x \in \mathbb{R}$, it exists some $x_0 \in \mathbb{R}$ such as $F(x_0) \neq G(x_0)$ and $\tau(F,G) = \theta(F) + n^{-1/2}A$ where A is some constant for each n of the sequence $X_1, X_2, ..., X_n$.

Theorem 2. *Under $\mathcal{H}_{1,n}$, if $E|h(X_i,Y_j)|^{2+\delta} < \infty$, $\int\int_{\mathbb{R}^2} |h(x,y)|^{2+\delta}dF(x)dG(y) < \infty$ for some $\delta > 0$ and condition (1), then the distribution of $\{Z_n(\lambda); 0 \leq \lambda \leq 1\}_{n\in\mathbb{N}}$ converges in distribution towards a Gaussian process with mean $(1-\lambda)\lambda A$ and with representation*

$$Z(\lambda) - (1-\lambda)\lambda A = (1-\lambda)W_1(\lambda) + \lambda(W_2(1) - W_2(\lambda)), \quad 0 \leq \lambda \leq 1$$

where $\{W_1(\lambda), W_2(\lambda)\}_{0\leq\lambda\leq 1}$ is the two-dimensional Brownian motion with mean zero and covariance function $cov(W_k(s), W_l(t)) = \min(s,t)\sigma_{k,l}$, $k, l = 1, 2$.

Sketch of the proof Similar to the prove of Theorem 1 by changing the covariance structure of the Brownian motion. □

Corollary 1. *Suppose that $F_1(x) = F_2(x) = \ldots = F_{[n\lambda_0]}(x) = F(x)$, $F_{[n\lambda_0]+1}(x) = \ldots = F_n(x) = G(x)$ for all $x \in \mathbb{R}$ and it exist a constant B such as $G(x) = F(x + n^{-1/2}B)$ and suppose also that the kernel function h is twice derivative and $\int\int(\partial h(x,y)/\partial y)dF(x)dG(y) < \infty$ and $\partial^2 h(x,y)/\partial^2 y$ is bounded, then we have the alternative $\mathcal{H}_{1,n}$.*

Proof It is easy deduced from the Taylor-Young formula. □

Let

$$g^{(2)}(x,y) = h(x,y) - h_1^{(2)}(x) - h_2^{(2)}(y) + \tau(F,G).$$

Theorem 3. *We assume that $\mathcal{H}_1$ holds, the conditions of integrability of Theorem 2 and condition (1) are satisfied, then*

$$n^{-1/2}Z_n^*(t) \xrightarrow[n\to\infty]{a.s.} \begin{cases} \theta(F)t(\lambda_0 - t) + \tau(F,G)t(1-\lambda_0), & 0 \leq t \leq \lambda_0 \\ \mu(G)(t-\lambda_0)(1-t) + \tau(F,G)\lambda_0(1-t), & \lambda_0 \leq t < 1. \end{cases} \quad (3)$$

Sketch of the proof Let $1 \leq [(n+1)t] \leq [n\lambda_0]$, then

$$\begin{aligned} Z_n^*(t) &= n^{-3/2} \sum_{1\leq i<j\leq [n\lambda_0]} h(X_i, X_j) + n^{-3/2} \sum_{i=1}^{[n\lambda_0]} \sum_{j=[n\lambda_0]+1}^{n} h(X_i, X_j) \\ &\quad -n^{-3/2}\Bigg\{ \sum_{1\leq i<j\leq [(n+1)t]} h(X_i, X_j) + \sum_{[(n+1)t]+1\leq i<j\leq [n\lambda_0]} h(X_i, X_j) \\ &\quad + \sum_{[(n+1)t]+1\leq i\leq [n\lambda_0]} \sum_{[n\lambda_0]+1\leq j\leq n} h(X_i, X_j)\Bigg\} \\ &= R_n^{(1)} + R_n^{(2)} - \{R_n^{(3)} + R_n^{(4)} + R_n^{(5)}\}. \end{aligned}$$

From the Hoeffding decomposition (2), from Markov inequality, from Borel-Cantelli Lemma, from Lemma 2 in Appendix and Lemma 2 of Yoshihara (1976), we get the convergence of Z_n^*.

We deduce from Theorem 3, the following Corollary.

Corollary 2. *We assume that $\mathcal{H}_0$ holds, and the condition of Theorem 1, then*

$$n^{-1/2} Z_n^*(t) \xrightarrow{a.s.}_{n\to\infty} \theta(F) t(1-t).$$

4 Appendix

Proposition 1. *Under the conditions of Theorem 1, we have*

$$n^{-3/2} \sup_{0\leq\lambda\leq 1} \left| \sum_{i=1}^{[n\lambda]} \sum_{j=[n\lambda]+1}^{n} g(X_i, X_j) \right| \xrightarrow{P}_{n\to\infty} 0.$$

Under the conditions of Theorem 2, we have

$$n^{-3/2} \sup_{0\leq\lambda\leq 1} \left| \sum_{i=1}^{[n\lambda]} \sum_{j=[n\lambda]+1}^{n} g(X_i, Y_j) \right| \xrightarrow{P}_{n\to\infty} 0.$$

Sketch of the proof Without loss of generality, we proof it only under the conditions of Theorem 1. The proof of Proposition 1 needs two lemmas.

Lemma 1. *Under the conditions of Theorem 1, there exists a Constant $C > 0$ such that*

$$E\left(\sum_{i=1}^{[n\lambda]} \sum_{j=[n\lambda]+1}^{n} g(X_i, X_j) \right)^2 \leq Cst[n\lambda](n - [n\lambda]).$$

Lemma 2. *Under the conditions of Theorem 1, we have*

$$E\Big(\Big|G_n(\lambda) - G_n(\lambda')\Big|^2\Big) \leq \frac{Cst}{n}(\lambda - \lambda'), \text{ for all } 0 \leq \lambda' \leq \lambda \leq 1.$$

Proposition 2. *Let $X_1, X_2, ..., X_n, ...$ be a strictly stationary sequence of random variables satisfying the strong mixing condition, if $\sup_{n\geq 1} E\,|X_i|^{2+\delta} < \infty$ and if*

$$\sum_{i=1}^{\infty} (\alpha(i))^{\frac{\delta}{2+\delta}} < \infty$$

then $\sigma_*^2 < \infty$. If $\sigma_* > 0$, then the distribution function of $S_n(\lambda) = \frac{1}{\sigma_*\sqrt{n}} \sum_{i=1}^{[n\lambda]} X_i$ converges weakly to a Wiener measure on $(D, \mathcal{D})$ where

$$\sigma_*^2 = E(X_1^2) + 2\sum_{i=1}^{\infty} E(X_1 X_{i+1}) \tag{4}$$

and $\mathcal{D}$ is the σ-fields of Borel sets for the Skorohod topology.

Sketch of the proof The first part of the Proposition 2 is to prove that the finite dimensional distributions of S_n converges weakly to those of W and the second part of the proof is to prove that S_n is tight. □

Proposition 3. *Under the conditions of Theorem 1, we have*

$$\Big\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{[n\lambda]} \begin{pmatrix} h_1^{(1)}(X_i) \\ h_2^{(1)}(X_i) \end{pmatrix} \Big\}_{0\leq\lambda\leq 1} \to_{n\to\infty} \Big\{ \begin{pmatrix} W_1(\lambda) \\ W_2(\lambda) \end{pmatrix} \Big\}_{0\leq\lambda\leq 1}. \tag{5}$$

Under the conditions of Theorem 2, we have

$$\Big\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{[n\lambda]} \begin{pmatrix} h_1^{(2)}(X_i) \\ h_2^{(2)}(Y_i) \end{pmatrix} \Big\}_{0\leq\lambda\leq 1} \to_{n\to\infty} \Big\{ \begin{pmatrix} W_1(\lambda) \\ W_2(\lambda) \end{pmatrix} \Big\}_{0\leq\lambda\leq 1}. \tag{6}$$

Sketch of the proof The proofs of (5) and (6) are similar. We will prove it only for (5). To prove (6), we need to establish finite dimensional convergence and tightness.

Bibliography

Csörgő, M. and Horváth, L. (1988). Invariance principales for changepoint problems, *J. Multivariate Anal.*, 27, pp. 151-168.

Csörgő, M. and Horváth, L. (1997). Limit Theorems in Change-Point Analysis, *John Wiley & Sons*, New York; Chichester.

Dehling, H., Fried, R., Garcia, I. and Wendler, M. (2015). Change-Point Detection Under Dependence Based on Two-Sample U-Statistics, *Asymptotic Laws and Methods in Stochastics*, Fields Institute Communications, Springer, New York, NY, 76, pp. 195-220.

Yoshihara, K. (1976). Limiting behavior of U-statistics for stationary absolutely regular processes, *Z. Wahrscheinlichkeitstheorie Verw. Gebiete*, 35, pp. 237-252.

Estimating drift parameters in a non-ergodic Gaussian Vasicek-type model

Khalifa Es-Sebaiy [1] & Mohammed Es.Sebaiy [2]

[1] *Department of Mathematics, Faculty of Science, Kuwait University, Kuwait. Email: khalifasbai@gmail.com, khalifa.essebaiy@ku.edu.kw*

[2] *Cadi Ayyad University, Marrakech, Moroccco. Email: mohammedsebaiy@gmail.com*

Abstract. We study a problem of parameter estimation for a non-ergodic Gaussian Vasicek-type model defined as $dX_t = \theta(\mu + X_t)dt + dG_t,\ t \geq 0$ with unknown parameters $\theta > 0$, $\mu \in \mathbb{R}$ and $\alpha := \theta\mu$, where G is a Gaussian process. We provide least square-type estimators $(\widetilde{\theta}_T, \widetilde{\mu}_T)$ and $(\widetilde{\theta}_T, \widetilde{\alpha}_T)$, respectively, for (θ, μ) and (θ, α) based a continuous-time observation of $\{X_t,\ t \in [0,T]\}$ as $T \to \infty$. Our aim is to derive some sufficient conditions on the driving Gaussian process G in order to ensure the strongly consistency and the joint asymptotic distribution of $(\widetilde{\theta}_T, \widetilde{\mu}_T)$ and $(\widetilde{\theta}_T, \widetilde{\alpha}_T)$. Moreover, we obtain that the limit distribution of $\widetilde{\theta}_T$ is a Cauchy-type distribution, and $\widetilde{\mu}_T$ and $\widetilde{\alpha}_T$ are asymptotically normal. We apply our result to fractional Vasicek, subfractional Vasicek and bifractional Vasicek processes. This work extends the results of [7] studied in the case where $\mu = 0$.

Keywords. Parametric estimation, Fractional Gaussian processes, Young integral.

1 Introduction

Let $G := \{G_t, t \geq 0\}$ be a centered Gaussian process satisfying the following assumption

$(\mathcal{A}_1)$ There exist constants $c > 0$ and $\gamma \in (0,1)$ such that for every $s, t \geq 0$,

$$G_0 = 0, \qquad E\left[(G_t - G_s)^2\right] \leq c\,|t-s|^{2\gamma}.$$

Note that, if $(\mathcal{A}_1)$ holds, then by the Kolmogorov-Centsov theorem, we can conclude that for all $\varepsilon \in (0, \gamma)$, the process G admits a modification with $(\gamma - \varepsilon)-$Hölder continuous paths, still denoted G in the sequel.

In the present paper, our goal is to estimate jointly the drift parameters of the Gaussian Vasicek-type (also called mean-reverting Ornstein-Uhlenbeck) process $X := \{X_t, t \geq 0\}$ that is defined as the unique (pathwise) solution to

$$X_0 = 0, \qquad dX_t = \theta\,(\mu + X_t)\,dt + dG_t, \quad t \geq 0, \tag{1}$$

where $\theta > 0$ and $\mu \in \mathbb{R}$ are considered as unknown parameters.

In recent years, the study of various problems related to the model (1) has attracted interest. When G is a standard Brownian motion, the model (1) with $\mu = 0$ was originally proposed by Ornstein and Uhlenbeck and then it was generalized by Vasicek, see [13]. In the finance context, and if the driving process G is a fractional Brownian motion (fBm) with Hurst parameter $H \in (0, 1)$, the paper [10] on rough volatility contends that the short-time behavior indicates that the Hurst parameter H in the volatility is less than $1/2$. Also, several papers have studied models which have long-memory, that is the memory parameter H in the volatility is greater than $1/2$; see for example [2, 3, 4, 5].

An example of interesting problem related to (1) is the statistical estimation of μ and θ when one observes the whole trajectory of X. In order to estimate the unknown parameters θ and μ when the whole trajectory of X defined in (1) is observed, we will first consider the classical least squares estimators (LSEs) $\widehat{\theta}_T$ and $\widehat{\alpha}_T$ for θ and $\alpha := \mu\theta$, respectively. By minimizing (formally) the function

$$F(\theta, \alpha) = \int_0^T \left| \dot{X}_s - (\alpha + \theta X_s) \right|^2 ds,$$

we obtain

$$\widehat{\theta}_T = \frac{T \int_0^T X_s dX_s - X_T \int_0^T X_s ds}{T \int_0^T X_s^2 ds - \left(\int_0^T X_s ds \right)^2} \tag{2}$$

and

$$\widehat{\alpha}_T = \frac{X_T \int_0^T X_s^2 ds - \int_0^T X_s dX_s \int_0^T X_s ds}{T \int_0^T X_s^2 ds - \left(\int_0^T X_s ds \right)^2}. \tag{3}$$

More precisely, $(\widehat{\theta}_T, \widehat{\alpha}_T)$ is the solution of the system $\frac{\partial F}{\partial \theta}(\theta, \alpha) = 0$, $\frac{\partial F}{\partial \alpha}(\theta, \alpha) = 0$. Moreover, the expressions given in (2) and (3) are well-defined for $\frac{1}{2} < \gamma < 1$, since the stochastic integral $\int_0^T X_s dX_s$ is well-defined in the Young sense for $\frac{1}{2} < \gamma < 1$ only, by using $(\mathcal{A}_1)$. In this case we can write $\int_0^T X_s dX_s = \frac{1}{2} X_T^2$ for $\frac{1}{2} < \gamma < 1$. Thus, we can extend the estimators $\widehat{\theta}_T$ and $\widehat{\alpha}_T$ to all $0 < \gamma < 1$, as follows

$$\widetilde{\theta}_T = \frac{\frac{1}{2} T X_T^2 - X_T \int_0^T X_s ds}{T \int_0^T X_s^2 ds - \left(\int_0^T X_s ds \right)^2} \tag{4}$$

and

$$\widetilde{\alpha}_T = \frac{X_T \int_0^T X_s^2 ds - \frac{1}{2} X_T^2 \int_0^T X_s ds}{T \int_0^T X_s^2 ds - \left(\int_0^T X_s ds \right)^2}. \tag{5}$$

Furthermore, we can obtain a least squares-type estimator $\widetilde{\mu}_T$ for μ, that is the statistic

$$\widetilde{\mu}_T = \widetilde{\alpha}_T / \widetilde{\theta}_T = \frac{\int_0^T X_s^2 ds - \frac{1}{2} X_T \int_0^T X_s ds}{\frac{1}{2} T X_T - \int_0^T X_s ds}. \tag{6}$$

Now, notice that the estimators (4), (5) and (6) are well-defined for all $\gamma \in (0,1)$, and not only for $\gamma \in (\frac{1}{2},1)$. This then allows us to consider $(\widetilde{\theta}_T, \widetilde{\mu}_T)$ as estimator to estimate the parameters (θ,μ) of the equation (1), and $(\widetilde{\theta}_T, \widetilde{\alpha}_T)$ as estimator to estimate the parameters (θ,α) of the process (1) in the form

$$X_0 = 0, \quad dX_t = (\alpha + \theta X_t)\, dt + dG_t, \quad t \geq 0, \tag{7}$$

for all $\gamma \in (0,1)$.

We apply our approach to some Vasicek Gaussian processes as follows:
Fractional Vasicek process:
Suppose that the process G given in (1) is a fractional Brownian motion with Hurst parameter $H \in (0,1)$. When $H \in (\frac{1}{2},1)$, the parameter estimation for θ and μ has been studied in [**?**] by using the LSEs $\widehat{\theta}_T$ and $\widehat{\mu}_T$ which coincide, respectively, with $\widetilde{\theta}_T$ and $\widetilde{\mu}_T$ for $H \in (\frac{1}{2},1)$. Here we present a study valid for all $H \in (0,1)$. Moreover, we study the joint asymptotic distribution of $(\widetilde{\theta}_T, \widetilde{\mu}_T)$.

Subfractional Vasicek process:
Assume that the process G given in (1) is a subfractional Brownian motion S^H with parameter $H \in (0,1)$, that is, S^H is a centered Gaussian process with covariance function

$$E\left(S_t^H S_s^H\right) = t^{2H} + s^{2H} - \frac{1}{2}\left((t+s)^{2H} + |t-s|^{2H}\right); \quad s,\ t \geq 0.$$

For $H > \frac{1}{2}$, using the LSEs $\widehat{\theta}_T$ and $\widehat{\mu}_T$ which also coincide, respectively, with $\widetilde{\theta}_T$ and $\widetilde{\mu}_T$, the statistical estimation for θ and μ has been discussed in [14]. But the proof of the asymptotic distribution (3.32) of $\widehat{\mu}_T$ in [14] relies on a possible flawed technique because as $T \to \infty$, the value of the limit given in (3.32) depends on T, which gives a contradiction. Here, we give a solution for this problem, and we extend the results of [14] to all $H \in (0,1)$, and we also study the joint asymptotic distribution of $(\widetilde{\theta}_T, \widetilde{\mu}_T)$.

Bifractional Vasicek process:
To the best of our knowledge there is no study of the problem of estimating the drift of (1) in the case when G is a bifractional Brownian motion $B^{H,K}$ with parameters $(H,K) \in (0,1)^2$, that is, $B^{H,K}$ is a centered Gaussian process with the covariance function

$$E(B_s^{H,K} B_t^{H,K}) = \frac{1}{2^K}\left(\left(t^{2H} + s^{2H}\right)^K - |t-s|^{2HK}\right); \quad s,t \geq 0.$$

Here we analyzed this question.

2 Main Results

Here we analyze the asymptotic behavior of the LSEs $(\widetilde{\theta}_T, \widetilde{\mu}_T)$ and $(\widetilde{\theta}_T, \widetilde{\alpha}_T)$.

2.1 Strong consistency

In the following theorem we prove the strong consistency of the estimators $\widetilde{\theta}_T$, $\widetilde{\mu}_T$ and $\widetilde{\alpha}_T$.

Theorem 2.1 *Assume that* $(\mathcal{A}_1)$ *holds and let* $\widetilde{\theta}_T$ *and* $\widetilde{\mu}_T$ *be given by (4) and (6) for every* $T \geq 0$. *Then*

$$\widetilde{\theta}_T \longrightarrow \theta, \tag{8}$$

and

$$\widetilde{\mu}_T \longrightarrow \mu \tag{9}$$

almost surely, as $T \to \infty$. *As a consequence, we deduce that the estimator* $\widetilde{\alpha}_T$ *given in (5) is also strongly consistent, that is*

$$\widetilde{\alpha}_T = \widetilde{\mu}_T \widetilde{\theta}_T \longrightarrow \alpha = \mu\theta$$

almost surely, as $T \to \infty$.

2.2 Asymptotic distribution

In this section the following assumptions are required:

$(\mathcal{A}_2)$ There exist $\lambda_G > 0$ and $\eta \in (0, 1)$ such that, as $T \to \infty$

$$\frac{E\left(G_T^2\right)}{T^{2\eta}} \longrightarrow \lambda_G^2.$$

$(\mathcal{A}_3)$ The limiting variance of $e^{-\theta T}\int_0^T e^{\theta s} dG_s$ exists as $T \to \infty$ i.e., there exists a constant $\sigma_G > 0$ such that

$$\lim_{T\to\infty} E\left[\left(e^{-\theta T}\int_0^T e^{\theta s} dG_s\right)^2\right] \longrightarrow \sigma_G^2.$$

$(\mathcal{A}_4)$ For all fixed $s \geq 0$,

$$\lim_{T\to\infty} E\left(G_s e^{-\theta T}\int_0^T e^{\theta r} dG_r\right) = 0.$$

$(\mathcal{A}_5)$ For all fixed $s \geq 0$,

$$\lim_{T\to\infty} \frac{E\left(G_s G_T\right)}{T^{\eta}} = 0, \quad \lim_{T\to\infty} E\left(\frac{G_T}{T^{\eta}} e^{-\theta T}\int_0^T e^{\theta r} dG_r\right) = 0.$$

Recall that if $X \sim \mathcal{N}(m_1, \sigma_1)$ and $Y \sim \mathcal{N}(m_2, \sigma_2)$ are two independent random variables, then X/Y follows a Cauchy-type distribution. For a motivation and further references, we refer the reader to [12], as well as [11]. Notice also that if $N \sim \mathcal{N}(0,1)$ is independent of G, then N is independent of ζ_∞, since $\zeta_\infty = \theta \int_0^\infty e^{-\theta s} G_s ds$ is a functional of G.

Theorem 2.2 *Assume that* $(\mathcal{A}_1)$*–*$(\mathcal{A}_4)$ *hold. If* $N_1 \sim \mathcal{N}(0,1)$*,* $N_2 \sim \mathcal{N}(0,1)$ *and* G *are independent, then as* $T \to \infty$*,*

$$e^{\theta T}(\widetilde{\theta}_T - \theta) \xrightarrow{Law} \frac{2\theta\sigma_G N_2}{\mu + \zeta_\infty}, \tag{10}$$

$$T^{1-\eta}\left(\widetilde{\mu}_T - \mu\right) \xrightarrow{Law} \mathcal{N}\left(0, \frac{\lambda_G^2}{\theta^2}\right), \tag{11}$$

$$T^{1-\eta}\left(\widetilde{\alpha}_T - \alpha\right) \xrightarrow{Law} \mathcal{N}\left(0, \lambda_G^2\right). \tag{12}$$

Moreover, if $(\mathcal{A}_5)$ *holds, then as* $T \to \infty$*,*

$$\left(e^{\theta T}(\widetilde{\theta}_T - \theta), T^{1-\eta}\left(\widetilde{\mu}_T - \mu\right)\right) \xrightarrow{Law} \left(\frac{2\theta\sigma_G N_2}{\mu + \zeta_\infty}, \frac{\lambda_G}{\theta} N_1\right), \tag{13}$$

$$\left(e^{\theta T}(\widetilde{\theta}_T - \theta), T^{1-\eta}\left(\widetilde{\alpha}_T - \mu\right)\right) \xrightarrow{Law} \left(\frac{2\theta\sigma_G N_2}{\mu + \zeta_\infty}, \lambda_G N_1\right). \tag{14}$$

References

[1] Alazemi, F., Alsenafi, A., Es-Sebaiy, K. (2019). Parameter estimation for Gaussian mean-reverting Ornstein-Uhlenbeck processes of the second kind: non-ergodic case. Stochastics and Dynamics. In press, DOI: 10.1142/S0219493720500112

[2] Chronopoulou, A., Viens, F. (2012) Estimation and pricing under long-memory stochastic volatility. Annals of Finance 8, 379-403.

[3] Chronopoulou, A., Viens, F. (2012) Stochastic volatility and option pricing with long-memory in discrete and continuous time. Quantitative Finance 12, 635-649.

[4] Comte, F., Coutin, L., Renault, E. (2012) Affine fractional stochastic volatility models. Annals of Finance 8, 337-378.

[5] Comte, F., Renault, E. (1998). Long Memory in Continuous-time Stochastic Volatility Models. Mathematical Finance, 8(4):291-323.

[6] Douissi, S., Es-Sebaiy, K., Viens, F. (2019). Berry-Esséen bounds for parameter estimation of general Gaussian processes. ALEA, Lat. Am. J. Probab. Math. Stat. 16, 633-664.

[7] El Machkouri, M., Es-Sebaiy, K., Ouknine, Y. (2016). Least squares estimator for non-ergodic Ornstein-Uhlenbeck processes driven by Gaussian processes. Journal of the Korean Statistical Society 45, 329-341.

[8] El Onsy, B., Es-Sebaiy, K., Viens, F. (2017). Parameter Estimation for a partially observed Ornstein-Uhlenbeck process with long-memory noise. Stochastics, 89(2), 431-468.

[9] Es-Sebaiy, K., Viens, F. (2019). Optimal rates for parameter estimation of stationary Gaussian processes. Stochastic Processes and their Applications, 129(9), 3018-3054.

[10] Gatheral, J., Jaisson, T., Rosenbaum, M. (2018). Volatility Is Rough. Quantitative Finance, 18(6), 933-949.

[11] Marsaglia, G. (1965). Ratios of normal variables and ratios of sums of uniform variables. J. Amer. Statist. Asso. 60:193-204.

[12] Pham-Gia, T., Turkkan, N., Marchand, E. (2006). Density of the ratio of two normal random variables and applications. Communications in Statistics-Theory and Methods, 35(9), 1569-1591.

[13] Vasicek, O. (1977). An equilibrium characterization of the term structure. J. Finance Econ. 5(2): p. 177-188.

[14] Xiao, W., Zhang, X., Zuo, Y. (2018). Least squares estimation for the drift parameters in the sub-fractional Vasicek processes. Journal of Statistical Planning and Inference, Vol. 197, Pages 141-155.

Formulation Probabiliste des Moindre Carrés Partiels

Lola Etiévant[1], Vivian Viallon[2]

[1] *Univ Lyon, Université Claude Bernard Lyon 1, CNRS UMR 5208, Institut Camille Jordan, 43 boulevard du 11 novembre 1918, F-69622 Villeurbanne, France.*
lola.etievant@univ-lyon1.fr
[2] *International Agency for Research on Cancer, Nutritional Methodology and Biostatistics Group, Lyon, France. ViallonV@iarc.fr*

Résumé. Les Moindre Carrés Partiels (PLS, pour Partial Least Squares) réfèrent à une famille de méthodes de réduction de dimension, qui identifient deux ensembles de composantes de covariance maximale, afin de résumer les relations existant entre deux ensembles de variables observées $x \in \mathbb{R}^p$ et $y \in \mathbb{R}^q$, où $p \geq 1, q \geq 1$. el Bouhaddani et al. (2018) ont récemment proposé une formulation probabiliste de la PLS. Sous les contraintes qu'ils considèrent pour les paramètres de leur modèle, celui-ci peut-être vu comme une formulation probabiliste de la version PLS-SVD de la PLS. Cependant, nous établissons que ces contraintes sont trop restrictives, au sens où elles limitent le modèle à un ensemble de lois du couple (x, y) pour lesquelles les composantes de covariance maximale (solutions de la PLS-SVD) sont aussi respectivement de variance maximale (solutions de deux Analyses en Composantes Principales appliquées séparément sur chacun des deux ensembles de variables x et y). Nous proposons alors une extension du modèle, pour obtenir une version probabiliste plus générale de la PLS-SVD, qui n'est pas limitée à ces seules lois. Nous présentons des résultats de simulation numérique qui illustrent notamment les limites du modèle proposé par el Bouhaddani et al.

Mots-clés. Moindre carrés partiels, PLS, formulation probabiliste, identifiabilité.

Abstract. Partial Least Squares (PLS) refer to a class of dimension-reduction techniques relying on the identification of two sets of components with maximal covariance, in order to model the relationship between two sets of observed variables $x \in \mathbb{R}^p$ and $y \in \mathbb{R}^q$, with $p \geq 1, q \geq 1$. el Bouhaddani et al. (2018) have recently proposed a probabilistic formulation of PLS. Under the constraints they consider for the parameters of their model, their model can be seen as a probabilistic formulation of PLS-SVD. However, we establish that these constraints are too restrictive as they define a set of distributions of (x, y), for which the components with maximal covariance (solutions of the PLS-SVD), are also necessarily of respective maximal variances (solutions of the Principal Components Analyses of x and y, respectively) Then, we propose an extension of the model, which corresponds to a more general probabilistic formulation of PLS-SVD, and which is no longer restricted to these sole distributions. We present numerical examples, which especially illustrate the limitations of the model proposed by el Bouhaddani et al.

Keywords. Partial least squares, PLS, probabilistic formulation, identifiability.

1 Introduction

Les Moindre Carrés Partiels (PLS, pour Partial Least Squares) font généralement référence à une famille de méthodes de réduction de dimension, qui identifient deux ensembles de composantes de covariance maximale, afin de résumer les relations existant entre deux ensembles de variables observées $x \in \mathbb{R}^p$ et $y \in \mathbb{R}^q$, où $p \geq 1, q \geq 1$. Les versions les plus communes incluent la PLS Régression, et des versions symétriques, telles que la PLS-W2A ou encore la PLS-SVD (Jöreskog and Wold, 1982, Rosipal and Krämer, 2006, Wegelin, 2000, Wold, 1985). La PLS-SVD repose sur la Décomposition en Valeurs Singulière (SVD, pour Singular Value Decomposition) de la matrice de covariance de x et y, et les poids (qui permettront la construction des composantes) sont les vecteurs singuliers de cette matrice de covariance. Les PLS Régression et PLS-W2A fonctionnent quant à elles de manière itérative, appliquant à chaque étape une technique de déflation qui garantit aux composantes des propriétés d'orthogonalité.

Des versions probabilistes des méthodes de réduction de dimension ont été proposées, notamment pour l'Analyse en Composantes Principales (PCA, pour Principal Component Analysis) (Tipping and Bishop, 1999), l'Analyse Canonique des Corrélations (CCA, pour Canonical Correlation Analysis) (Bach and Jordan, 2005), et la PLS (el Bouhaddani et al., 2018, Li et al., 2015, Zheng et al., 2016). Ces versions probabilistes reposent sur des équations structurelles, selon lesquelles les variables observées sont « générées » à partir de combinaisons linéaires de variables latentes, auxquelles s'ajoute un bruit Gaussien. L'estimation des paramètres dans ces modèles à variables latentes est généralement opérée via des algorithmes de type EM (Expectation-Maximization) (Dempster et al., 1977). Ces formulations probabilistes présentent plusieurs intérêts par rapport aux formalisations classiques, comme par exemple la prise en compte d'éventuelles données manquantes (Tipping and Bishop, 1999). D'autre part, elles ouvrent naturellement la voie aux méthodes d'estimation reposant sur des versions pénalisées de la vraisemblance, par exemple pour encourager la parcimonie de certains paramètres (Guan and Dy, 2009, Park et al., 2017).

Ici, nous nous intéressons au modèle PPLS (Probabilistic PLS) proposé récemment par el Bouhaddani et al. (2018). Notons I_p la matrice identité de taille $p \times p$, et 0_p le vecteur $(0, \ldots, 0)$ de taille p. Le modèle est défini par les trois équations suivantes

$$x = tW^\top + e, \quad y = uC^\top + f, \quad u = tB + h, \tag{1}$$

sous les contraintes :

(a) $t \sim \mathcal{N}(0, \Sigma_T)$, (b) Σ_T est une matrice diagonale, (c) $e \sim \mathcal{N}(0_p, \sigma_e^2 I_p)$,

(d) $f \sim \mathcal{N}(0_q, \sigma_f^2 I_q)$, (e) $h \sim \mathcal{N}(0_r, \sigma_h^2 I_r)$, (f) $r < \min(p, q)$

(g) W et C sont des matrices semi-orthogonales, de tailles $p \times r$ et $q \times r$ respectivement,

(h) B est diagonale de taille $r \times r$ (à éléments diagonaux strictement positifs),

(i) Les éléments diagonaux de $\Sigma_T B$ sont ordonnés de manière strictement décroissante.

Les paramètres du modèle PPLS sont $W = (W_1, \cdots, W_r)$, $C = (C_1, \cdots, C_r)$, B, Σ_T, σ_e^2, σ_f^2 et σ_h^2. En particulier, W et C correspondent aux matrices de poids. Les deux premières équations en (1) décrivent comment les variables observées x et y sont générées à partir des variables latentes t et u ; la troisième équation décrit quant à elle l'association entre ces deux variables latentes. Sous ce modèle, nous avons en particulier $\mathrm{Cov}(x, y) = W\Sigma_T BC^\top$. el Bouhaddani et al. établissent alors que W et C correspondent aux vecteurs singuliers (à gauche et à droite, respectivement) associés aux r plus grandes valeurs singulières de la matrice $\mathrm{Cov}(x, y)$. Bien qu'ils n'en fassent pas mention, ce résultat suggère que leur modèle correspond à une formulation probabiliste de la PLS-SVD. Dans la section suivante, nous montrons cependant que les contraintes (c) et (d) d'isotropie des variances des termes d'erreurs Gaussiens, sont telles qu'elles restreignent le modèle PPLS à un ensemble de lois « triviales », sous lesquelles les poids de la PLS-SVD définissent deux ensembles de composantes qui sont certes de covariance maximale, mais aussi respectivement de variance maximale.

2 Limites et extension du modèle PPLS

Sous les contraintes (a)-(i), el Bouhaddani et al. (2018) établissent l'identifiabilité des paramètres de leur modèle PPLS ; en particulier les matrices de poids W et C sont identifiables, au signe près, et correspondent aux r premiers vecteurs singuliers de la matrice $\mathrm{Cov}(x, y)$. Cependant, les contraintes (c) et (d) d'isotropie des variances des termes d'erreurs Gaussiens assurent d'une part, que $\mathrm{Var}(x) = W\Sigma_T W^\top + \sigma_e^2 I_p$, et d'autre part, que $\mathrm{Var}(y) = C(\Sigma_T B^2 + \sigma_h^e I_r)C^\top + \sigma_f^2 I_q$. Il en découle que les colonnes de W et C sont également vecteurs propres associées aux r plus grandes valeurs propres des matrices $\mathrm{Var}(x)$ et $\mathrm{Var}(y)$, respectivement. Autrement dit, le modèle PPLS définit un ensemble de lois du couple (x, y) pour lesquelles les deux ensembles de composantes de covariance maximale, obtenues par PLS-SVD sur (x, y), sont aussi de variance maximale, respectivement. En particulier, lorsque les r plus grandes valeurs propres des matrices $\mathrm{Var}(x)$ et $\mathrm{Var}(y)$ sont de multiplicité algébrique égale à un (c'est-à-dire lorsque les éléments diagonaux de Σ_T et $\Sigma_T B^2$, respectivement, sont distincts), les vecteurs propres associés sont définis de manière unique (au signe près): pour ces lois, les lois marginales de x et y suffisent donc pour l'identification des r couples de composantes de covariance maximale. Et de fait dans ce cas, deux PCAs, appliquées séparément sur x et y, permettent d'identifier ces composantes. Pour des lois du couple (x, y) qui ne vérifient pas cette propriété, le modèle PPLS est nécessairement mal spécifié, et il n'est alors pas garanti que les estimations des matrices de poids W et C retournées par l'algorithme EM proposé par el Bouhadanni et al. aient un rapport quelconque avec les poids de la PLS-SVD (ce que confirment nos résultats en Section 3).

Nous proposons alors une extension du modèle PPLS, qui peut être vue comme une formulation probabiliste de la PLS-SVD pour une famille de lois plus générales que celle

couverte par le modèle PPLS de el Bouhaddani et al. (2018). À cet effet, nous assouplissons dans un premier temps les contraintes (c) et (d) d'isotropie des variances des termes d'erreurs Gaussiens, afin d'obtenir un modèle similaire au modèle PPLS, mais pour lequel les paramètres ne peuvent en général pas être identifiés en utilisant uniquement les lois marginales de x et de y. En particulier, pour proposer un modèle aussi général que possible, nous supposons simplement que les bruits e et f sont de matrices de variance quelconques (semi-définies positives). Nous proposons donc de remplacer les contraintes (c) et (d) du modèle PPLS par les conditions (c*) et (d*) :

(c*) $e \sim \mathcal{N}(0_p, \Psi_e)$, avec Ψ_e une matrice semi-définie positive de taille $p \times p$,

(d*) $f \sim \mathcal{N}(0_q, \Psi_f)$, avec Ψ_f une matrice semi-définie positive de taille $q \times q$,

Or, pour garantir l'identifiabilité de notre modèle, nous ne pouvons conserver les deux ensembles de variables latentes t et u, et devons considérer une écriture avec un seul ensemble de variables latentes (comme c'est aussi le cas dans la version probabiliste de la CCA proposée par Bach and Jordan (2005)). Nous proposons donc finalement le modèle suivant :

$$x = tW^\top + e, \quad y = tC^\top + f, \tag{2}$$

sous les contraintes (a), (b), (c*), (d*), (f), (g) et (i*), où (i*) est l'analogue de (i) dans le cas où un unique ensemble de variables latentes est considéré :

(i*) Les éléments diagonaux de Σ_T sont ordonnés de manière strictement décroissant.

Nous établissons notamment que les matrices de poids W et C sont identifiables (au signe près), correspondent aux vecteurs singuliers de la matrice $\mathrm{Cov}(x, y)$, et ne peuvent en général pas être obtenues à partir des seules lois marginales de x et y.

3 Illustrations numériques

Nous considérons deux schémas de simulations numériques pour illustrer nos résultats, et en particulier pour illustrer le comportement des estimations de W et C obtenues avec l'algorithme EM conçu par el Bouhaddani et al. (2018) pour leur modèle PPLS. Nous nous plaçons dans le cas où $p = q = 20$ pour les dimensions des variables observées, et où $r = 3$ pour le nombre de variables latentes. Nous faisons varier le nombre d'observations $n \in \{50, 250, 500, 1000, 5000\}$. Pour le premier schéma, nous générons des données sous le modèle de el Bouhaddani et al., en utilisant les mêmes paramètres que ceux choisis pour leurs simulations; en particulier, les éléments de diagonaux de Σ_T et $\Sigma_T B^2$ sont respectivement distincts. Pour le second schéma, nous générons des données sous un modèle similaire, mais avec des termes d'erreur e et f de matrices de variances semi-définies positives quelconques. La colonne gauche de la Figure 1 présente les résultats de comparaisons entre les vraies matrices de poids et celles obtenues par (i) l'algorithme EM de la PPLS, (ii) deux PCAs distinctes sur x et y, (iii) PLS-SVD sur (x, y), et (iv) PLS-W2A sur (x, y). La colonne droite de la Figure 1 présente les résultats de comparaisons entre les matrices de poids obtenues par l'algorithme EM de la PPLS et celles obtenues par

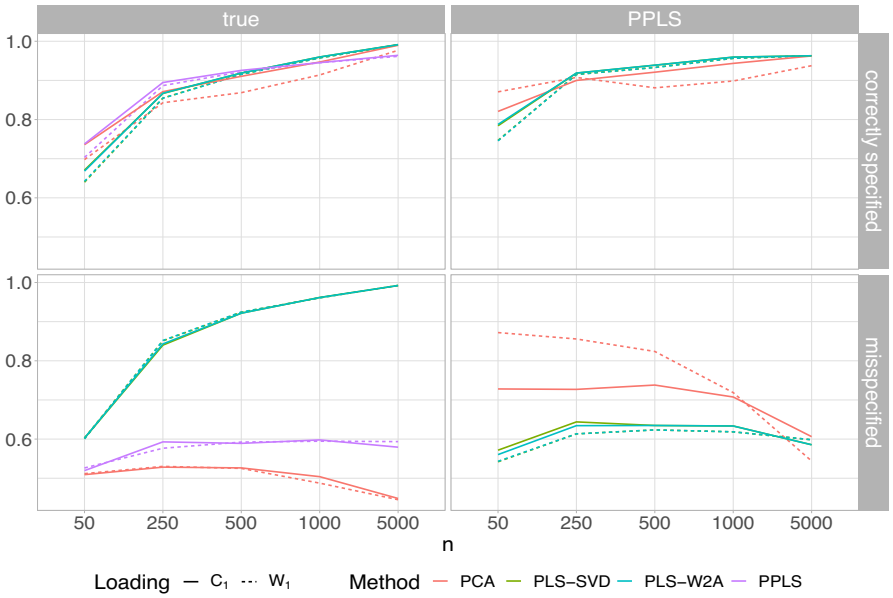


Figure 1: Produit scalaire médian (en valeur absolue) : (*gauche*) entre la première colonne des vraies matrices de poids W_1 (et C_1) et celles obtenues par l'algorithme EM du modèle PPLS, par deux PCAs distinctes sur x et y, par PLS-SVD sur x et y, et par PLS-W2A sur x et y. (*droite*) entre la première colonne des matrices de poids obtenues par l'algorithme EM du modèle PPLS et celles obtenues par deux PCAs distinctes sur x et y, par PLS-SVD sur x et y, et par PLS-W2A sur x et y. Les résultats calculés sur 1000 simulations, pour $p = q = 20$, $r = 3$ et $n \in \{50, 250, 500, 1000, 5000\}$. Ils sont présentés sur la première ligne lorsque le modèle PPLS est correctement spécifié (données générées sous le modèle rappelé en Équation (1)), et sur la seconde ligne lorsqu'il ne l'est pas (données générées sous un modèle similaire mais avec e et f qui ne sont pas de variances isotropes).

(*i*) par deux PCAs distinctes sur x et y, (*ii*) PLS-SVD sur x et y, et (*iii*) PLS-W2A sur x et y. Ces résultats sont moyennés sur 1000 simulations sous chacun des deux schémas. Nous présentons ici les résultats pour les premières colonnes des poids (W_1 et C_1), les résultats pour les $r - 1$ autres colonnes étant similaires. Lorsque le modèle PPLS est correctement spécifié (première ligne), les méthodes PLS (PPLS, PLS-SVD et PLS-W2A) identifient les vraies matrices de poids W et C, mais deux simples PCAs sur x et y permettent de les identifier également. Lorsque le modèle PPLS est mal spécifié, l'algorithme EM de la PPLS ne permet plus d'identifier les vraies matrices de poids, pas plus que les deux simples PCAs, contrairement par exemple à l'approche PLS-SVD (graphe en bas à gauche). D'un autre côté le graphe en bas à droite indique que les poids retournés par l'EM de la PPLS sont généralement plus proches de ceux des deux PCAs que de ceux de la PLS-SVD ou de la PLS-W2A. De fait, lorsqu'il est appliqué à des données réelles et que le modèle PPLS

est mal spécifié, l'algorithme EM de la PPLS n'est en général pas capable de capturer la part commune à x et y.

Bibliographie

Bach, F. and Jordan, M. (2005). A probabilistic interpretation of canonical correlation analysis.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society. Series B (Methodological), 39(1):1–38.

el Bouhaddani, S., Uh, H.-W., Hayward, C., Jongbloed, G., and Houwing, J. (2018). Probabilistic partial least squares model: identifiability, estimation and application. Journal of Multivariate Analysis, 167.

Guan, Y. and Dy, J. (2009). Sparse probabilistic principal component analysis. Journal of Machine Learning Research - Proceedings Track, 5:185–192.

Jöreskog, K. and Wold, H. (1982). Soft modeling: The basic design and some extensions. In Joreskog, K. and Wold, H., editors, Systems under indirect observation. Causality, structure, prediction. Part II, pages 1–54.

Li, S., Nyagilo, J., Dave, D., Wang, W., Zhang, B., and Gao, J. (2015). Probabilistic partial least squares regression for quantitative analysis of raman spectra. International Journal of Data Mining and Bioinformatics, 11:223.

Park, C., Wang, M., and Mo, E. (2017). Probabilistic penalized principal component analysis. Communications for Statistical Applications and Methods, 24:143–154.

Rosipal, R. and Krämer, N. (2006). Overview and recent advances in partial least squares. In Saunders, C., Grobelnik, M., Gunn, S., and Shawe-Taylor, J., editors, Subspace, Latent Structure and Feature Selection, pages 34–51. Springer Berlin Heidelberg.

Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 61(3):611–622.

Wegelin, J. A. (2000). A survey of partial least squares (pls) methods, with emphasis on the two-block case. Technical report, Univ. of Washington, Dept. of Statistics.

Wold, H. (1985). Partial least squares. In Kotz, S. and Johnson, N., editors, Encyclopedia of Statistical Sciences. Volume 6, pages 581–591.

Zheng, J., Song, Z., and Ge, Z. (2016). Probabilistic learning of partial least squares regression model: Theory and industrial applications. Chemometrics and Intelligent Laboratory Systems, 158:80 – 90.

Bayesian Inference on Local Distributions of Multi-dimensional Curves

Anis Fradi [1,2] & Chafik Samir [1,3]

[1] *CNRS-LIMOS (UMR 6158), University Clermont Auvergne, France*
[2] *MAPSFA (LR11ES35), Faculty of Sciences of Monastir, Tunisia*
[3] *Institut de Mathématiques de Toulouse, France*
anis.fradi@etu.uca.fr & chafik.samir@uca.fr

Résumé. Nous introduisons un nouveau modèle statistique pour l'analyse des courbes multidimensionnelles définies sur $I = [0, 1]$. Nous nous intéressons à l'inférence bayésienne de regroupement de courbes modélisées pour des processus gaussiens en sous-populations homogènes. En pratique, nous nous limitons à modéliser les courbes avec des lois gaussiennes multivariées car nous ne pouvons observer que leurs discrétisations finies. Pour atteindre un tel objectif, nous utilisons la géométrie statistique des fonctions de reparamétrisation identifiées avec les distributions locales des courbes. Afin d'explorer l'inférence sur toutes les fonctions de reparamétrisation, nous faisons le lien avec la sphère de Hilbert. Cela nous permet de réduire la complexité de l'espace de reparamétrisations et facilite la l'optimisation. Nous donnons également la loi à postériori des reparamétrisations et discutons de certaines propriétés asymptotiques. Enfin, l'intérêt pratique de la méthode proposée est illustré par une application potentielle sur les courbes multidimensionnelles. Une direction future d'intérêt est de construire une extension théorique pour des domaines plus complexes avec de nouveaux aspects de l'apprentissage et l'inférence sur des variétés de grande dimension.

Mots-clés. Inférence bayésienne, processus gaussien, courbes multidimensionnelles, sphère de Hilbert

Abstract. We introduce a new statistical model for analyzing multi-dimensional curves defined on $I = [0, 1]$. We are interested in the Bayesian inference of grouping parameterized curves modeled with Gaussian processes into homogeneous sub-populations. In practice, we restrict ourselves to model the curves with multivariate Gaussians since we can only observe their finite discretizations. To reach such goal, we use the statistical geometry of reparametrization functions identified with local distributions on curves. In order to explore the inference on all reparametrization functions, we make the connection with the Hilbert sphere. This allows us to reduce the complexity of the space of reparametrizations and facilitates the optimization task. We also give the posterior on these reparametrizations and discuss some asymptotic properties. Finally, the practical interest of the proposed method is illustrated with a potential application on multi-dimensional curves. A future direction of interest is to build theoretical extension for more complex domains for new aspects of manifold learning and inference on high-dimensional manifolds.

Keywords. Bayesian inference, Gaussian process, multidimensional shapes of curves, Hilbert sphere

1 Introduction

In the case of clustering multidimensional curves [1] with values in $\mathbb{R}^d$ for $d \geq 2$, landmarks are not always available and continuous representations are successful in this respect [2]. Unlike other approaches, the distance between two curves must take into account their reparameterizations. The reader can refer to [3] for survey on shapes of curves as elements of Riemannian manifolds [4]. Recently, [5] have introduced a novel representation which include smooth curves in Euclidean spaces. In this work, we adapt this representation to our context due to its several advantages and we note q_i instead of the original observed curves β_i, $i = 1, ..., N$.

The proposed method consist in generalizing the clustering of multidimensional curves with a Bayesian inference on reparametrizations [6]. Since we can not observe all q_i directly, we note its discretized version $q_i(t_1), ..., q_i(t_m)$, supposed to be drawn from multivariate Gaussians for the likelihood term. If we assume that we have K classes of curves, the posterior probability on reparametrizations $\gamma^1, ..., \gamma^K$ becomes more affordable on the resulting square-root densities $\psi^1, ..., \psi^K$ belonging to the Hilbert sphere $\mathcal{H}$ [7], which can be performed using a Hamiltonian Monte Carlo sampling. To our knowledge, this is the first attempt to model the population background from nonparametric distributions.

2 Statistical geometry of probability distributions

Before we give details of our method for curve clustering we will introduce some notions of of probability distributions. We define the space of cumulative density distributions, identified with reparametrizations space for curves, by

$$\Gamma = \Big\{ \gamma : I \to I \ \mid \ \gamma(0) = 0, \ \ \gamma(1) = 1, \ \text{ and } \ \dot{\gamma} \ \text{ is nonnegative} \Big\} \tag{1}$$

In this paper, we assume that each class of curves results from its own local distributions. So the optimization in reparametrizations space is well needed for the clustering task. Due to the complexity of this space, we make the connection with the Hilbert sphere: If we consider the transformation $\psi(t) = \sqrt{\dot{\gamma}(t)}$ for all $t \in I$ then the space of functions ψ becomes

$$\mathcal{H} = \Big\{ \psi \in \mathbb{L}^2(I) \ \mid \ \psi \text{ is nonnegative, \ and } \ ||\psi||_{\mathbb{L}^2} = \Big(\int_I \psi^2(t) \mathrm{d}t \Big)^{1/2} = 1 \Big\} \tag{2}$$

equipped with the Fisher-Rao metric

$$\big\langle g_1, g_2 \big\rangle_{\mathbb{L}^2} = \int_I g_1(t) g_2(t) \mathrm{d}t \tag{3}$$

for $g_1, g_2 \in \mathcal{T}_\psi(\mathcal{H})$. Since $\mathcal{H}$ is identified with the Hilbert upper-hemisphere, each ψ defines a unique reparametrization (isometrically) γ, satisfying

$$\gamma(t) = \int_0^t \psi(s)^2 \mathrm{d}s, \quad \forall t \in I \tag{4}$$

The advantage of working in $\mathcal{H}$ rather than Γ is that $\mathcal{H}$ is endowed with more familiar geometric features making the optimization more efficient. For example, the overlap between $\psi_1, \psi_2 \in \mathcal{H}$ is given by the angle (arccos) of the shortest arc (geodesic) between them on $\mathcal{H}$, i.e., $d_{\mathcal{H}}(\psi_1, \psi_2) = \cos^{-1}\left(\left\langle \psi_1, \psi_2 \right\rangle_{\mathbb{L}^2}\right)$.

3 Bayesian clustering model

Let β be a smooth curve defined on I with values in $\mathbb{R}^d$, i.e., $\beta : I \to \mathbb{R}^d$ and having first derivatives in $\mathbb{L}^2\big(I, \mathbb{R}^d\big)$. We represent each curve using its q-function defined by

$$\begin{array}{rcl} q : I & \to & \mathbb{R}^d \\ t & \mapsto & q(t) = \dot{\beta}(t)||\dot{\beta}(t)||_2^{-1/2} \quad \text{if } \dot{\beta}(t) \neq 0, 0 \text{ otherwise} \end{array} \tag{5}$$

We note that q is well defined and is an element of $\mathbb{L}^2\big(I, \mathbb{R}^d\big)$. If the domain of β is transformed with a reparametrization $\gamma \in \Gamma$ then $\beta \circ \gamma$ is represented by q^* satisfying $q^*(t) = \sqrt{\dot{\gamma}(t)}q(\gamma(t))$. The advantage of this representation is that it is invariant to translations, reparametrizations and takes into account the uniform scaling. In addition, the Fréchet mean of $\{q_i^*\}_{i=1}^N$ is

$$\tilde{q} = \underset{q \in \mathbb{L}^2(I, \mathbb{R}^d)}{\operatorname{argmin}} \sum_{i=1}^{N} \inf_{\gamma_i \in \Gamma} ||q - q_i^*||^2 \text{ where } ||q|| = \int_I \Big(||q(t)||_2^2\Big)^{1/2} \mathrm{d}t \tag{6}$$

In this section, we will reformulate our problem of clustering $q_1^*, ..., q_N^*$ into K populations. For each q_i^*, we draw a class $C_i = k$ with values in $\{1, ..., K\}$ under the probability $p(C_i = k) = \pi_k$. Consequently, the probability that q_i^* belongs to the class k is $p(q_i^*|C_i = k)$. In practice, we observe a discretization $q_i^*(t_h) \in \mathbb{R}^d$, $h = 1, ..., m$ and we note $T = (t_1, ..., t_m)$. We assume that $q_i^*(T)|C_i = k \sim \mathcal{N}\big(\tilde{q}^k(T), \sigma^2\mathcal{I}\big)$ where $\sigma^2 > 0$ is the variance and $\mathcal{I}$ is the $md \times md$ identity matrix. This work deals with estimating the optimal reparametrization for the k-th class denoted by γ^k. The probability that q_i^* belongs to k-th class is

$$p(q_i|\gamma^k, \tilde{q}^k(T), \sigma^2) \propto \exp\big(-\frac{1}{2\sigma^2}||q_i^*(T) - \tilde{q}^k(T)||_2^2\big) \tag{7}$$

Consequently, if we note $D = \{q_i\}_{i=1}^N$, the log-likelihood term is proportional to

$$\begin{array}{rcl} \log p(D|\gamma^1, ..., \gamma^K, \pi_1, ..., \pi_K, \tilde{q}^1(T), ..., \tilde{q}^K(T), \sigma^2) & \propto & \displaystyle\sum_{i=1}^{N} \log\Big(\sum_{k=1}^{K} \pi_k \exp\big(-\frac{1}{2\sigma^2}||q_i^*(T) \\ & & -\tilde{q}^k(T)||_2^2\big)\Big) \end{array} \tag{8}$$

For the prior laws of γ^ks, we simply use the term

$$p(\gamma^k) \propto \exp(-d_\Gamma(\gamma^k, \mathrm{id})^2) \tag{9}$$

where $d_\Gamma(.,.)$ is the geodesic distance defined on Γ and id is the identity function, i.e., $\mathrm{id}(t) = t$ for all $t \in I$. This means that γ^k is modeled with a Gaussian prior law: $\gamma^k \sim \mathcal{N}(\mathrm{id}, 1)$ of mean $\gamma \equiv \mathrm{id}$

and standard variance 1. Since there is no explicit expression of $d_\Gamma(.,.)$, we establish the isometric bijective map between Γ and $\mathcal{H}$ from Eq.(4), i.e., $d_\Gamma(\gamma_1, \gamma_2) = d_\mathcal{H}(\psi_1, \psi_2)$. Consequently, the prior on γ^k becomes a more simple prior on ψ^k, satisfying

$$p(\psi^k) \propto \exp\big(-d_\mathcal{H}(\psi^k, 1)^2\big) = \exp\big(-\cos^{-1}\big(\langle\psi^k, 1\rangle_{\mathbb{L}^2}\big)^2\big) \tag{10}$$

since $\sqrt{\dot{t}} = 1$. Moreover, we assume that all $\psi^k s$ are independent, implying

$$p(\psi^1, ..., \psi^K) \propto \exp\big(-\sum_{k=1}^{K} \cos^{-1}\big(\langle\psi^k, 1\rangle_{\mathbb{L}^2}\big)^2\big) \tag{11}$$

Given $D, \pi_1, ..., \pi_K, \tilde{q}^1(T), ...\tilde{q}^K(T)$ and σ^2, the log-posterior of $\psi^1, ..., \psi^K$ is then

$$\begin{aligned}\log p(\psi^1, ..., \psi^K | D, \pi_1, ..., \pi_K, \tilde{q}^1(T), ..., \tilde{q}^K(T), \sigma^2) \quad \propto \quad & \sum_{i=1}^{N} \log\Big(\sum_{k=1}^{K} \pi_k \exp\big(-\frac{1}{2\sigma^2}||q_i^*(T) \\ & -\tilde{q}^k(T)||_2^2\big)\Big) - \sum_{k=1}^{K} \cos^{-1}\big(\langle\psi^k, 1\rangle_{\mathbb{L}^2}\big)^2\end{aligned} \tag{12}$$

4 Application

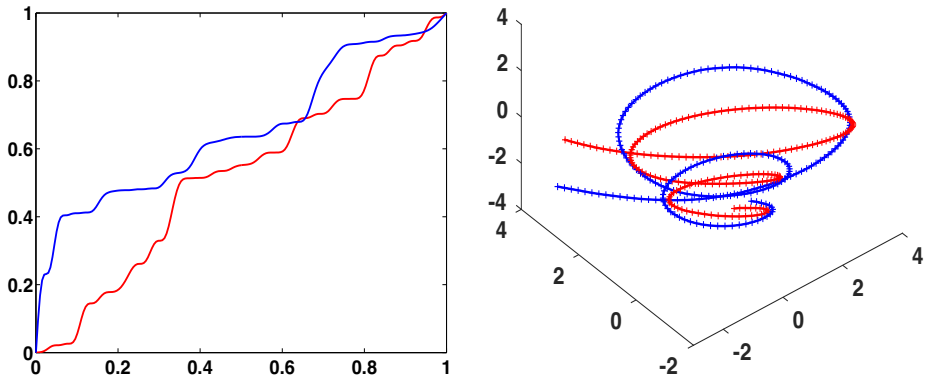
We validate the proposed method on a dataset of 3D curves used to classify humans according to their gender [8]. We compare the proposed method against: GMM, Euclidean kmeans, geodesic kmeans and geodesic kmedoids. All performance rates are reported in Table 1. Accordingly, we point out that the proposed method performs much better than the other baseline methods with a great margin. For more details, Figure 1 (a) illustrates the optimal reparametrization for k-th class γ^k when estimating the corresponding square-root densities ψ^k. In Figure 1 (b), we plot the Fréchet mean $\tilde{q}^k$ of observed cochlea for both classes.

5 Conclusion

In this paper, we have introduced a new Bayesian clustering for multidimensional curves. We have considered that each class has its own unknown local distribution and we have formulated the problem to estimate them jointly. The proposed method is nonparametric in the sense that we do not assume that the prior belongs to any fixed class of distributions. We have tested our model on a real dataset in comparison which some current existing methods. We showed several benefits and better accuracy when estimating the optimal reparametrization for each class and dealing with the Hilbert sphere.

Table 1: Accuracy rates (AR), specificities (SP) and sensibilities (SE) for cochlea.

Methods	AR	SP	SE
GMM	58.3%	58.07%	58.5%
Euclidean kmeans	58.5%	58.44%	58.5%
Geodesic kmeans	74.9%	74.4%	75.45%
Geodesic kmedoids	89.2%	89.8%	88.41%
Proposed	**95.8**%	**94**%	**97.73**%



(a) (b)

Figure 1: The optimal reparametrizations (a) and the Fréchet means of curves (b). Class of female in red and class of male in blue.

References

[1] C. Huang, M. Styner and H. Zhu. (2015), Clustering high-dimensional landmark-based two-dimensional shape data, *Journal of the American Statistical Association*, pp. 946–961.

[2] D. G. Kendall. (1984), Shape manifolds, procrustean metrics, and complex projective spaces, *Bulletin of the London Mathematical Society*.

[3] I. L. Dryden and K. V. Mardia. (2016), Statistical shape analysis, with applications in R, Second Edition, *John Wiley and Sons*.

[4] B. Abhishek and B. Rabi. (2012), Nonparametric inference on manifolds: with applications to shape spaces, *Cambridge University Press*.

[5] A. Srivastava, E. Klassen, S. H. Joshi and I. H. Jermyn. (2011), Shape analysis of elastic curves in Euclidean spaces, *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1415–1428.

[6] A. Srivastava and E. Klassen. (2016), Functional and shape data analysis, *Springer, New York, NY.*

[7] A. Berlinet and C. T. Agnan. (2003), Reproducing kernel Hilbert spaces in probability and statistics, *Springer US.*

[8] M. Vater and K. Manfred. (2011), Comparative aspects of cochlear functional organization in mammals, *Hearing research*, pp. 89-99.

Gradient Boosting adapté à la regression des paramètres d'une loi Pareto généralisée

Sébastien Farkas [1]

[1] *Sorbonne Université, CNRS, Laboratoire de Probabilités, Statistique et Modélisation, LPSM, 4 place Jussieu, F-75005 Paris, France. E-mail: sebastien.farkas@sorbonne-universite.fr*

Résumé. Dans un contexte de valeurs extrêmes, nous proposons d'améliorer les performances d'une méthodologie déjà proposée de régression des coefficients d'une loi Pareto généralisée par arbre de régression. Les méthodologies de forêt aléatoire et de gradient boosting ont en effet permis de construire des estimateurs plus robuste que ceux dérivés des arbres de régression. Nous adaptons ces méthodologies au contexte de la régression des coefficients d'une loi Pareto généralisée.

Mots-clés. Théorie des Valeurs Extrêmes, Forêt aléatoire, Gradient Boosting ...

Abstract. We propose to improve the performance of an existing proposed methodology of regression of the coefficients of a generalized Pareto distribution by regression tree. Random forest and gradient boosting are methodologies that allowed the construction of more robust estimators than those derived from regression trees. We adapt these approachs to the context of the regression of coefficients of a generalized Pareto distribution.

Keywords. Extreme Value Theory, Random Forest, Gradient Boosting ...

1 Introduction

1.1 Données extrêmes, loi de Pareto généralisée et régression

La Théorie des Valeurs Extrêmes (TVE) permet, notamment, de répondre à une interrogation qui présente, dans certains domaines, un intérêt certain : à partir d'un échantillon i.i.d $(Z_i)_{i\in 1,\dots,N}$, comment quantifier la probabilité que Z_1 soit supérieur à q, avec q relativement proche, voir supérieur au maximum sur l'échantillon ?

Une des approches de la Théorie des Valeurs Extrêmes consiste en l'étude des données en excès par rapport à un seuil u. En effet, une lecture informelle du théorème démontré par Bakema et De Haan (1974) est que la fonction de survie des excès $\overline{F}_u(z) = P[Z_1 - u > z \mid Z_1 > u]$ peut s'approcher, sous une hypothèse sur $\overline{F}$ dîte de variation régulière et lorsque $u \to +\infty$, par une fonction de survie 1 définissant une loi Pareto généralisée dont le paramètre σ est dît d'échelle et le paramètre γ est dît de forme. En pratique, l'hypothèse

de variation régulière est supposée vérifiée et u est choisi à la fois suffisamment élevé afin de justifier l'utilisation la précédente approximition, et dans le même temps suffisament faible pour obtenir un nombre d'excès n, parmi les N données, permettant l'adéquation d'une loi Pareto généralisée (GPD). L'ajustement d'une GPD peut alors se faire sur les excès notés $(Y_i)_{i\in 1,\dots,n}$ et il est alors possible de répondre à la problématique initialement soulevée, et également, par exemple, de calculer des quantiles extrêmes.

$$\overline{H}_{\sigma,\gamma}(y) = \left(1+\gamma\frac{y}{\sigma}\right)^{-1/\gamma},\ y>0 \tag{1}$$

L'observation des données $(Y_i)_{i\in 1,\dots,n}$ peut, dans certains cas de figure, être conjointe à celle de covariables $(X_i)_{i\in 1,\dots,n}$, où chaque X_i est élément de $\mathcal{X}\subset\mathbb{R}^d$. Il devient alors intéressant d'étudier le comportement de la queue de distribution de $Y \mid X$. En écrivant, $(\sigma(\mathbf{x}),\gamma(\mathbf{x})) = m(\mathbf{x})$, un problème de régression apparaît alors, consitant à déterminer la fonction f. Plusieurs articles, dont Beirlant et al. (2013), proposent des approches pour estimer f dans des cadres non paramétrique, paramétrique ou semi-paramétrique. Une alternative à ces méthodes est présentée plus explicitement ci-après.

1.2 Un arbre adapté aux valeurs extrêmes

Les arbres de régression permettent de construire un estimateur précis, interprétable et flexible aux structures non-linéaires. L'objectif est d'estimer la fonction de regression m^* définie par $m^* = \arg\min_{m\in\mathcal{M}} E[\phi(Y,m(\mathbf{X}))]$, selon une certaine fonction de perte ϕ. Leurs constructions se décomposent en deux étapes: d'abord la croissance puis l'élagage. La croissance consiste en un partitionnement récursif et binaire de $\mathcal{X}$. À l'étape k, l'estimateur s'écrit $\hat{m}^k(X) = \sum_{j=1}^k \hat{m}_j^k(X)\mathbf{1}_{X\in A_j}$, où $\{A_1,\dots,A_k\}$ est une partition de $\mathcal{X}$, et $\hat{m}_j^k(X) = \arg\min_{m\in\mathcal{M}}\sum_{i=1}^n \phi(Y_i,m(X_i))\mathbf{1}_{X_i\in A_j}$ minimise l'erreur empirique sur chaque A_j. La partition $\{A_1,\dots,A_k\}$ est alors modifiée en partitionnant A_{j^*} par (A_{j^*1},A_{j^*2}) de manière binaire (la coupe de A_{j^*} se fait suivant une unique coordonnée de $\mathcal{X}$) pour minimiser (selon j^* et la coupe de A_{j^*}) l'erreur empirique du nouvel estimateur (défini par la nouvelle partition). Chaque coupe crée ainsi un noeud dans l'arbre dont la structure résulte sur des feuilles: des groupes d'individus homogènes relativement à ϕ. Ensuite, l'élagage a pour objectif d'éviter le phénomène de surapprentissage en sélectionant, dans la suite des estimateurs $\hat{m}^k$, le meilleur au sens d'un critère pénalisant, au moyen d'un échantillon test ou d'une valisation croisée, la complexité de l'arbre, $\hat{m}$. Le choix de la fonction de perte est libre. Toutefois, la perte quadratique est la plus communément utilisée et l'estimateur $\hat{m}$ peut alors s'écrire $\hat{m} = \sum_{j=1}^k \beta_j\mathbf{1}_{X\in A_j}$, avec $\{\beta_1,\dots,\beta_k\}\in\mathbb{R}^k$.

Dans une récente pré publication de Farkas et al. (2020), la méthodologie d'arbre de régression est adaptée dans le cadre de l'étude des excès $(Y_i)_{i\in 1,\dots,n}$, en utilisant une fonction de perte égale à l'opposé de la vraissemblance des excès à une loi Pareto généralisée, $\phi(y,m(\mathbf{x})) = -\log(\sigma(\mathbf{x})) - \left(\frac{1}{\gamma(\mathbf{x})}+1\right)\log\left(1+\frac{y\gamma(\mathbf{x})}{\sigma(\mathbf{x})}\right)$. Nous ferons référence à cette méthode par l'acronyme GPRT (Generalized Pareto Regression Trees). Ainsi, lors

de la première phase de croissance de l'arbre, chaque noeud crée une nouvelle partition de $\mathcal{X}$ qui améliore l'adéquation des excès à une loi Pareto généralisée sur chacun de ses segments. La fonction de régression estimée $\hat{m}$ permet de caractériser les liens entre les covariables et le comportement en queue de distribution de Y. Toutefois, les estimateurs issus des arbres de régression sont sensibles aux données d'entraînement et des méthodes d'amélioration ont justement été développées pour proposer des estimateurs davantage robustes. Ainsi proposons-nous de les appliquer dans le contexte des valeurs extrêmes.

2 Amélioration de la performance de l'arbre par Gradient Boosting

L'implémentation de la méthode Extreme Gradient Boosting, introduit par Chen et al. (2016), a popularisé l'approche Gradient Boosting. Elle consiste également en la création d'une famille d'estimateurs, mais cette fois-ci ordonnée et dépendante. La construction de l'estimateur $\hat{m}_{k+1}$ dépend en effet de l'estimateur précédent $\hat{m}_k$, avec comme objectif d'en diminuer l'erreur empirique $L(\hat{m}_k) = \sum_{i=1}^{n} \phi(Y_i, \hat{m}(X_i))\mathbf{1}_{X_i \in A_j}$ et de construire, au fur et à mesure, une suite d'estimateurs convergent vers la solution optimale du problème de régression m^*. Un candidat idéal pourrait être obtenu un réalisant, à partir de $\hat{m}_k$, un (petit) pas w dans la direction opposé du gradient de l'erreur empirique $L(\hat{m}_k)$. Néanmoins, rien ne garanti l'appartenance de ce candidat à notre classe d'estimateurs $\mathcal{M}$. La stratégie réside alors dans la recherche du meilleur arbre de régression permettant d'approcher ce candidat, mais son implémentation classique repose sur l'hypothèse d'une prédiction réelle. Or, dans le cas des GPRT, l'estimateur obtenu est un couple. Nous devrons ainsi adapter l'implémentation pour permettre une fléxibilité supplémentaire. Des expériences par simulation permettront de tester la pertinence d'un tel algorithme.

Bibliographie

Athey, S., Tibshirani, J. and Wager, S. (2019). Generalized random forests. *The Annals of statistics*, 2, 1148–1178.

Balkema, A. A. and De Haan, L. (1974). Residual life time at great age. *The Annals of probability*, 792–804.

Beirlant, J., and Goegebeur, Y. (2003). Regression with response distributions of Pareto-type. *Computational statistics & data analysis*, 4, 595–619.

Biau, G., and Scornet, E. (2016). A random forest guided tour. *TEST*, 25, 197–227.

Chen, T. and Guestrin, C (2016). XGBoost: A Scalable Tree Boosting System. *KDD*, 785–794.

Farkas, S., Lopez, O. and Thomas, M. (2020). Cyber claim analysis through Generalized Pareto Regression Trees with applications to insurance. *preprint*.

Divergence Wasserstein par lots

Kilian Fatras [1] & Younes Zine [2] & Rémi Flamary [3] & Rémi Gribonval [2,4] & Nicolas Courty [1]

[1] *Univ Bretagne Sud, Inria, CNRS, IRISA, France et kilian.fatras@irisa.fr*
[2] *Univ Rennes, Inria, CNRS, IRISA, France et younes.zine@ens-rennes.fr*
[3] *Univ Côte d'Azur, OCA, UMR 7293, CNRS, Laboratoire Lagrange, France et remi.flamary@unice.fr*
[4] *Univ Lyon, Inria, CNRS, ENS de Lyon, UCB Lyon 1, LIP UMR 5668, F-69342, Lyon, France et remi.gribonval@inria.fr*
[1] *Univ Bretagne Sud, Inria, CNRS, IRISA, France et nicolas.courty@irisa.fr*

Résumé. Le transport optimal est un outil puissant afin de comparer des mesures entre elles et à cette fin, il a trouvé plusieurs applications en apprentissage automatique. Malheureusement, il souffre d'une complexité de calcul en $\mathcal{O}(n^3 log(n))$ ce qui empêche son utilisation directe dans un contexte de Big Data. Afin de diminuer ce coût de calcul, les praticiens optent pour une stratégie par sous-lots, *i.e.* en moyennant les résultats de sous problèmes. Malheureusement, cette stratégie induit un biais qui casse l'axiome de séparabilité dans la définition d'une distance. Nous proposons dans cet article une nouvelle fonction basée sur la distance de Wasserstein par lots, qui respecte l'axiome de séparabilité et qui possède un estimateur non biaisé. Ce résultat est appuyé par des simulations numériques empiriques. Enfin, nous mentionnons une question ouverte sur la positivité de notre nouvelle fonction.

Mots-clés. Distance de Wasserstein, Transport optimal, Divergence, Borne de concentrations, batch

Abstract. Optimal transport distances are powerful tools to compare probability distributions and have found many applications in machine learning. Yet they suffer from an $\mathcal{O}(n^3 log(n))$ computational complexity that prevents their direct use on large scale datasets. To overcome this computational challenge, practitioners estimate these distances on minibatches *i.e.,* they average the outcome of several smaller optimal transport problems. Unfortunately, the minibatch paradigm implies a bias which breaks the distance seperaty axiom. In this article, we propose a new cost function based on minibatch Optimal Transport distances, which respects the distance separability axiom and which has an unbiased estimator. We highlight several theoretical results supported by numerical experiments. Finally, we describe open questions on the positivity of our new loss function.

Keywords. Wasserstein distance, Optimal transport, Divergence, Concentration bounds, minibatch

1 Introduction

Le transport optimal connaît un fort succès pour certaines applications d'apprentissage automatique. Sa capacité à comparer des objets à travers un coût permet son utilisation pour des modèles génératifs [Arjovsky et al., 2017] ou encore en adaptation de domaine [Courty et al., 2017]. En effet, lorsque le coût est une distance, le transport optimal devient lui-même une distance appelée distance de Wasserstein. Formellement, cette distance prend deux mesures et un coût pour calculer la quantité suivante :

$$W_c(\alpha, \beta) = \min_{\pi \in \boldsymbol{U}(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(\boldsymbol{x}, \boldsymbol{y}) d\pi(\boldsymbol{x}, \boldsymbol{y}) \tag{1}$$

où $\boldsymbol{U}(\alpha, \beta)$ représente l'espace des contraintes. La formulation ci-dessus peut être vue comme une généralization de la distance de Wasserstein W_1 où la fonction de coût n'est pas la distance euclidienne mais est donnée par c. Sa complexité de calcul étant prohibitive à une application sur des jeux de données volumineux ($\mathcal{O}(n^3 log(n))$ entre distributions empirique avec n échantillons), les praticiens ont mis en place des stratégies de lots qui consistent à calculer et à moyenner le transport optimal sur des sous-problèmes du problème original. Cette pratique, bien qu'efficace, change le problème initial, car elle revient à calculer

$$WL_c(\alpha, \beta) = \mathbb{E}_{(X,Y)\sim\alpha^{\otimes m}\otimes\beta^{\otimes m}}[W_c(X, Y)], \tag{2}$$

où $\alpha^{\otimes m}$ est la loi produit (m fois) de α. Cependant, cette stratégie par lots brise le premier axiome d'une distance qui est intrinsèque à la distance de Wasserstein, ainsi lorsque les objets sont identiques, la "distance" de Wasserstein par lots $WL_c(\alpha, \alpha)$ est non nulle.

En pratique, nous ne connaissons pas les mesures α et β mais nous avons accès à des mesures empiriques α_n et β_n composées de n données. Ainsi nous pouvons estimer $WL_c(\alpha_n, \beta_n)$ à travers une partie des lots sur les données empiriques avec l'estimateur incomplet suivant :

$$\widetilde{U}_{W_c}^{k,m}(\alpha_n, \beta_n) := k^{-1} \sum_{(A,B)\in D_k} W_c(A, B) \tag{3}$$

où k désigne le nombre de lots et m la taille des lots et D_k est l'ensemble de cardinalité k dont les élements sont les sous lots tirés uniformément. Lorsque les données sont *iid*, cet estimateur est un estimateur non biaisé de $WL_c(\alpha, \beta)$.

La quantité $\widetilde{U}_{W_c}^{k,m}$ définit une U-statistique incomplète à deux populations et d'ordre $2n$. Or comme prouvé récemment par [Fatras et al., 2020], nous avons une inégalité de concentration entre l'estimateur et sa moyenne $WL_c(\alpha, \beta)$.

Theorem 1 (Borne de déviation, (Fatras et al., 2020)) *Fixons $\delta \in (0, 1)$, $k \geqslant 1$ et m, considérons deux mesures α, β à support compact. Nous avons une borne de déviation entre $\widetilde{U}_{W_c}^{k,m}(\alpha_n, \beta_n)$ et $WL_c(\alpha, \beta)$ qui dépend du nombre de données empiriques*

n et du nombre de lots k, avec probabilité d'au moins $1-\delta$ *sur le tirage de* α_n, β_n *et* D_k*, nous avons :*

$$|\widetilde{U}^{k,m}_{W_c}(\alpha_n,\beta_n) - WL_c(\alpha,\beta)| \leq M\left(\sqrt{\frac{\log(2/\delta)}{2\lfloor n/m\rfloor}} + \sqrt{\frac{2}{k}\log(2/\delta)}\right) \tag{4}$$

où M *est une constante et* $\lfloor n\rfloor$ *la partie entière de* n*.*

2 Divergence Wasserstein par lots

Nous proposons de corriger la distance de Wasserstein par lots de la manière suivante :

$$DWL_c(\alpha,\beta) := WL_c(\alpha,\beta) - 1/2(WL_c(\alpha,\alpha) + WL_c(\beta,\beta)). \tag{5}$$

Cette stratégie est inspirée de la divergence de Sinkhorn [Genevay et al., 2018] qui corrige le biais induit par le transport optimal entropique. Le lecteur pourra vérifier que la fonction DWL est nulle lorsque l'on a la même mesure à l'entrée et que c est une distance. Elle hérite aussi de la symétrie de la distance de Wasserstein. Cependant, plusieurs questions subsistent a priori. Est-ce que $\widetilde{U}^{k,m}_{W_c}$ estime facilement $DWL_c(\alpha,\beta)$ lorsque l'on a accès à des données empirique ? Est-ce que cet estimateur est toujours positif?

Concentration Nous voulons étendre l'inégalité (4) en enlevant la dépendance stochastique aux données empiriques α_n et β_n ainsi que la dépendance aux tirages des lots. Pour cela nous avons le résultat suivant :

Theorem 2 *Sous les hypothèses du Théorème 1, l'inégalité suivante est valable :*

$$\mathbb{E}[|\widetilde{U}^{k,m}_{W_c}(\alpha_n,\beta_n) - WL_c(\alpha,\beta)|] \leqslant 8\cdot M\max(\sqrt{\frac{1}{2\lfloor n/m\rfloor}}, \sqrt{\frac{2}{k}}) \tag{6}$$

Proof 1 *Nous rappellons la formule : pour une variable aléatoire réelle* X*, si* $X \geqslant 0$ *alors* $\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X \geq \lambda)d\lambda$*. Nous dénotons par* X *la variable aléatoire* $|\widetilde{U}^{k,m}_{W_c}(\alpha_n,\beta_n) - WL_c(\alpha,\beta)|$*. Le Théorème 1 s'écrit :*

$$\mathbb{P}\left(X > C\sqrt{\log(2/\delta)}\right) \leqslant \delta$$

où $C := 2M\max(\sqrt{\frac{1}{2\lfloor n/m\rfloor}}, \sqrt{\frac{2}{k}})$*. Nous pouvons la réécrire comme :*

$$\mathbb{P}(X > \lambda) \leqslant \exp\left(-\frac{\lambda^2}{C^2}\right)$$

Ainsi, en utilisant la formule ci-dessus :

$$\begin{aligned}\mathbb{E}[X] &\leqslant \int_0^\infty \exp\left(-\frac{\lambda^2}{C^2}\right)d\lambda \\ &= C\int_0^\infty \exp(-u^2)du \leqslant 4C\end{aligned}$$

ce qui complète la preuve.

Nous venons d'étendre la borne de déviation entre $\widetilde{U}_{W_c}^{k,m}$ et $WL_c(\alpha,\beta)$ avec un contrôle en espérance à la place d'un contrôle en probabilité. Or $DWL_c(\alpha,\beta)$ est composé de 3 termes de la forme $WL_c(.,.)$. Ainsi, nous pouvons utiliser ce théorème sur chaque terme de $DWL_c(\alpha,\beta)$ afin de prouver une inégalité de concentration similaire.

Ce résultat est important car il démontre que l'estimation de la fonction DWL sur des mesures continues n'est pas atteinte par la malédiction de la dimension. Or il est bien connu que l'estimation de la distance de Wasserstein entre deux mesures est atteinte par cette malédiction [Weed and Bach, 2019], ce qui fait que DWL est une candidate sérieuse pour des applications d'apprentissage automatique.

Expérimentation numérique Afin de mettre en évidence cette concentration qui ne dépend pas de la dimension, nous réalisons les expériences suivantes. Nous tirons deux échantillons α_n et $\alpha_n^{'}$ de taille n suivant la loi uniforme $U(0,1)$. La quantité $DWL_c(\alpha_n,\alpha_n^{'})$ doit converger vers 0 lorsque n tend vers l'infini. Afin de diminuer la variance dans le tirage des lots, nous fixons le nombre de lots k à 10^5. Pour la première expérience, nous fixons la taille des lots m à 128 et nous faisons varier la valeur de la dimension d des données entre 2, 7 et 10. La deuxième expérience revient cette fois-ci à fixer la dimension et à faire varier la taille des lots, cette dernière prend les valeurs 64, 128, 256. Dans les deux expériences nous observons que les courbes décroissent bien en $\mathcal{O}(n^{-1/2})$.

Positivité Nous détaillons à présent la partie positivité de $DWL(\alpha,\beta)$ qui reste une question ouverte. Lorsqu'on utilise la distance de Wasserstein avec un coût euclidien:

$$W_p(\alpha,\beta) = \min_{\pi\in\boldsymbol{U}(\alpha,\beta)}\left(\int_{\mathcal{X}\times\mathcal{Y}} ||\boldsymbol{x}-\boldsymbol{y}||_2^p d\pi(\boldsymbol{x},\boldsymbol{y})\right)^{1/p},$$

nous n'avons pas trouvé de contre-exemples. Cependant, nous avons remarqué qu'elle n'est pas vraie lorsque W_c n'est pas une distance. Par exemple pour W_2^2, nous avons le contre-exemple suivant. Nous disposons des points aux positions 0, $\pi/2$, π et $3\pi/2$ sur le cercle unité. Ces points définissent notre mesure α et nous perturbons cette distribution par un facteur ε. Ainsi la mesure β est définie par les points suivants : ε, $\pi/2+\varepsilon$, $\pi+\varepsilon$ et $3\pi/2+\varepsilon$. Or lorsque nous calculons $DWL(\alpha,\beta)$ pour une perturbation $\varepsilon = 0.1$, nous

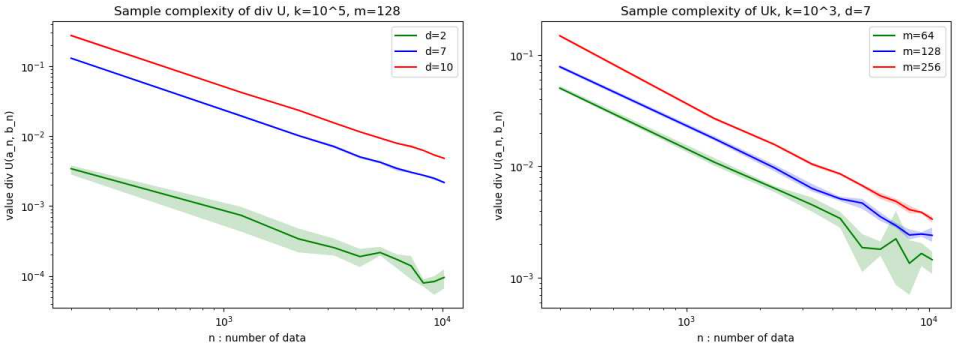


Figure 1: Convergence de $DWL_c(\alpha_n, \alpha_n^{'})$ vers 0 avec m fixé (gauche) et avec d fixé. (droite)

obtenons -0.029. Cependant lorsqu'on utilise W_p (avec p compris entre 1 et 5), cet exemple devient positif, ceci suggère donc que DWL_c est une divergence dès lors que W_c est une distance.

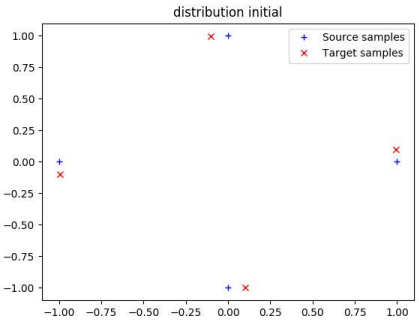


Figure 2: Exemple pour lequel $DWL(\alpha, \beta)$ est négatif, avec un coût c quadratique.

3 Conclusion

Dans ce papier, nous avons décrit une des principale limites à l'utilisation de la distance de Wasserstein par lot, à savoir la perte de l'axiome de séparabilité. Afin de corriger cette perte, nous avons élaboré une nouvelle fonction basée sur la distance de Wasserstein par lots. Nous avons démontré qu'elle possède des propriétés désirables dans un contexte de données empiriques, à savoir qu'elle ne souffre pas du fléau de la dimension. Nous avons

aussi détaillé la principale question ouverte sur la positivité qu'il reste à démontrer et nous avons mis en évidence certaines restrictions.

Bibliographie

Arjovsky, M. and Chintala, S. and Bottou, L. (2017), *Wasserstein Generative Adversarial Networks*, Proceedings of the 34th International Conference on Machine Learning.

Courty, N. and Flamary, R. and Tuia, D. and Rakotomamonjy, A. (2017), *Optimal Transport for Domain Adaptation*, IEEE Transactions on Pattern Analysis and Machine Intelligence.

Fatras, K. and Zine, Y. and Flamary, R. and Gribonval, R. and Courty, N. (2020), *Learning with minibatch Wasserstein : asymptotic and gradient properties*, AISTATS 2020.

Genevay, A. and Peyre, G. and Cuturi, M. (2018). Learning Generative Models with Sinkhorn Divergences, *AISTATS 2018*

Hoeffding, W. (1963). "Probability Inequalities for Sums of Bounded Random Variables", *Journal of the American Statistical Association*

Weed, J., and Bach, F. (2019). "Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance," *Bernoulli*, 25(4A), 2620–2648

Partitionnement de données incomplètes en utilisant l'imputation multiple et un clustering par consensus

Lilith Faucheux [1,2] & Sylvie Chevret [1,3]

[1] *Université de Paris, Sorbonne Paris Cité, Center of research in Epidemiology and Statistics, INSERM UMR-1153, ECSTRRA Team, 1 avenue Claude Vellefaux F-75010 Paris, France* [2] *lilith.faucheux@inserm.fr* [3] *sylvie.chevret@paris7.jussieu.fr*

Résumé. Les algorithmes de clustering permettent de proposer une partition des observations sur des critères de proximité. Une de leurs limites, communément admise, est qu'ils nécessitent un jeu de données complet. En présence de données manquantes, Faucheux *et al* (2020) ont proposé d'utiliser en premier lieu une procédure d'imputation multiple, puis d'appliquer `Multicons`, un algorithme de *clustering* par consensus pour combiner les partitions obtenues sur chaque jeu de données imputé, à l'instar des règles de Rubin moyennant les estimations d'un paramètre de population. Dans le cadre de cette méthode, nous montrons, à l'aide de simulations, qu'effectuer un consensus intermédiaire sur chaque jeu imputé avant tout consensus global améliore la qualité de la partition, bien que faiblement. De plus, de meilleures performances sont obtenues en augmentant le nombre d'imputations ainsi qu'en générant plus d'une partition par jeu imputé, sans avantage visible pour plus de 3 partitions.

Mots-clés. Partitionnement, consensus, données manquantes, imputation multiple.

Abstract. A major drawback of commonly used clustering algorithms is their requirement of a complete dataset. To overcome this issue, in a previous work (Faucheux *et al* (2020)), we proposed to first perform multiple imputation, then to use a consensus clustering algorithm (`Multicons`) to combine the partitions as one would use Rubin's rules when pooling the various estimates of a population parameter from an incomplete dataset. In this work, we showed via a simulation study that performing an additional intermediate consensus in each imputed dataset improved the quality of the partition, though slightly. Moreover, increasing the number of imputations also improved the quality of the partition. Lastly, generating more than one partition within each imputed dataset further improved the partition, though no benefit seemed to appear with more than 3 partitions.

Keywords. Clustering, consensus, missing data, multiple imputation.

1 Introduction

Le *clustering* est une approche de partitionnement non supervisé, utilisée pour regrouper des observations dont la classe n'est pas préalablement définie, sur des critères de proximité. De nombreux algorithmes de *clustering* ont été proposés, mais ils présentent plusieurs limites. Ils sont sensibles aux conditions initiales, notamment lorsque le nombre de groupes n'est pas connu *a priori*. De plus, ils requièrent un jeu de données complet.

Récemment proposé, le *clustering* par consensus permet de s'affranchir de la sensibilité aux conditions initiales. Il consiste à combiner plusieurs partitions d'un jeu de données en une partition finale, consensuelle, supposée être plus robuste au choix des paramètres et à leur initialisation que les algorithmes classiques (Vega-Pons et Ruiz-Shulcloper (2011)).

Pour avoir un jeu de données complet, la pratique courante est d'analyser seulement les cas complets (*complete cases analysis*, CCA). Little et Rubin (2002) ont pourtant montré que, dans le cadre d'une analyse supervisée (régression), quand le mécanisme de données manquantes n'est pas complètement aléatoire (*missing completely at random*, MCAR), l'estimation des coefficients est biaisée.

Différentes méthodes ont été proposées pour un *clustering* avec données manquantes. Faucheux *et al* (2020) ont utilisé `Multicons`, un algorithme de *clustering* par consensus (Al Najdi *et al* (2016)), après imputation multiple. Cela permet de combiner les partitions obtenues pour chaque jeu imputé, à l'instar des règles de Rubin dans le cas d'un paramètre de population. Cependant, il reste une incertitude sur le nombre d'imputations à effectuer, et la place du consensus dans l'algorithme. Nous avons donc conduit une étude de simulation pour tenter de répondre à ces interrogations.

2 *Clustering* après imputation multiple

L'algorithme `Multicons` utilise la technique d'extraction des itemsets fréquents fermés (*Frequent Closed Itemsets*) afin d'identifier des motifs, des similarités, entre des partitions. En variant le nombre de partitions nécessaires pour obtenir un motif, une hiérarchie de partitions est générée. La partition finale est celle qui est la plus similaire aux partitions initiales (selon l'indice de Jaccard), avec un nombre de groupes qui peut être différent de celui, variable, des partitions initiales.

Dans le cas d'un jeu incomplet, Faucheux *et al* (2020) ont combiné une procédure d'imputation multiple avec l'algorithme `Multicons`, en trois étapes :

Imputation multiple - m jeux de données complets sont générés par appariement d'après la moyenne prédite par des modèles de régressions linéaires séquentiels (MICE, van Buuren et Groothuis-Oudshoorn (2011)).

Clustering indépendants - Pour chacun des m jeux de données imputés, indépendemment, le nombre de groupes est estimé par le critère CH de Calinski et Harabasz (1974) puis p partitions sont obtenues par K-means.

Consensus - Les $m \times p$ partitions sont combinées en une partition finale en utilisant l'algorithme `Multicons`. Les groupes de cardinalité inférieure à 10 sont retirés de la

partition finale et leurs observations considérées comme non classées.

Le consensus peut alors être obtenu soit une fois sur l'ensemble des $m \times p$ partitions ("consensus global"), soit répété : (i) sur les p partitions de chaque jeu imputé, puis (ii) sur les m partitions consensuelles ("consensus intermédiaire").

3 Étude de simulation

3.1 Génération des données

Les données ont été simulées de manière à reproduire un contexte expérimental (détaillé dans Faucheux *et al* (2020)). Des jeux de données de 500 individus indépendants répartis en trois groupes de tailles respectives $n_1 = 166$, $n_2 = n_3 = 167$ ont été simulés.

Données complètes - Conditionnellement aux groupes prédéfinis k $(= 1, 2, 3)$, trois variables normales $\boldsymbol{X}$=$(X_1, X_2, X_3) \sim \mathcal{N}(\boldsymbol{\mu_k}, \boldsymbol{\Sigma})$ ont été générées, où $\boldsymbol{\mu_k}$ est un 3-vecteur de moyennes dépendant du groupe, et $\boldsymbol{\Sigma}$ une matrice de covariance $0.3 \times \mathbf{I_3}$. A noter que si les variables sont indépendantes, conditionnellement au groupe, une structure de corrélation entre elles est induite par la position relative des groupes.

Données manquantes - Des données manquantes ont ensuite été générées sur X_1 selon une distribution de Bernoulli de paramètre p, variable selon trois mécanismes : (i) pour le mécanisme MCAR, p a été fixé à une prévalence de 30%, (ii) pour le mécanisme *manquant aléatoirement* (MAR), p dépendait de X_2 par $\text{logit}(p) = \beta_{02}+\beta_2 X_2$, (iii) pour le mécanisme *manquant non aléatoirement* (MNAR), p dépendait de X_1 par $\text{logit}(p) = \beta_{01} + \beta_1 X_1$. Les intercepts (β_{01}, β_{02}) ont été fixés de manière à contrôler la prévalence des données manquantes sur X_1 à 30%, et les coefficients $\beta_1 = \beta_2 = 2$ afin d'obtenir des dépendances, respectivement à X_2 et X_1, modérées.

Scénarios - Trois scénarios de séparation des groupes ont été simulés en variant les moyennes $\boldsymbol{\mu_k}$ des groupes. Dans le Scénario 1, les groupes sont non alignés et complètement distincts. Dans le Scénario 2, les groupes sont non alignés mais se rapprochent. Dans le Scénario 3, les trois groupes sont alignés, avec une superposition importante des groupes. L'ordre de grandeur du chevauchement moyen de la distribution multivariée entre deux groupe est 10^{-10} pour le Scénario 1, 10^{-4} pour le Scénario 2, et pour le Scénario 3, 10^{-2} et 10^{-5}, respectivement pour deux groupes consécutifs et non consécutifs. Une représentation des trois scénarios est illustrée sur la figure 1.

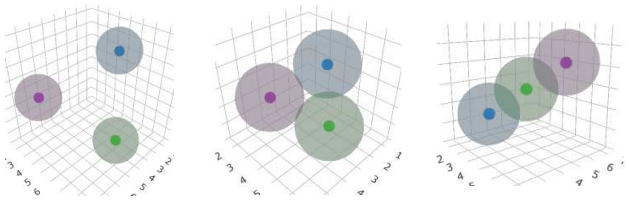
Pour chaque scénario, nous avons généré $N = 1000$ réplications indépendantes de jeux de données $\boldsymbol{X}$, soit complets, soit incomplets selon l'un des trois mécanismes décrits.

3.2 Procédures de *clustering*

Différentes valeurs de paramètres de la procédure décrite en partie 2 ont été évaluées :

Nombre d'imputations - $m = 3, 5$ jeux de données ont été imputés.

Nombre de partitions - $p = 1, 3, 5, 10, 20$ partitions ont été obtenues pour chaque jeu de données imputé. Le nombre total de partitions est donné par $m \times p$.



Scénario 1 Scénario 2 Scénario 3

Figure 1: Représentation des centres de groupes simulés et de leur région de recouvrement à 95% pour les trois scénarios. Les paramètres de moyenne sont disponibles dans Faucheux *et al* (2020).

Consensus - Un consensus intermédiaire a été évalué, ainsi qu'un consensus global lorsque le nombre de partitions total n'était pas trop élevé ($m \times p \leq 15$). Cette restriction est imposée par les limites de calcul de l'algorithme de consensus.

3.3 Evalutation des performances

Les performances ont été évaluées par l'indice de Rand ajusté (ARI) (Vinh *et al* (2010)). Cet indice quantifie la concordance entre la partition P estimée et la partition P_{ref} réelle des données, variant de 0 à 1 lorsque les deux partitions sont identiques. Pour quantifier la difficulté du partitionnement, un K-means a été effectué sur les données complètes.

3.4 Résultats

Construction de la partition finale

La Figure 2A présente l'ARI obtenu avec consensus "intermédiaire" ou seulement "global" sur un même jeu de données, en fonction du scénario, du mécanisme de données manquantes et du nombre d'imputations. Pour toutes les situations étudiées, l'ajout d'un consensus intermédiaire ne modifie que peu la qualité de la partition finale, avec un ARI identique dans 87.9% des cas. Néanmoins, pour les Scénarios 1 et 2, on observe de rares simulations où le consensus global seul atteint une meilleure partition qu'après un consensus intermédiaire (1.2% et 0.5% des simulations ont une différence d'ARI supérieure à 0.01 respectivement pour le Scénario 1 et 2). En revanche, l'ajout d'un consensus intermédiaire améliore ou ne modifie pas la qualité de la partition par rapport a un consensus global seul dans la plupart des simulations (97.4%, 95.3% et 93.6% des cas respectivement pour le Scénario 1, 2 et 3). Seuls les résultats de l'algorithme avec consensus intermédiaire seront présentés par la suite.

Nombre d'imputations versus partitions

La différence d'ARI selon le nombre d'imputations pour un nombre total de partitions égal à $m \times p = 15$ est représentée sur la figure 2B. Aucun impact n'est observé pour le Scénario 1. Pour les Scénarios 2 et 3, un nombre élevé d'imputations m plutôt que de

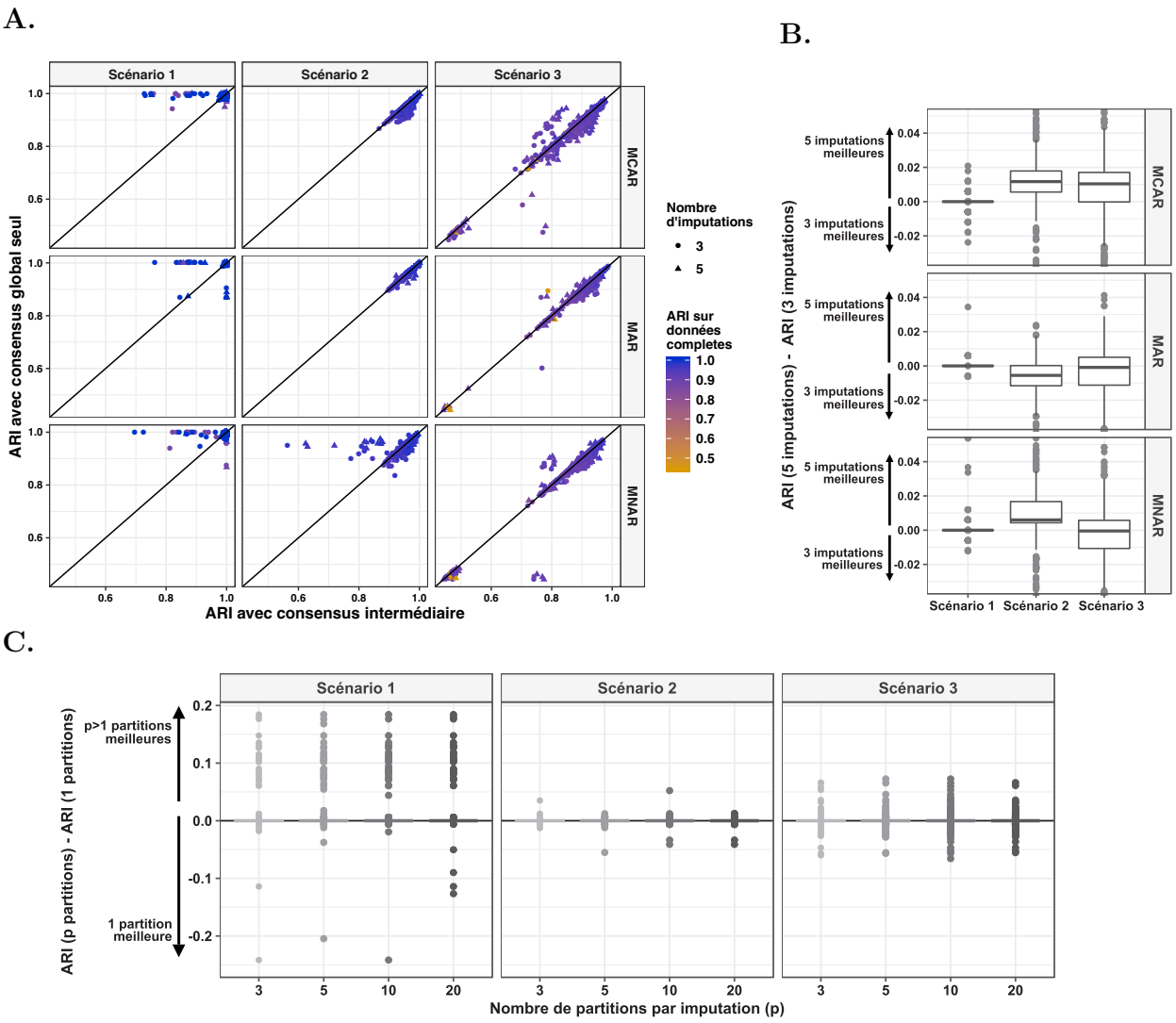


Figure 2: Évaluation des performances du *clustering* selon le consensus (”intermédiaire” ou ”global”), le nombre d'imputations (m) et de partitions (p), et le mécanisme de données manquantes (MCAR, MAR et MNAR), pour les 3 scénarios.
A. Évaluation du consensus : ARI obtenus après consensus ”intermédiaire” ou ”global”, avec $m = 3$ et $p = 5$ ou $m = 5$ et $p = 3$. **B.** Évaluation du nombre d'imputations avec consensus intermédiaire selon le mécanisme de données manquantes : Différence d'ARI avec $m = 3$ et $p = 5$ ou $m = 5$ et $p = 3$. **C.** Évaluation du nombre de partitions sur chacun de $m = 5$ jeux MCAR imputés avec consensus intermédiaire : Différences d'ARI sur p et 1 partition.

partitions sur chaque jeu imputé p conduit à de meilleures performances sauf dans le cas MAR, notamment dans le Scénario 2. Par la suite, les résultats présentés concerneront tous un consensus intermédiaire sur $m = 5$ jeux imputés.

Nombre de partitions

La figure 2C représente la différence d'ARI entre les partitions finales obtenues avec $p > 1$ et $p = 1$ partition(s) sur chaque jeu imputé. Aucun impact n'ayant été observé selon le mécanisme ayant généré les données manquantes, la figure ne présente que les résultats pour le cas MCAR. L'ARI obtenu avec $p = 1$ et $p = 3$ partitions par imputation était identique dans 95.6%, 92.5%, et 80.7% des cas, respectivement dans le Scénario 1, 2, et 3. Parmi les rares cas non identiques, $p = 3$ partitions a donné une meilleure partition finale que $p = 1$ partition dans 80%, 55%, et 53% des cas, respectivement dans le Scénario 1, 2, et 3. Enfin, dans l'ensemble des cas de figures étudiés, aucune différence de performance n'est apparue au delà de $p = 3$ partitions par imputation.

3.5 Conclusion

Au total, dans le cadre des simulations effectuées, nous recommandons l'utilisation d'un consensus intermédiaire, en évitant un nombre trop faible d'imputations. La réalisation de plus d'une partition par imputation, qui apporte rarement un bénéfice mais est peu coûteuse en temps, peut également être recommandée.

Bibliographie

Al-Najdi, A., Pasquier, N., and Precioso, F. (2016). Frequent closed patterns based multiple consensus clustering, In *International Conference on Artificial Intelligence and Soft Computing, Springer International Publishing*, 14–26.

Calinski, T., and Harabasz, J. (1974). A dendrite method for cluster analysis, *Communications in Statistics - Theory and Methods*, 3, 1–27.

Faucheux, L., Resche-Rigon, M., Curis, E., Soumelis, V. and Chevret, S. (2020). Clustering with missing and left-censored data: A simulation study comparing multiple-imputation-based procedures, *Biometrical Journal*, 1–22.

Little, R. J. A. and Rubin, D. B. (2002), *Statistical analysis with missing data*, Wiley, New York.

Van Buuren, S., and Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R, *Journal of Statistical Software*, 45, 1–67.

Vega-Pons, S. and Ruiz-Shulcloper, J. (2011). A survey of clustering ensemble algorithms, *International Journal of Pattern Recognition and Artificial Intelligence*, 25, 337–372.

Vinh, N. X., Epps, J. and Bailey, J. (2010). Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance, *Journal of Machine Learning Research*, 11, 2837–2854.

Learning with signatures: estimation in the expected signature model.

Adeline Fermanian [1]

[1] *Sorbonne Université, Laboratoire de Probabilités, Statistique et Modélisation, 4 place Jussieu, 75005 Paris, France, adeline.fermanian@sorbonne-universite.fr*

Résumé. Des données séquentielles ou temporelles sont utilisées dans de nombreux domaines de recherche, tels que la finance quantitative, la médecine ou la vision par ordinateur. Nous nous intéresserons à une nouvelle approche de l'apprentissage séquentiel, appelée méthode par signature, qui a son origine dans la théorie des chemins rugueux. Son principe de base est de représenter des chemins multidimensionnels par un ensemble gradué de leurs intégrales itérées, appelé la signature. Les signatures encodent les propriétés géométriques des données considérées comme des chemins. Elles caractérisent ces chemins (à une classe d'équivalence près) et les fonctions linéaires sur la signature approximent arbitrairement bien toute fonction continue. Ces bonnes propriétés font de la signature une approche pertinente en apprentissage. En outre, elles ont présenté des résultats concluants pour une série d'applications (voir, par exemple, Lyons et al, 2014 ; Yang et al, 2015). Dans notre exposé, nous présenterons en détail le modèle linéaire sur la signature, introduit pour la première fois par Levin et al (2013), définirons les estimateurs dans ce modèle et montrerons leur convergence presque sûre à une vitesse exponentielle.

Mots-clés. Données séquentielles, signature, modèle linéaire fonctionnel ...

Abstract. Sequential or temporal data arise in many fields of research, such as quantitative finance, medicine, or computer vision. We will be concerned with a novel approach for sequential learning, called the signature method, and rooted in rough path theory. Its basic principle is to represent multidimensional paths by a graded feature set of their iterated integrals, called the signature. Signatures encode geometric properties of input functions, considered as paths. They characterize these paths (up to a negligible equivalence class) and linear functions on the signature approximate arbitrarily well continuous functions of paths. These appealing properties make signatures a relevant approach in statistical learning. Moreover, they have achieved a series of successful applications in machine learning (see, e.g., Lyons et al, 2014; Yang et al, 2015). In our talk, we will present in detail the expected signature model, first introduced by Levin et al (2013), define estimators in this model and show their almost sure convergence at an exponential rate.

Keywords. Sequential data, signatures, functional linear model ...

1 Introduction

In a classical regression setting, a real output Y is described by a finite number of predictors. A typical example would be to model the price of a house as a linear function of several characteristics such as surface area, number of rooms, location, and so on. These predictors are typically encoded as a vector in $\mathbb{R}^p$, $p \in \mathbb{N}^*$. However, some applications do not fall within this setting. For example, in medicine, a classical task consists in predicting the state of a patient (e.g., ill or not) from the recording of several physiological variables over a period of time. The input data is then a multidimensional time series, and not a vector. Similarly, sound recognition or stock market prediction tasks both consist of learning from time series, possibly multidimensional. Then, the question arises of extending the linear model to this more general setting, where one wants to predict from a functional input, of the form $X : [0, 1] \to \mathbb{R}^d$, $d \in \mathbb{N}^*$.

The approaches undertaken in the literature are diverse. On the one hand, input functions may be modeled as realizations of stochastic processes, so that these models depend only on a finite number of coefficients which can be estimated with parametric statistics tools. In particular, in time series analysis, several versions of the ARIMA models haven been studied, with a focus on the modelling of time dependence. On the other hand, in functional data analysis (Ramsay and Silverman, 2005), input functions are traditionally represented on a set of basis functions, for example splines or the Fourier basis, and then coefficients of the projection of input functions on this basis are estimated. However, these models do not handle easily multidimensional series.

In the present article, we are interested in a novel approach to functional regression, called the expected signature model and introduced by Levin et al. (2013). Its principle is to represent input functions by their signatures, defined as an infinite series of their iterated integrals. Signatures date back from the 60s when Chen (1958) showed that a path can be faithfully represented by its iterated integrals and it has been at the center of Lyon's rough paths theory in the 90s. We refer the reader to the monographs by Lyons et al. (2007) and Friz and Victoir (2010) for recent accounts of the theory. It has recently gained attention from the machine learning community, due to some successful applications, for example Arribas et al. (2018), Lyons et al. (2014), Yang et al. (2015) or Yang et al. (2017). The main advantage of this approach is that it can handle multidimensional input functions, that is inputs X with values in $\mathbb{R}^d$, where d can be large, whereas traditional methods struggle to model interactions in multidimensional series. Moreover, it requires little assumptions on the regularity of X and encodes nonlinear geometric information about X. Finally, linear functions of the signature can approximate arbitrarily well any continuous function of X, making signatures a relevant feature set in a linear regression setting.

In the present article, we tackle the problem of estimation in a regression model on the signature. Any continuous function of X is approximated by a scalar product on the truncated signature. Therefore, the estimation of a regression function boils down to

the estimation of the coefficients of the scalar product on the truncated signature. The truncation order of the signature is a crucial parameter as it controls the complexity of the model. The main purpose of the article is to estimate this parameter.

2 The expected signature model

In this section, we present in detail the expected signature model, first introduced by Levin et al (2013). Let $X : [0,1] \to \mathbb{R}^d$, $X_t = (X_t^1, \dots, X_t^d)$, be a multidimensional path of bounded variation. Then, its signature is the infinite sequence defined by

$$S(X) = \big(1, S^{(1)}(X), \dots, S^{(d)}(X), S^{(1,1)}(X), S^{(1,2)}(X), \dots, S^{(i_1,\dots,i_k)}(X), \dots\big),$$

where, for any $I = (i_1, \dots, i_k) \subset \{1, \dots d\}^k$,

$$S^I(X) = \int_{0 \le u_1 < \dots < u_k \le 1} \cdots \int dX_{u_1}^{i_1} \dots dX_{u_k}^{i_k}.$$

In practice, the signature series is truncated at a certain order m and then denoted by $S^m(X) = \big(1, S^{(1)}(X), S^{(2)}(X), \dots, S^{(d,\dots,d)}(X)\big)$, where coefficients corresponding to an index I of length less than m are kept. $S^m(X)$ is then of length

$$s_d(m) = \sum_{k=0}^{m} d^k = \frac{d^{m+1} - 1}{d - 1}.$$

We place ourselves in a regression setting, that is we want to learn a random variable $Y \in \mathbb{R}$ from a random path $X : [0,1] \to \mathbb{R}^d$, assumed to be of bounded variation. Our model is then the following: we assume that there exists $m^* \in \mathbb{N}^*$, $\beta^* \in \mathbb{R}^{s_d(m^*)}$, such that

$$\mathbb{E}\left[Y|X\right] = \left\langle \beta_{m^*}^*, S^{m^*}(X) \right\rangle, \quad \text{and} \quad \mathrm{Var}(Y|X) = \sigma^2 < \infty. \tag{1}$$

This model assumes that Y is a linear function of signature features of X up to a certain unknown order m^*. The signature truncation order m^* is a key quantity in this model, as it controls the complexity. It is necessary for applications to have an idea of the magnitude of m^*. Therefore, our main focus is to build a consistent estimator of m^*. As we will see later, a simple estimator of $\beta_{m^*}^*$, and therefore of the regression function, is then also obtained.

3 Estimating the truncation order

We are now in a position to define the estimator of m^*. Let $D_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ be some i.i.d. observations drawn according to the law of (X, Y), we use the approach

of penalized empirical risk minimization. For the moment, let us fix a certain truncation order $m \in \mathbb{N}$, and let $\alpha > 0$ denote a fixed positive number. Then, for any $m \in \mathbb{N}$, the ball in $\mathbb{R}^{s_d(m)}$ of radius α centered at 0 is denoted by

$$B_{m,\alpha} = \big\{\beta \in \mathbb{R}^{s_d(m)} \mid \|\beta\| \leq \alpha\big\},$$

where $\|\cdot\|$ denoted the Euclidean norm. By a slight abuse of notation, the sequence $\big(B_{m,\alpha}\big)_{m\in\mathbb{N}}$ can be seen as a nested sequence of balls, i.e.,

$$B_{0,\alpha} \subset B_{1,\alpha} \subset \cdots \subset B_{m,\alpha} \subset B_{m+1,\alpha} \subset \cdots .$$

From now on, we will only consider coefficients within these balls. Therefore, we assume that the true coefficient $\beta^*_{m^*}$ lies within such a ball, i.e.,

$$\beta^*_{m^*} \in B_{m^*,\alpha}. \tag{2}$$

For a fixed truncation order m, the theoretical risk is defined by

$$\mathcal{R}_m(\beta) = \mathbb{E}\big(Y - \big\langle \beta, S^m(X)\big\rangle\big)^2.$$

The minimal theoretical risk for a certain truncation order m, denoted by $L(m)$ is then

$$L(m) = \inf_{\beta\in B_{m,\alpha}} \mathcal{R}_m(\beta).$$

Since the sets $B_{m,\alpha}$ are nested, L is a decreasing function of m. Its minimum is attained at $m = m^*$, and, provided $m \geq m^*$, $L(m)$ is then constant and equal to

$$\mathcal{R}(\beta^*_{m^*}) = \mathbb{E}\big(Y - \big\langle \beta^*_{m^*}, S^{m^*}(X)\big\rangle\big)^2 \leq \sigma^2.$$

Then, the empirical risk with signatures truncated at order m is defined by

$$\widehat{\mathcal{R}}_{m,n}(\beta) = \frac{1}{n}\sum_{i=1}^{n} \big(Y_i - \big\langle \beta, S^m(X_i)\big\rangle\big)^2,$$

where $\beta \in B_{m,\alpha}$. The minimum of $\widehat{\mathcal{R}}_{m,n}$ over $B_{m,\alpha}$ is denoted by $\widehat{L}_n(m)$ and defined as

$$\widehat{L}_n(m) = \min_{\beta\in B_{m,\alpha}} \widehat{\mathcal{R}}_{m,n}(\beta) = \widehat{\mathcal{R}}_{m,n}(\widehat{\beta}_m),$$

where $\widehat{\beta}_m$ denotes a point in $B_{m,\alpha}$ where the minimum is attained. Note that $\beta \mapsto \widehat{\mathcal{R}}_{m,n}(\beta)$ is a convex function so $\widehat{\beta}_m$ exists. We point out that minimizing $\widehat{\mathcal{R}}_{m,n}$ over $B_{m,\alpha}$ is equivalent to performing a Ridge regression with a certain regularization parameter which depends on α.

In short, for a fixed truncation order m, a Ridge regression gives us the best parameter $\widehat{\beta}_m$ to model Y as a linear form on the signature of X truncated at order m. We must now find a truncation order $\widehat{m}$ close to the true one m^*. Since the $B_{m,\alpha}$ are nested, the sequence $(\widehat{L}_n(m))_{m\in\mathbb{N}}$ decreases with m. Thus, increasing m makes the set of parameters larger and therefore decreases the empirical risk. An estimator of m^* can then be defined by a trade-off between this decreasing empirical risk and an increasing function that penalizes the number of coefficients. More precisely, we let

$$\widehat{m} = \inf\Big(\operatorname*{argmin}_{m\in\mathbb{N}}\big(\widehat{L}_n(m) + \mathrm{pen}_n(m)\big)\Big), \tag{3}$$

where $\mathrm{pen}_n(m)$ is an increasing function of m that will be defined in Theorem 1.

4 Result

In this section, we derive a penalization that ensures exponential convergence of $\widehat{m}$ to m^*. In addition to (1) and (2), we need the following assumption:

- (H) there exists $K_Y > 0$ and $K_X > 0$ such that almost surely $|Y| \leq K_Y$ and $\|X\|_{TV} \leq K_X$,

where $\|\cdot\|_{TV}$ denotes the total variation norm. (H) says that the trajectories have a length uniformly bounded by K_X, which is in practice a reasonable assumption. We shall also use the constant K, defined by

$$K = 2(K_Y + \alpha e^{K_X})e^{K_X}. \tag{4}$$

The main result of the section is the following.

Theorem 1. *Let $K_{\mathrm{pen}} > 0$, $0 < \rho < \frac{1}{2}$, and denote*

$$\mathrm{pen}_n(m) = K_{\mathrm{pen}} n^{-\rho}\sqrt{s_d(m)}.$$

Let n_0 be the smallest integer satisfying

$$n_0 \geq \left((432K\alpha\sqrt{\pi} + K_{\mathrm{pen}})\sqrt{s_d(m^*+1)}\Big(\frac{2}{L(m^*-1)-\sigma^2} + \frac{\sqrt{2}}{K_{\mathrm{pen}}\sqrt{d^{m^*+1}}}\Big)\right)^{1/\tilde{\rho}},$$

where $\tilde{\rho} = \min(\rho, \frac{1}{2} - \rho)$. Then, under assumption (H), for any $n \geq n_0$,

$$\mathbb{P}\left(\widehat{m} \neq m^*\right) \leq C_1 \exp\left(-C_2 n^{1-2\rho}\right),$$

where C_1 and C_2 are constants.

This of course implies the almost sure convergence of $\widehat{m}$ to m^*. The proof of Theorem 1 is based on chaining tail inequalities that bound uniformly the tails of the risk. We can see that the penalty decreases slowly with n (more slowly than a square-root) and, if $d \geq 2$, increases with m exponentially, i.e., as $d^{m/2}$.

With an estimator of $\widehat{m}$ at hand, one can simply choose to estimate $\beta^*_{m^*}$ by $\widehat{\beta}_{\widehat{m}}$, which gives an estimator of the regression function in model (1). As a by-product of Theorem 1, we get

$$\mathbb{E}\Big(\big\langle \hat{\beta}_{\widehat{m}}, S^{\widehat{m}}(X)\big\rangle - \big\langle \beta^*_{m^*}, S^{m^*}(X)\big\rangle\Big)^2 = O\Big(\frac{1}{\sqrt{n}}\Big).$$

We recall that our main objective was to estimate m^*, and the estimator of the regression function may of course be non optimal.

Bibliographie

Arribas, I. P., Goodwin, G. M., Geddes, J. R., Lyons, T., Saunders, K. E. (2018). A signature-based machine learning model for distinguishing bipolar disorder and borderline personality disorder. *Translational psychiatry*, 8(1), 1-7.

Chen, K. (1958), Integration of paths—a faithful representation of paths by non-commutative formal power series, *Transactions of the American Mathematical Society*, 89 (2), pp. 395–407.

Friz, P. K. and Victoir, N. B. (2010), *Multidimensional Stochastic Processes as Rough Paths: Theory and Applications*, Cambridge University Press, Cambridge.

Levin D., Lyons T., Ni H. (2013) Learning from the past, predicting the statistics for the future, learning an evolving system. arXiv:1309.0260

Lyons, T. J., Caruana, M., Lévy, T. (2007), *Differential equations driven by rough paths*, Springer, Berlin.

Lyons, T., Ni H., Oberhauser H. (2014) A feature set for streams and an application to high-frequency financial tick data. In: *Proceedings of the 2014 International Conference on Big Data Science and Computing*, pp. 5.

Ramsay, J. O. and Silverman, B. W. (2005), *Functional Data Analysis. 2nd Edition*, Springer, New York.

Yang W., Jin L., Liu M. (2015) Chinese character-level writer identification using path signature feature, DropStroke and deep CNN. In: *Proceedings of the 13th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 546–550.

Yang, W., Lyons T., Ni H., Schmid C., Jin L., Chang J. (2017), Leveraging the Path Signature for Skeleton-based Human Action Recognition, *arXiv preprint arXiv:1707.03993*.

Experimental Comparison of Semi-parametric, Parametric, and Machine Learning Models for Time-to-Event Analysis Through the Concordance Index

Camila Fernández[1], Chung Shue Chen[2], Pierre Gaillard[3], Alonso Silva[4]

[1] *University of Chile & Nokia Bell Labs, camila.fernandez@nokia.com*
[2] *Nokia Bell Labs, chung_shue.chen@nokia-bell-labs.com*
[3] *Inria, pierre.gaillard@inria.fr*
[4] *Safran Tech, alonso.silva-allende@safrangroup.com*

Résumé. Dans cette communication, nous faisons une comparaison des méthodes semi-paramétriques (modèle à risque proportionnel de Cox, modèle additif d'Aalen), paramétriques (modèle Weibull de temps de défaillance accéléré), et d'apprentissage automatique (Random Survival Forest, Gradient Boosting Cox proportional hazards loss, DeepSurv) à travers l'indice de concordance dans deux jeux de données (PBC et GBCSG2). Nous faisons deux comparaisons: une avec les hyperparamètres par défaut de ces méthodes et une avec les meilleurs hyperparamètres (ces hyperparamètres ont été trouvés par une recherche aléatoire automatique).

Keywords. Apprentissage automatique, Analyse de survie, Santé.

Abstract. In this paper, we make an experimental comparison of semi-parametric (Cox proportional hazards model, Aalen additive model), parametric (Weibull AFT model), and machine learning methods (Random Survival Forest, Gradient Boosting Cox proportional hazards loss, DeepSurv) through the concordance index on two different datasets (PBC and GBCSG2). We present two comparisons: one with the default hyperparameters of these methods and one with the best hyperparameters found by randomized search.

Keywords. Machine Learning, Survival Analysis, Health.

1 Introduction

Time-to-event analysis originated from the idea to predict the time until a certain critical event occurs. For example, in healthcare, the goal is usually to predict the time until a patient with a certain disease dies. Another example is maintenance where the objective is to predict the time until a component fails. There are many other examples that are of interest to time-to-event analysis such as predicting customer churn, predicting the time until a convicted criminal reoffends, etc. One of the main challenges of time-to-event analysis is right censoring, which means that the event of interest has only occurred for a

subset of the observations, making the problem different from typical regression problems in machine learning.

In this paper, we will use two datasets to perform this analysis. The first one is about patients diagnosed with breast cancer (GBCSG2) and the second one are patients diagnosed with primary biliary cirrhosis (PBC). For the first dataset the critical event of interest will be the recurrence of cancer while for the second one it will be the death of the patient.

In each dataset and for each sample we have an observed time that could be either the survival time or the censored time. A censored time will occur when the time of death has not been observed, and then, in this case this time corresponds to the last medical record of the patient. The censored time will be a lower bound for the survival time.

1.1 Survival and hazard functions

The fundamental task of time-to-event analysis is to estimate the probability distribution of time until some event of interest happens.

Consider a covariates/features vector X, a random variable that takes on values in the covariates/features space $\mathcal{X}$. Consider a survival time T, a non-negative real-valued random variable. Then, for a feature vector $x \in \mathcal{X}$, our aim is to estimate the conditional survival function:

$$S(t|x) := \mathbb{P}(T > t|X = x), \tag{1}$$

where $t \geq 0$ is the time and $\mathbb{P}$ is the probability function.

In order to estimate the conditional survival function $S(\cdot|x)$, we assume that we have access to n training samples, in which for the i-th sample we have: X_i the feature vector, δ_i the survival time indicator, which indicates whether we observe the survival time or the censoring time, and Y_i which is the survival time if $\delta_i = 1$ and the censoring time otherwise.

Many models have been proposed to estimate the conditional survival function $S(\cdot|x)$. The most standard approaches are the semi-parametric and parametric models, which assume a given structure of the hazard function:

$$h(t|x) := -\frac{\partial}{\partial t} \log S(t|x). \tag{2}$$

1.2 Concordance index

The concordance index, introduced by Harrell et al. (1996), is the most used performance metric for time-to-event analysis. It measures the fraction of pairs of subjects that are correctly ordered within the pairs that can be ordered. The highest (and best) value that can be obtained is 1, which means that there is complete agreement between the order of the observed and predicted times. The lowest value that can be obtained is 0, which denotes a perfectly wrong model, while a value of 0.5 means that it is a random model.

To calculate the concordance index we first take every pair in the test set such that the earlier observed time is not censored. Then we consider only pairs (i, j) such that $i < j$ and we also eliminate the pairs for which the times are tied unless at least one of them has an event indicator value of 1. Next, we compute for each pair (i, j) a score $C_{i,j}$ which for $Y_i \neq Y_j$ is 1 if the subject with earlier time (between i and j) has higher predicted risk (between i and j), is 0.5 if the risks are tied and 0 otherwise. For $Y_i = Y_j$ and $\delta_i = \delta_j = 1$ we set $C_{i,j} = 1$ if the risks are tied and 0.5 otherwise. If only one of δ_i or δ_j is 1 we set $C_{i,j} = 1$ if the predicted risk is higher for the subject with $\delta = 1$ and 0.5 otherwise.

Final we compute the concordance index as follows

$$\frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} C_{i,j}, \tag{3}$$

where $\mathcal{P}$ represents the set of eligible pairs (i, j).

2 Datasets description

2.1 German Breast Cancer Study Group dataset (GBCSG2)

The German Breast Cancer Study Group (GBCSG2) dataset, made available by Schumacher et al. (1994), studies the effects of hormone treatment on recurrence-free survival time. The event of interest is the recurrence of cancer time. The dataset has 686 samples and 8 covariates/features: age, estrogen receptor, hormonal therapy, menopausal status (premenopausal or postmenopausal), number of positive nodes, progesterone receptor, tumor grade, and tumor size. At the end of the study, there were 387 patients (56.4%) who were right censored (recurrence-free). In our experiments, we reserve 25% of the dataset as testing set.

2.2 Mayo Clinic Primary Biliary Cirrhossis dataset (PBC)

The Mayo Clinic Primary Biliary Cirrhosis dataset, made available by Therneau and Grambsch (2000), studies the effects of the drug D-penicillamine on the survival time. The event of interest is the death time. The dataset has 276 samples and 17 covariates/features: age, serum albumin, alkaline phosphotase, presence of ascites, aspartate aminotransferase, serum bilirunbin, serum cholesterol, urine copper, edema, presence of hepatomegaly or enlarged liver, case number, platelet count, standardised blood clotting time, sex, blood vessel malformations in the skin, histologic stage of disease, treatment and triglycerides. At the end of the study, there were 165 patients (59.8%) who were right censored (alive). In our experiments, we reserve 25% of the dataset as testing set.

3 Models

3.1 Semi-parametric model: Cox proportional hazards

Cox (1972) proposes a semi-parametric model, also known as Cox proportional hazards model, to estimate the conditional survival function. This model assumes that the log-hazard of a subject is a linear function of their m static covariates/features $h_i, i \in [m]$, and a population-level baseline hazard function $h_0(t)$ that changes over time:

$$h(t|x) = h_0(t) \exp\left(\sum_{i=1}^{m} h_i(x_i - \bar{x_i})\right). \tag{4}$$

The term 'proportional hazards' refers to the assumption of a constant relationship between the dependent variable and the regression coefficients. Also, this model is semi-parametric in the sense that the baseline hazard function $h_0(t)$ does not have to be specified and it can vary allowing a different parameter to be used for each unique survival time.

3.2 Semi-parametric model: Aalen's additive model

Aalen's additive model, proposed by Aalen (1989), estimates the hazard function but instead of being a multiplicative model as the Cox proportional hazards model, it is an additive model. The hazard function estimator is the following

$$h(t|x) = b_0(t) + \sum_{i=1}^{m} b_i(t)x_i. \tag{5}$$

3.3 Parametric model: Weibull Accelerated Failure Time model (Weibull AFT)

Consider we have two survival functions for each one of two different populations, $S_A(t)$ and $S_B(t)$ and an accelerated failure rate λ such that $S_A(t) = S_B(\frac{t}{\lambda})$ where λ can be modeled as a function of the covariates/features and it describes stretching out or contraction of the survival time:

$$\lambda(x) = \exp\left(b_0 + \sum_{i=1}^{m} b_i x_i\right) \tag{6}$$

Then, we suppose a Weibull form for the survival function $S(t)$ leading us to assume

$$h(t|x) = \left(\frac{t}{\lambda(x)}\right)^{\rho} \tag{7}$$

where ρ is an unknown parameter that must be fitted. This model is called Weibull accelerated failure time shortened as Weibull AFT model.

3.4 Machine learning model: Random Survival Forest

The random survival forest model, proposed by Ishwaran et al. (2008), is an extension of the random forest model, introduced by Breiman et al. (2001), that can take into account censoring. The randomness is introduced in two ways, first we use bootstrap samples of the dataset to grow the trees and second, at each node of the tree, we randomly choose a subset of variables as candidates for the split. The quality of a split is measured by the log-rank splitting rule. Then, we average the trees results which allows us to improve the accuracy and avoid overfitting.

3.5 Machine learning model: Random Survival Forest and Adaptative Nearest Neighbors

We also consider a random survival forest variation from Chen (2019). Each leaf will be associated to a different subset of the data set for which a Kaplan Meier survival estimator is applied, and so, each leaf is associated to a survival function estimate. Then, for a test point x we choose all the leaves that x belongs to and we only average the results of these leaves to obtain our final estimation.

3.6 Machine learning model: Gradient Boosting Cox Proportional Hazards Loss

The idea of gradient boosting was originated by Breiman and later developed by J.H. Friedman (2001). Gradient boosting is an additive model in which at each step it adds a new weak learner so that it minimizes a loss function. The model has principally three components, the loss function, the weak learner and the additive model. The loss function we aim to minimize will be the negative Cox's log partial likelihood, as proposed by Ridgeway (1999). At each step i we have an estimator $\hat{S}_i$ and we add an estimator $\hat{h}$ which will be originated by a decision tree and such that minimizes the loss function. Then, our estimator at the stage $i+1$ will be

$$\hat{S}_{i+1}(t|x) = \hat{S}_i(t|x) + \hat{h}(t|x). \tag{8}$$

3.7 DeepSurv

DeepSurv, proposed by Katzman et al. (2018), is a nonlinear version of Cox proportional hazards model. Cox proportional hazards is a semiparametric model that calculates the effects of observed covariates on the risk of an event occurring and it supposes that this

risk is a linear combination of the covariates. However, this linear assumption could not be accurate to the reality and too simplistic. DeepSurv proposes to use deep neural networks to learn a nonlinear relationship between covariates/features and an individual's risk of failure. DeepSurv is a multi-layer perceptron and it estimates for each feature x the risk function $\hat{r}_\theta(x)$ parametrized by the weights of the network θ. This function is the same function $\sum_{i=1}^{m} h_i(x_i - \bar{x_i})$ presented in the Cox proportional hazard model, but the difference is that in this case it is not assumed to be linear and it is given by minimizing the loss function of the neural network

$$l(\theta) = -\frac{1}{N} \sum_{i:\delta_i=1} \left(\hat{r}_\theta(x_i) - \log \sum_{j \in \mathcal{R}(Y_i)} e^{\hat{r}_\theta(x_j)} \right) + \lambda ||\theta||^2, \tag{9}$$

where λ is a regularization parameter, N is the number of uncensored subjects and $\mathcal{R}(t)$ is the set of subjects at risk at time t.

4 Results and Conclusions

We compared all the models described for the two datasets through the concordance index[1]. For each dataset, the experiment we performed is the following: we choose 25 different seeds for splitting the dataset, this generates 25 different partitions between training and validation sets. Then we run the model 25 times (one for each partition) and we make a boxplot with the distribution of the concordance indexes obtained. In the figures, we can observe the median of the obtained concordance indexes represented by the red lines and the average represented by the red triangles.

Figure 1 shows the comparison of the concordance indexes for PBC dataset where we can appreciate that random survival forest model fitted with a random search of the hyperparameters outperforms the other models. Figure 2 shows the comparison of the concordance indexes for GBCSG2 dataset and we can see that random survival forest with adaptive nearest neighbors outperforms the other models.

Furthermore, we can observe that traditional methods performed reasonably well for the small dataset PBC (see Cox proportional hazards with randomized search), but they underperformed against machine learning methods for the larger dataset (GBCSG2). We can also observe that the deep learning method (Deepsurv) did not perform better than random survival forest model and therefore the progress made by deep learning in other areas (computer vision, NLP, etc.) has not yet been replicated for time-to-event analysis.

Classical methods for predicting survival time are easier to interpret and to analyze the way in which each covariate/feature has an influence in the model. For the case of

[1]Our code is available at: https://github.com/CamilaFernandez8/Experimental-comparison-for-time-to-event-analysis-through-the-concordance-index

PBC dataset, random survival forest with random search outperforms Cox proportional hazards with random search by less than 1% while in GBCSG2 the method RSF with adaptive nearest neighbor increase the performance by 2.5% with respect to randomized search Cox proportional hazards model. Therefore, if this increment in performance is significant enough to compensate for the loss of easier interpretation of the model will highly depend on the application.

Acknowledgment

The work presented in this paper has been partially carried out at LINCS (www.lincs.fr).

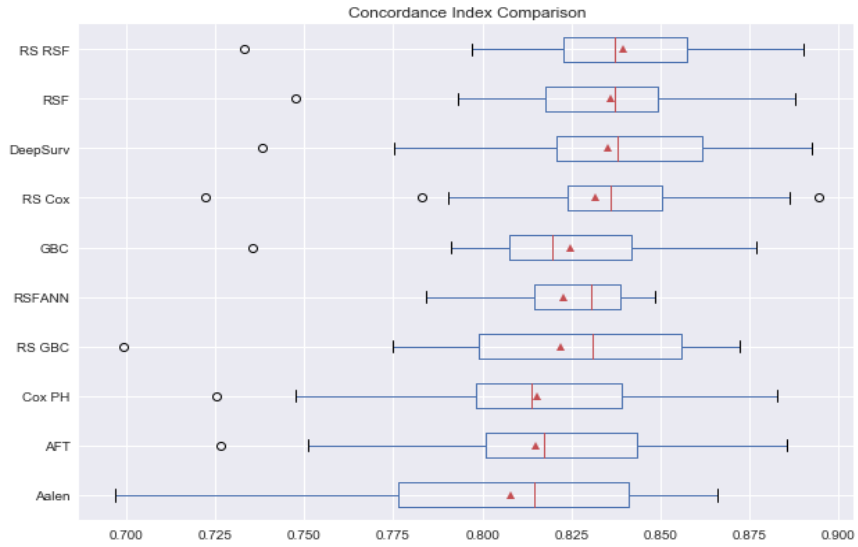


Figure 1: Concordance index comparison for PBC dataset.

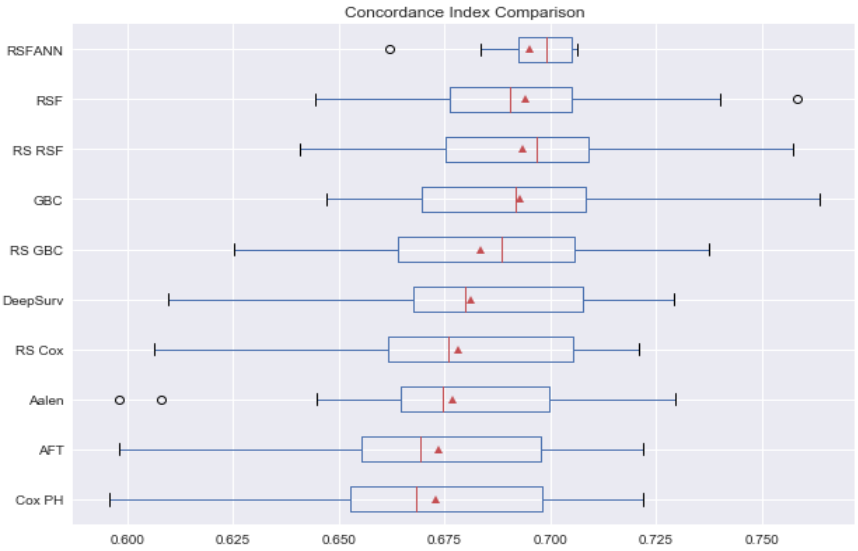


Figure 2: Concordance index comparison for GBCSG2 dataset.

Bibliography

Aalen, O. O. (1989). A linear regression model for the analysis of life times. *Statistics in Medicine*, vol. 8, pp. 907–925.
Breiman, L. (2001). Random Forest. *Machine Learning* 45 5-32.

Chen, G. (2019). Nearest Neighbor and Kernel Survival Analysis: Nonasymptotic Error Bounds and Strong Consistency Rates. arXiv preprint arXiv:1905.05285.
Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society*, Series B. 34 (2): 187–220.
Davidson-Pilon, C., et al. 2020, CamDavidsonPilon/lifelines:v0.23.9, doi: 10.5281/zenodo.805993.
Friedman, J.H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics,* pp.1189-1232
Harrell, F.E Jr., Lee, K. L., Mark, D. B. (1996), Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, vol. 15, no. 4, pp. 361–87.
Ishwaran, H., Kogalur, U. B., Blackstone, E. H., Lauer, M. S. (2008). Random survival forests. *Annals of Applied Statistics*, vol. 2, no. 3, pp. 841–860.
Katzman J.L., Shaham U., Cloninger A., Bates J., Jiang T., and Kluger Y. (2018). DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC medical research methodology* 18.1 (2018):24
Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python, JMLR 12, pp. 2825–2830.
Pölsterl, S. (2019). Scikit-survival:v0.11, doi:10.5281/zenodo.3352342.
Ridgeway, G. (1999). The state of boosting, Computer Science and Statistics, vol. 31, pp. 172–181.
Schumacher, M., Basert, G., Bojar, H., Huebner, K., Olschewski, M., Sauerbrei, W., Schmoor, C., Beyerle, C., Neumann, R.L.A. and Rauschecker H.F., for the German Breast Cancer Study Group (1994), Randomized 2 x 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. *Journal of Clinical Oncology*, 12, pp. 2086–2093.
Therneau, T., and Grambsch, P. (2000). Modeling Survival Data: Extending the Cox Model, Springer-Verlag, New York. ISBN: 0-387-98784-3.

Carte SOM profonde : Apprentissage joint de représentations et auto-organisation

Florent Forest [1,2] & Mustapha Lebbah [1] & Hanene Azzag [1] & Jérôme Lacaille [2]

[1] *LIPN, Université Sorbonne Paris Nord - 99 av J.-B. Clément, 93430 Villetaneuse*
[2] *Safran Aircraft Engines - Rond-point René Ravaud, 77550 Réau*
forest@lipn.univ-paris13.fr

Résumé. Dans la lignée des récentes avancées en apprentissage profond de représentations pour le clustering, ce travail (précédemment publié en anglais) présente le modèle DESOM (Deep Embedded SOM), combinant l'apprentisssage non supervisé de représentations et d'une carte auto-organisée de Kohonen (SOM). Le modèle, composé d'un auto-encodeur et d'une couche SOM, est optimisé conjointement, afin de régulariser l'espace latent et améliorer la performance de la carte SOM. Nous évaluons les performances de classification et de visualisation ainsi que les bénéfices de l'apprentissage joint.

Mots-clés. carte auto-organisée, clustering, apprentissage profond, auto-encodeur

Abstract. In the wake of recent advances in joint clustering and deep learning, we introduce the Deep Embedded Self-Organizing Map, a model that jointly learns representations and the code vectors of a self-organizing map. Our model is composed of an autoencoder and a custom SOM layer that are optimized in a joint training procedure, motivated by the idea that the SOM prior could help learning *SOM-friendly* representations. We evaluate SOM-based models in terms of clustering quality and unsupervised clustering accuracy, and study the benefits of joint training.

Keywords. self-organizing map, clustering, representation learning, autoencoder

1 Introduction

Après les succès des réseaux de neurones en apprentissage supervisé, de récents travaux se sont tournés vers l'apprentissage de représentations pour des tâches non supervisées, en particulier le clustering. Les algorithmes classiques sont basés sur des mesures de similarité qui se révèlent inefficaces en grande dimension. Une solution est de réduire d'abord la dimensionalité des données, puis d'effectuer le clustering dans un espace de basse dimension. La réduction est obtenue par des méthodes linéaires comme l'ACP ou non linéaires comme les auto-encodeurs profonds. Cette approche consiste, dans un premier temps, à optimiser un critère de perte d'information (en général via une erreur de reconstruction), puis dans un deuxième temps, optimiser un critère de partitionnement des données via un algorithme de clustering. À l'inverse, les approches dites de deep clustering (Song et al (2014), Xie et al (2015), Guo et al (2017), Yang et al (2016), Jiang

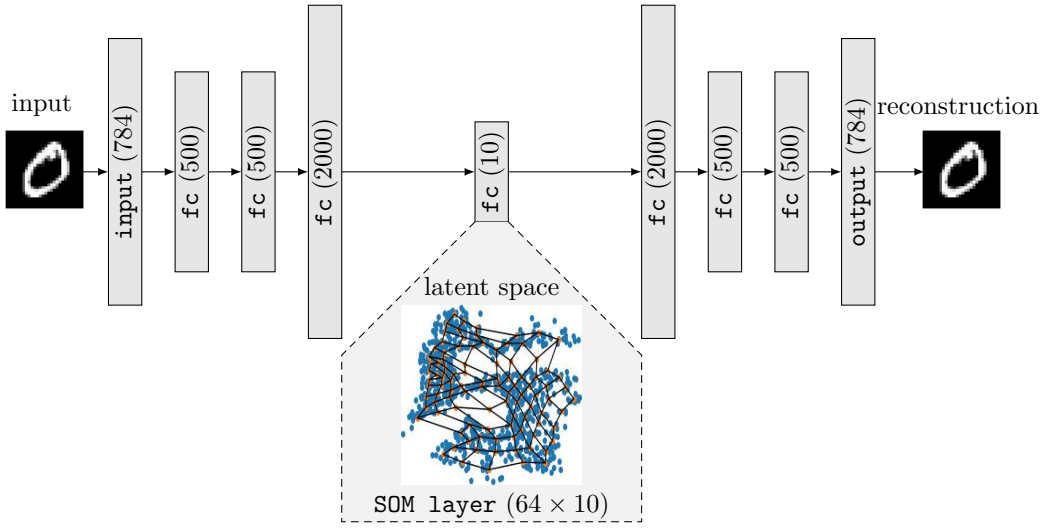


Figure 1: Architecture DESOM avec une carte 8×8.

et al (2017), Harchaoui et al (2018)) considèrent ces deux étapes comme une tâche jointe, et tentent d'apprendre un espace latent intégrant la structure de clusters.

La carte auto-organisée (SOM) (Kohonen, 1982) permet de classifier et visualiser des données mulivariées sur un grille de faible dimension, constituée de neurones chacun associé à un vecteur prototype de l'espace d'origine. Par contrainte topologique de l'apprentissage, des neurones voisins sur la carte correspondent à des prototypes proches dans l'espace de grande dimension.

Ce travail présente DESOM (Deep Embedded SOM) (Forest et al (2019)), un modèle apprenant conjointement une SOM et un espace latent de projection. La transformaton entre les espaces d'origine et latent est représentée par un auto-encodeur (AE). Les prototypes sont situés dans l'espace latent et décodés au moment de la visualisation et interprétation. Cette approche possède certains avantages : (1) les représentations riches apprises via les auto-encodeurs améliorent les performances de SOM, (2) l'auto-organisation et l'apprentissage de représentations s'effectue comme une tâche jointe, améliorant la qualité de la classification et réduisant la durée d'entraînement. À notre connaissance, le seul travail similaire est SOM-VAE (Fortuin et al (2019)), avec d'importantes différences au niveau de l'espace latent (discret, alors que celui de DESOM est continu), du voisinage (fixe alors que nous utilisons un voisinage gaussien avec décroissance du rayon) et de l'auto-encodeur (variationnel au lieu de déterministe).

2 Proposition

L'architecture proposée est illustrée Fig. 1. La carte SOM est composée de K neurones, associés aux prototypes $\{\mathbf{m}_k\}_{1 \leq k \leq K}$. Soit $\delta(\cdot, \cdot)$ la distance topologique entre deux neu-

rones sur la carte. Nous adoptons une fonction voisinage gaussienne $\mathcal{K}^T(d) = e^{-d^2/T^2}$, dont le rayon est contrôlé par un paramètre de température T. Ce dernier décroît exponentiellement durant l'apprentissage. Les paramètres des réseaux encodeur et décodeur sont notés respectivement $\mathbf{W_e}$ et $\mathbf{W_d}$. $\mathbf{z}_i = \mathbf{f}_{\mathbf{W_e}}(\mathbf{x}_i)$ représente la projection d'un point $\mathbf{x}_i$ sur l'espace latent, et sa reconstruction via le décodeur est notée $\tilde{\mathbf{x}}_i = \mathbf{g}_{\mathbf{W_d}}(\mathbf{z}_i)$. Nous définissons une fonction coût composée de deux termes :

$$\mathcal{L}(\mathbf{W_e}, \mathbf{W_d}, \mathbf{m}_1, \ldots, \mathbf{m}_K, \chi) = \mathcal{L}_r(\mathbf{W_e}, \mathbf{W_d}) + \gamma \mathcal{L}_{som}(\mathbf{W_e}, \mathbf{m}_1, \ldots, \mathbf{m}_K, \chi)$$

Le premier terme $\mathcal{L}_r$ est une erreur quadratique moyenne de reconstruction. Le second terme $\mathcal{L}_{som}$ est la fonction coût de SOM. Celle-ci dépend des $\{\mathbf{m}_k\}$ et de la fonction d'affectation $\chi(\mathbf{z}) = \text{argmin}_k ||\mathbf{z} - \mathbf{m}_k||^2$.

$$\mathcal{L}_{som} = \sum_i \sum_{k=1}^{K} \mathcal{K}^T \left(\delta(\chi(\mathbf{f}_{\mathbf{W_e}}(\mathbf{x}_i)), k)\right) ||\mathbf{f}_{\mathbf{W_e}}(\mathbf{x}_i) - \mathbf{m}_k||^2$$

Le coefficient γ définit le poids relatif du coût de reconstruction et de SOM. Notre procédure jointe fixe l'affectation χ (non diférentiable) entre chaque étape d'optimization. Ainsi, il est possible de définir des coefficients constants $w_{i,k} \equiv \mathcal{K}^T \left(\delta(\chi(\mathbf{f}_{\mathbf{W_e}}(\mathbf{x}_i)), k)\right)$, grâce auxquels les dérivées partielles de la fonction coût s'expriment simplement. Le parcours des gradients est illustré Fig. 2.

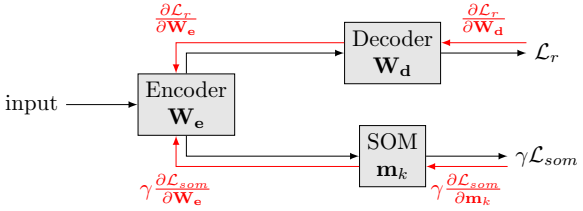


Figure 2: Parcours des gradients de DESOM.

3 Implémentation

DESOM est implémenté sous le framework Keras et le code est open-source[1]. La principale nouveauté est une couche SOM, paramétrée par une matrice $K \times L$, avec K le nombre de neurones et L la dimensionalité latente. Cette couche calcule les distances euclidiennes entre les données en entrée et les prototypes, ce qui permet d'exprimer la fonction coût SOM comme une somme pondérée par les poids $w_{i,k}$. La procédure d'entraînement est détaillée dans l'algorithme 1.

[1] `https://github.com/FlorentF9/DESOM`

Input: jeu d'apprentissage; topologie SOM; T_{max}; T_{min}; *iterations*; *batchSize*
Output: poids de l'AE $\mathbf{W_e}$, $\mathbf{W_d}$; vecteurs référents SOM $\{\mathbf{m}_k\}$
Initialiser AE (Glorot uniform) et SOM (échantillon aléatoire) ;
for $iter = 1, \ldots, iterations$ **do**
  $T \leftarrow T_{max} \left(T_{min}/T_{max}\right)^{iter/iterations}$;
  Charger le prochain batch d'entraînement ;
  Calculer les distances à la carte SOM et les poids $w_{i,k}$;
  Effectuer une étape de descente de gradient sur le batch ;
end

Algorithm 1: Procédure d'entraînement de DESOM

4 Expériences

Les expériences utilisent des jeux de données classiques en classification automatique : MNIST, Fashion-MNIST (images) et Reuters-10k (texte). Les clusterings obtenus sont évalués de manière quantitative par la pureté et le NMI (information mutuelle normalisée). La discrimation des classes est évaluée en classifiant les prototypes par k-means, puis en mesurant la précision de classification non supervisée par rapport aux classes cibles. Qualitativement, nous vérifions l'organisation topologique et la qualité visuelle de la carte.

Nous comparons **minisom**, une implémentation standard de SOM[2]; **kerasom**, notre implémentationn de SOM sous Keras (i.e. DESOM sans auto-encodeur); **AE+minisom** et **AE+kerasom**, avec AE and SOM entraînés séparément; finalement, **DESOM.** Nous incluons aussi SOM-VAE.

4.1 Paramètres d'entraînement

Chaque modèle est entraîné 10000 itérations avec une taille de batch de 256 et l'optimiseur Adam. Les températures initiale et finale sont $T_{max} = 8.0$ and $T_{min} = 0.1$. L'architecture de l'auto-encodeur ($[500, 500, 2000, 10]$) et la taille de la carte (8×8) sont choisis pour être comparables avec de précédents travaux. L'hyper-paramètre γ est fixé empiriquement à 0.001, sans validation croisée pour rester dans un cadre non supervisé. Une valeur trop élevée mène à des solutions dégénérées pour l'encodeur, dû au fait que le coût SOM est plus facilement optimisé que l'erreur de reconstruction. De plus, le modèle est peu sensible à la valeur de γ tant qu'elle reste dans cet ordre de grandeur. Un pré-entraînement de l'AE est bénéfique dans la plupart des approches de clustering profond. Cependant, la fonction coût SOM produit de forts gradients en début d'entraînement, perturbant l'encodeur. Par conséquent, aucun pré-entraînement n'est utilisé ici.

[2] `https://github.com/JustGlowing/minisom`

Méthode	MNIST		Fashion-MNIST		REUTERS-10k	
	pur	nmi	pur	nmi	pur	nmi
minisom (8×8)	0.637	0.430	0.646	0.494	0.690	0.230
kerasom (8×8)	0.826	0.565	0.717	0.512	0.697	0.324
AE+minisom (8×8)	0.871	0.616	0.734	0.531	0.690	0.235
AE+kerasom (8×8)	**0.939**	**0.661**	**0.764**	**0.539**	0.777	0.306
SOM-VAE (8×8)	0.868	0.595	0.739	0.520	-	-
DESOM (8×8)	**0.939**	0.657	0.752	**0.538**	**0.849**	**0.381**

Table 1: Purité et NMI (moyenne sur 10 entraînements). Meilleur résultat et résultats statistiquement équivalents (p-value > 0.05) en gras.

Méthode	MNIST	Fashion-MNIST	REUTERS-10k
AE+kerasom (8×8) + k-means	**76.06**	44.87	36.61
DESOM (8×8) + k-means	**76.11**	**56.02**	**57.18**

Table 2: Précision de classification non supervisée (%) (moyenne sur 10 entraînements).

4.2 Résultats quantitatifs et qualitatifs

Les résultats (Tab. 1) confirment l'intérêt de la réduction de dimension par AE. Les modèles les plus performants sont AE+kerasom et DESOM. L'apprentissage joint n'améliore pas systématiquement la pureté et le NMI mais reste a minima comparable, et plus rapide à entraîner. Étonnament, kerasom obtient de meilleurs résultats que minisom, suggèrant que l'optimisation par descente de gradient stochastique améliore l'entraînement de SOM, découverte faite aussi par Fortuin et al (2019). Enfin, DESOM surclasse son concurrent direct SOM-VAE. En terme de discrimination des classes (Tab. 2), DESOM obtient les meilleurs scores, montrant que l'apprentissage joint a permis d'apprendre un espace latent régularisé plus adapté pour la classification des prototypes de la carte.

Les visualisations des prototypes décodés de DESOM (Fig. 3) forment des régions continues et bien organisées correspondant aux différentes classes. De plus, les images prototypes apprises par une SOM standard sont floues, à cause du calcul de moyennes dans l'espace original, ce qui n'est pas le cas pour DESOM.

5 Conclusion et travaux futurs

DESOM est la première approche qui entraîne conjointement un auto-encodeur et une carte SOM dans un espace latent continu. La carte obtenue est organisée, compétitive en terme de clustering, et ne nécessite pas de pré-entraînement. Des travaux sont encore nécessaires pour étudier les hyper-paramètres, et le modèle pourrait être étendu en une variante variationnelle afin d'obtenir un modèle génératif.

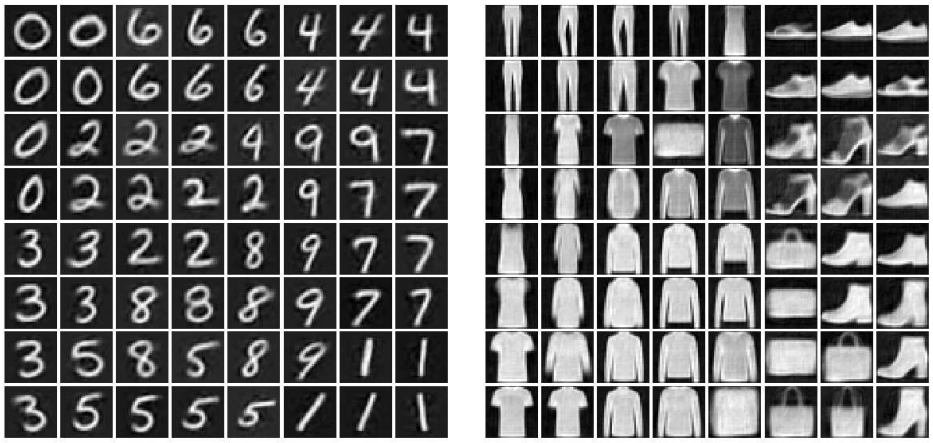


Figure 3: Carte DESOM de MNIST (gauche) et Fashion-MNIST (droite).

Bibliographie

Forest, F., Lebbah, M., Azzag, H. et Lacaille, J. (2019). Deep Embedded SOM: Joint Representation Learning and Self-Organization, *ESANN*.

Song, C., Huang, Y., Liu, F., Wang, Z. et Wang, L. (2014). Deep auto-encoder based clustering, *Intelligent Data Analysis*.

Xie, J., Girshick, R. et Farhadi, A. (2015), Unsupervised Deep Embedding for Clustering Analysis, *ICML*.

Guo, X., Gao, L., Liu, X. et Yin, J. (2017), Improved deep embedded clustering with local structure preservation, *IJCAI*.

Yang, B., Fu, X., Sidiropoulos, N. D. et Hong, M. (2016), Towards K-means-friendly Spaces: Simultaneous Deep Learning and Clustering, *ICML*.

Jiang, Z., Zheng, Y., Tan, H., Tang, B. et Zhou, H. (2017) Variational Deep Embedding : An Unsupervised and Generative Approach to Clustering, *IJCAI*.

Harchaoui, W., Mattei, P., Alamansa, A. et Bouveyron, C. (2018), Wasserstein Adversarial Mixture Clustering.

Kohonen, T. (1982), Self-organized formation of topologically correct feature maps, *Biological Cybernetics*.

Fortuin, V., Hüser, M., Locatello, F., Strathmann, H. et Rätsch, G. (2019), Deep Self-Organization: Interpretable Discrete Representation Learning on Time Series, *ICLR*.

Algorithmes stochastiques pour le transport optimal appliqué au traitement de données de cytométrie en flux.

Paul Freulon [1] & Jérémie Bigot [1] & Boris P. Hejblum [2]

[1] *Institut de Mathématiques de Bordeaux & CNRS (UMR 5251), Université de Bordeaux, Talence, France - paul.freulon@math.u-bordeaux.fr ; jeremie.bigot@math.u-bordeaux.fr*
[2] *Université de Bordeaux – ISPED, Inserm Bordeaux Population Health U1219 – Inria BSO SISTM, Bordeaux, France ; Vaccine Research Institute, Créteil, France - boris.hejblum@u-bordeaux.fr*

Résumé. L'analyse automatisée de données de cytométrie en flux est un domaine de recherche actif. Nous proposons une nouvelle méthode utilisant le transport optimal pour l'estimation directe des proportions de différentes populations cellulaires présentes au sein d'un échantillon biologique à partir de mesures de cytométrie en flux. Notre méthode s'appuie sur la distance de Wasserstein pour évaluer la proximité entre deux échantillons caractérisés par une série de mesures de cytométrie en flux, prenant ainsi en compte le potentiel décalage d'une même population cellulaire entre les échantillons (causés par la variabilité technique de la cytométrie en flux). Pour approcher le calcul de cette distance et de son gradient, nous utilisons la procédure d'optimisation stochastique de Robbins-Monro. La distance de Wasserstein permet également d'estimer les pondérations associées à chaque terme dans un modèle de mélange entre une distribution source (dont la segmentation en sous-populations cellulaires est connue) et une distribution cible non-segmentée.

Mots-clés. Algorithmes stochastiques, Apprentissage supervisé, Cytométrie en flux, Distance de Wasserstein, Modèles de mélange, Régularisation, Transport optimal

Abstract. The automated analysis of flow cytometry measurements is an active research field. We introduce a new methodology using optimal transport to directly estimate the different cell population proportions from a biological sample characterized with flow cytometry measurements. We rely on the Wasserstein metric to compare cytometry measurements from different samples, thus accounting for possible mis-alignment of a given cell population across sample (due to technical variability from the flow cytometry technology). To approximate the Wasserstein metric and its gradient, we use the Robbins-Monro algorithm. In this work, the Wasserstein metric is used to estimate the weights of a mixture model between a source distribution (with known segmentation into cell sub-populations) and a target distribution with unknown segmentation.

Keywords. Flow cytometry, Mixture model, Optimal Transport, Regularization, Stochastic algorithms, Supervised learning, Wasserstein metric

1 Introduction

La cytométrie en flux est une technologie de mesure à haut débit qui permet de rapidement caractériser un grand nombre de cellules selon leur morphologie et selon de nombreux marqueurs intra- et extra-cellulaires à partir d'un échantillon biologique. C'est une technologie qui est par exemple utilisée pour le suivi de patients atteints du VIH afin de mesurer leur taux de lymphocyte CD4. En pratique, des fluorochromes sont couplés à des molécules complémentaires des marqueurs cellulaires ciblés, qui sont ensuite mis en présences des cellules extraites de l'échantillon. Le cytomètre analyse le rayonnement lumineux des cellules une par une afin de déterminer si les marqueurs biologiques ciblés sont à chaque fois présents (les fuorochromes se sont attachés à la cellule à proportion de l'abondance du marqueur cellulaire) ou absents. Du point de vue de la modélisation statistique, le passage des n cellules d'un échantillon biologique dans le cytomètre génère une série de n observations : $X_1, ..., X_n$ appartenant à $\mathbb{R}^d$, où d est compris entre 4 et 50. Chaque observation $X_i \in \mathbb{R}^d$ correspond au passage d'une cellule i devant le rayon laser du cytomètre, et la valeur $X_i^{(m)}$ correspond à l'intensité de la fluorescence associée au marqueur m.

L'approche de référence pour l'analyse des données de cytométrie est l'analyse manuelle dans le but de quantifier des groupes de cellules (étape dite de « segmentation » ou « labellisation »). Cette méthode consiste à entourer par des formes géométriques tracée à la main des sous-populations cellulaires d'intérêt à partir de projections successives des données en 2 dimensions. Ce processus étant fastidieux, peu reproductible et couteux, un travail important a été effectué ces dernières années pour proposer des méthodes automatiques de segmentation des données de cytométrie (Commenges et al., 2018; Hejblum et al., 2019). Ici, nous considérons directement le problème d'intérêt dans le traitement des données de cytometrie en flux, à savoir l'estimation des proportions de sous-populations cellulaires pour des données de cytométrie non-labellisées. Pour cela nous nous plaçons dans un cadre d'apprentissage supervisé à partir de la connaissance d'une segmentation sur un jeu de données de référence.

Notre approche s'appuyant sur la distance de Wasserstein, nous commençons par introduire cette distance en Section 2 ainsi que son approximation via la procédure d'optimisation stochastique de Robbins-Monro. Ensuite, nous proposons un estimateur des proportions par classe pour un nouveau jeu de données non-labellisé. Nous présentons enfin quelques résultats de l'application de notre méthode sur des données réelles de cytométrie en flux en Section 3.

2 Distance de Wasserstein et algorithmes stochastiques pour le transport optimal

2.1 Distance de Wasserstein

Le transport optimal est un outil permettant de comparer deux mesures de probabilités entre elles. Cette comparaison se fait en cherchant à déplacer à moindre coût une mesure source α vers une mesure source β. L'intérêt de cette distance pour l'analyse de données est maintenant établie, et Peyré et al. (2019) détaille de nombreuses applications du transport optimal pour l'apprentissage automatique. Dans le cadre de ce travail, nous nous sommes restreints au transport optimal entre deux mesures discrètes. En notant $\alpha = \sum_{i=1}^{I} a_i \delta_{x_i}$ la mesure source, et $\beta = \sum_{j=1}^{J} b_j \delta_{y_j}$ la mesure cible, le problème de Kantorovich consiste en la résolution du problème de minimisation suivant :

$$W(\alpha, \beta) = \min_{P \in U(a,b)} \sum_{i,j} C_{i,j} P_{i,j} = \min_{P \in U(a,b)} \langle C, P \rangle.$$

où $U(a,b) = \{P \in \mathbb{R}_+^{I \times J} : P.1_m = a \text{ et } P^T.1_n = b\}$ désigne l'ensemble des plans de transport de α vers β, et $C \in \mathbb{R}_+^{n \times m}$ est la matrice de coût où le coefficient $C_{i,j}$ représente le coût d'envoyer une unité de matière de x_i vers y_j. Lorsque la matrice de coût est définie à partir d'une distance, le problème de Kantorovich permet de définir une distance entre mesures de probabilités appelée distance de Wasserstein. Dans toute la suite de ce travail la matrice de coût sera définie à partir du carré de la distance euclidienne.

Le facteur limitant l'utilisation de la distance de Wasserstein est souvent son coût calculatoire élevé. En effet, si les deux mesures α et β ont toutes les deux un support de taille N, alors la résolution du problème de Kantorovich nécessite $O(N^3 \log(N))$ opérations. Afin de limiter ce coût calculatoire, Cuturi (2013) propose de régulariser le problème de Kantorovich par l'ajout de l'entropie $H : \mathbb{R}_+^{n \times m} \to \mathbb{R}_+^{n \times m}$ définie par $H(P) = \sum_{i,j} P_{i,j}(\log(P_{i,j}) - 1)$ pour $P \in \mathbb{R}_+^{*\, n \times m}$. Pour $\varepsilon > 0$, on appelle alors « problème régularisé » le problème de minimisation suivant :

$$W^{\varepsilon}(\alpha, \beta) = \min_{P \in U(a,b)} \langle C, P \rangle + \varepsilon H(P) \tag{1}$$

et distance de Wasserstein régularisée la quantité $W^{\varepsilon}(\alpha, \beta)$. L'ajout de l'entropie permet ainsi de calculer une approximation $W^{\varepsilon}(\alpha, \beta)$ de la distance de Wasserstein $W(\alpha, \beta)$ en seulement $O(N^2 \log(N))$ opérations (par un algorithme dit « de Sinkhorn »).

2.2 Algorithmes stochastiques

Soit $\varepsilon > 0$. Genevay et al. (2016) montrent alors que le calcul de $W^{\varepsilon}(\alpha, \beta)$ peut se reformuler comme un problème de maximisation d'une espérance :

$$W_{\varepsilon}(\alpha, \beta) = \max_{v \in \mathbb{R}^J} \mathbb{E}[h_{\varepsilon}(X, v)] \tag{2}$$

où X est une variable aléatoire de loi α, et pour tout $x \in \mathbb{R}^d$ et tout $v \in \mathbb{R}^J$:

$$h_\varepsilon(x, v) = \sum_{j=1}^{J} b_j v_j - \varepsilon \log \left(\sum_{j=1}^{J} \exp \left(\frac{v_j - c(x, y_j)}{\varepsilon} \right) \right) - \varepsilon.$$

Cette reformulation avec une espérance permet d'utiliser l'optimisation stochastique et en particulier l'algorithme de Robbins-Monro pour résoudre ce problème. D'après Bercu and Bigot (2018), la formulation (2) permet de déduire un estimateur $\widehat{W}_n$ de $W^\varepsilon(\alpha, \beta)$. De plus, cette procédure permet d'éviter le stockage mémoire de l'intégralité de la matrice de coût car l'évaluation de la fonction h_ε ne requiert qu'une seule ligne de la matrice de coût. L'échantillonnage ainsi réalisé intrinséquement à chaque itération de l'algorithme stochastique se fait selon la distribution empirique des observations, ce qui nous semble pertinent dans un cadre statistique.

3 Application aux données de cytométrie en flux

3.1 Estimation des proportions cellulaires

Considérons un premier jeu de données $X_1^s, ..., X_I^s$, dit « jeu de données source », pour lequel les observations se répartissent en K classes $C_1, ..., C_K$. Nous supposons disponible pour ce jeu de données une segmentation en K sous-populations cellulaires (obtenue par une approche manuelle par exemple) – et donc les proportions par classe aussi. Considérons un second jeu de données $X_1^t, ..., X_J^t$, dit « jeu de données cible », pour lequel l'information de classe n'est en revanche pas disponible. Notre méthode propose un estimateur $\widehat{h} \in \Sigma_K = \{h \in \mathbb{R}_+^K : \sum_{k=1}^K h_k = 1\}$ des proportions par classe dans le jeu de données cible à partir de celles connues dans le jeu de données source en repondérant la distribution source afin de la rapprocher, au sens de la distance de Wasserstein, de la distribution cible, en suivant l'approche de Redko et al. (2018). On note $\alpha = \frac{1}{I} \sum_{i=1}^I \delta_{X_i^s}$ la distribution source et $\beta = \frac{1}{J} \sum_{j=1}^J \delta_{X_j^t}$ la distribution cible. La classification des observations étant disponible pour le jeu de données cible, nous pouvons réécrire la distribution source $\alpha = \sum_{k=1}^K \alpha_k$ où le terme α_k représente la distribution associée à la classe k, c'est à dire à une certaine population cellulaire dans le cadre de la cytométrie. À partir de cette écriture comme un modèle de mélange de la distribution source, nous pouvons repondérer la distribution α avec les proportions $h \in \Sigma_K$: $\alpha(h) = \sum_{k=1}^K h_k \alpha_k$. Nous proposons alors d'estimer les proportions par classe dans le jeu de données cible par :

$$\widehat{h} = \underset{h \in \Sigma_K}{\arg\min}\, W^\varepsilon(\alpha(h), \beta). \tag{3}$$

Le problème d'optimisation (3) est résolu grâce une méthode de gradient, où le gradient de la fonction $F : h \mapsto W^\varepsilon(\alpha(h), \beta)$ est estimé à partir des méthodes stochastiques décrites précédemment.

3.2 Résultats sur les données *HIPC*

Nous analysons les données de cytométrie en flux générées par le programme *HIPC* (Brusic et al., 2014). Ces mesures ont été réalisées dans 7 centres différents à partir d'échantillons biologiques de lymphocytes T de trois patients différents, avec 3 réplicats techniques chacun. Il en resulte 62 jeux de données disposant d'une classification cellulaire obtenue manuellement – dont on déduit directement donc des proportions cellulaires. Ces proportions nous servirons de référence pour évaluer les performances de notre estimateur.

Nous présentons les résultats de notre méthode lorsque les données de cytométrie sont réparties en deux grandes classes : CD4 et CD8. À partir d'un jeu de données de cytométrie analysé dans le centre de Stanford, il s'agit d'estimer les proportions de CD4 et de CD8 dans un autre échantillon biologique pour lequel les mesures de cytométrie ont été réalisées dans un autre centre. Les résultats présentés dans la Figure 1 indique clairement l'intérêt de l'approche par minimisation de la fonctionnelle (3) pour l'estimation des proportions de CD4 et CD8 dans le jeu de données cible.

4 Conclusion et Perspectives

La méthode que nous proposons pour traiter automatiquement les données de cytométrie a l'avantage de ne nécessiter qu'un seul jeu d'entraînement sans passer par une phase de classification des observations. Les proportions des différentes populations cellulaires est l'élément d'intérêt clinique. Une perspective d'amélioration est l'utilisation de plusieurs jeux de données sources pour lesquels l'information de classe est accessible.

Références

Bercu, B. and Bigot, J. (2018). Asymptotic distribution and convergence rates of stochastic algorithms for entropic optimal transportation between probability measures. *arXiv preprint arXiv :1812.09150*.

Brusic, V., Gottardo, R., Kleinstein, S. H., Davis, M. M., Hafler, D. A., Quill, H., Palucka, A. K., Poland, G. A., Pulendran, B., Reinherz, E. L., et al. (2014). Computational resources for high-dimensional immune analysis from the human immunology project consortium. *Nature biotechnology*, 32(2) :146.

Commenges, D., Alkhassim, C., Gottardo, R., Hejblum, B., and Thiébaut, R. (2018). Cytometree : A binary tree algorithm for automatic gating in cytometry analysis. *Cytometry Part A*, 93(11) :1132–1140.

Cuturi, M. (2013). Sinkhorn distances : Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300.

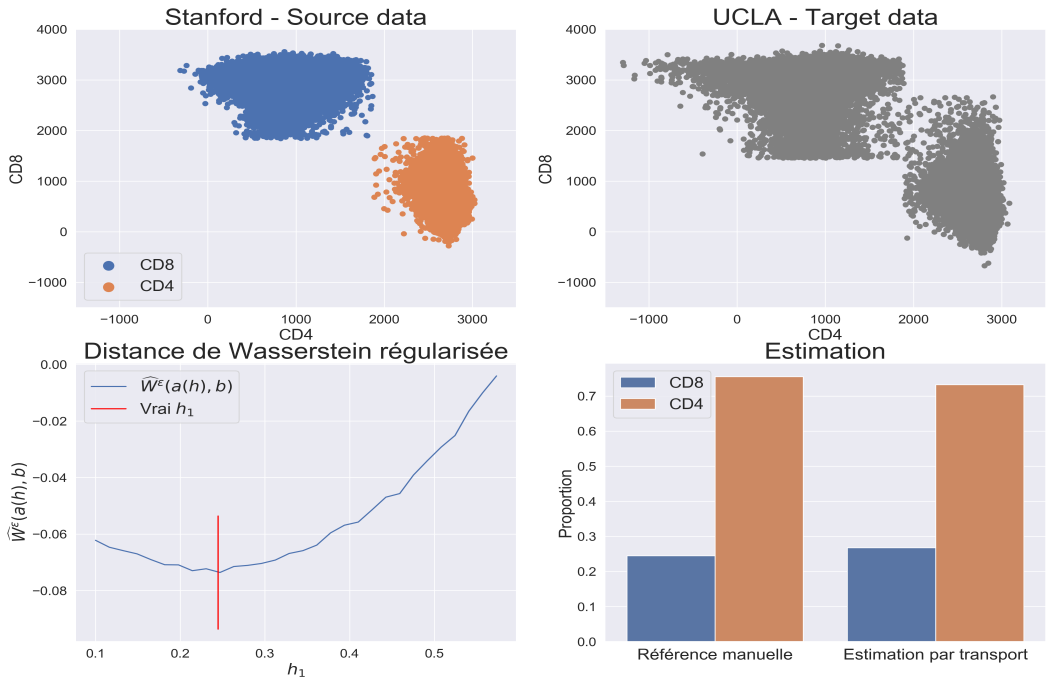


FIGURE 1 – En haut à Gauche : Données sources, la classification des observations est connue - En haut à droite : Données cibles non-labellisées. En bas à gauche : Evolution de la fonctionnelle $h_1 \mapsto W^{\varepsilon}(\alpha(h), \beta)$ pour $h = (h_1, 1 - h_1)$ en fonction de la pondération h_1 associée à la classe des CD8 dans le jeu de données source. En bas à droite : Résultats. Le vecteur des proportions obtenu par une analyse manuelle est $h = (0.245, 0.755)$, avec notre méthode, nous obtenons $\hat{h} = (0.268, 0.732)$

Genevay, A., Cuturi, M., Peyré, G., and Bach, F. (2016). Stochastic optimization for large-scale optimal transport. In *Advances in neural information processing systems*, pages 3440–3448.

Hejblum, B. P., Alkhassim, C., Gottardo, R., Caron, F., and Thiébaut, R. (2019). Sequential Dirichlet process mixtures of multivariate skew t-distributions for model-based clustering of flow cytometry data. *The Annals of Applied Statistics*, 13(1) :638–660.

Peyré, G., Cuturi, M., et al. (2019). Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6) :355–607.

Redko, I., Courty, N., Flamary, R., and Tuia, D. (2018). Optimal transport for multi-source domain adaptation under target shift. *arXiv preprint arXiv :1803.04899*.

Data annotation with active learning: application to environmental surveys

Chloé Friguet[1], Romain Dambreville[2], Ewa Kijak[3], Mathieu Laroze[4] & Sébastien Lefèvre[5]

[1] *Univ. Bretagne-Sud, UMR 6074, IRISA, Vannes, France - chloe.friguet@irisa.fr*
[2] *WIPSEA, Rennes, France - romain.dambreville@wipsea.fr*
[3] *Univ. Rennes, UMR 6074, IRISA, Rennes, France - ewa.kijak@irisa.fr*
[4] *Univ. Rennes, UMR 6074, IRISA et WIPSEA, Rennes, France - mathieu.laroze@irisa.fr*
[5] *Univ. Bretagne-Sud, UMR 6074, IRISA, Vannes, France - sebastien.lefevre@irisa.fr*

Résumé. Une procédure d'apprentissage actif est proposée pour réduire le coût d'annotation d'images aériennes pour des suivis environnementaux. La sélection des instances à étiqueter à chaque étape du processus actif est contrainte à l'appartenance à un groupe, une image (ou une partie d'image) dans notre cas. Un score pour classer les images et identifier celle qui doit être annotée à chaque itération est défini, en fonction de l'incertitude et des performances de détection du classifieur. Les performances de plusieurs stratégies concernant le gain d'interaction avec l'utilisateur sont discutées à partir d'une expérience sur des données d'images réelles collectées pour une étude environnementale.
Mots-clés. Apprentissage actif, annotation d'images aériennes, suivis environnementaux

Abstract. An active learning framework is introduced to deal with reducing the annotation cost for aerial images in environmental surveys. The selection of the queried instances at each step of the active process is here constrained by requiring that they belong to a group, an image (or a part of it) in our case. A score to rank the images and identify the one that should be annotated at each iteration is defined, based on both classifier uncertainty and performances. The performances of several strategies regarding the interaction gain are discussed based on an experiment on real image data collected for an environmental survey.
Keywords. Active learning, aerial images labelling, environmental surveys

1 Active learning and selective sampling

Machine Learning (ML) aims at deriving algorithms that can automatically learn from available data and make predictions. Active Learning (AL) is related to semi-supervised ML in which a learning algorithm is able to interact with the user to get some information about the label of new data during the training step. It is motivated by situations in which it is easy to collect unlabelled data but costly (time, money, tedious task) to (manually) obtain their labels. It stems from the idea that we should only acquire labels that actually improve our ability to make accurate predictions. More formally, a supervised model Θ

is trained incrementally on a training dataset X. A query criterion Q searches over the unlabelled dataset $\mathcal{U}$ and queries an oracle O to get a label feedback for a selected instance $\mathbf{x}^*$. The new labelled instance is added to the labelled dataset $\mathcal{L}$ and removed from $\mathcal{U}$. The model Θ is then re-trained on the augmented labelled dataset and the process is repeated until an ending criteria is met. Instances that are more useful than others for some performances have to be identified to create an optimal training dataset: well chosen, fewer representative instances are needed to achieve a similar performance. This selection process has been investigated as *selective sampling* [5]. The importance of an instance is related to a high level of both the information and uncertainty relatively to the trained model, considering therefore a trade-off between *informativeness* (ability to reduce the uncertainty of a statistical model) and *representativeness* (ability to represent the whole input data space) of the selection process [1].

2 Active learning to ease annotation

Motivating context: labelling objects in aerial images Nowadays, remote sensing technologies greatly ease environmental assessment over large study areas using aerial images, e.g. for monitoring and counting animals or ships. In the fields of both machine learning and image processing, many algorithms have been developed to tackle the complex task of object detection and to fasten and automate the counting processes. In practice, each image is divided into patches and object detection can be restated as a binary classification issue, to predict if the object of interest is on a patch or not, over the whole set of images $\mathcal{S}$. Most of these procedures are then supervised, and need to have prior ground truth available for each patch. However, manually labelling the patches requires, even for an expert, a time-consuming and tedious process. To assist the annotation process, an active learning approach can be used [6, 4], allowing interaction with the expert such as label confirmation or correction for each patch, at the query step. Note that in our context of environmental surveys, the data are unbalanced: there are only a few objects on each image, so most of patches are negatives. Moreover, the patch labelling is not easy even for an expert, and this query for a single patch usually requires some contextual information through the visualisation of the surrounding patches, that can therefore also be labelled at the same step. To address this challenge, the proposed method aims at assisting the annotation process by introducing an active learning procedure, querying the expert with groups of patches taken from the same image to ease annotation.

Proposed method The proposed active process is detailed with pseudo-code in Algo. 1. In our context, the initial data are a set $\mathcal{S} = \{I_0; \cdots ; I_K\}$ of $K+1$ images, each one being composed N patches mapped into a set of p features (see Figure 1). The input of the active selection algorithm is then a $(K+1)-$set of $N \times p$-matrix of instances $x_i^k \in \mathbb{R}^p$. The output is a class value for each instance denoted $y_i \in \mathcal{Y} = \{0; 1\}$, 1 being the positive class.

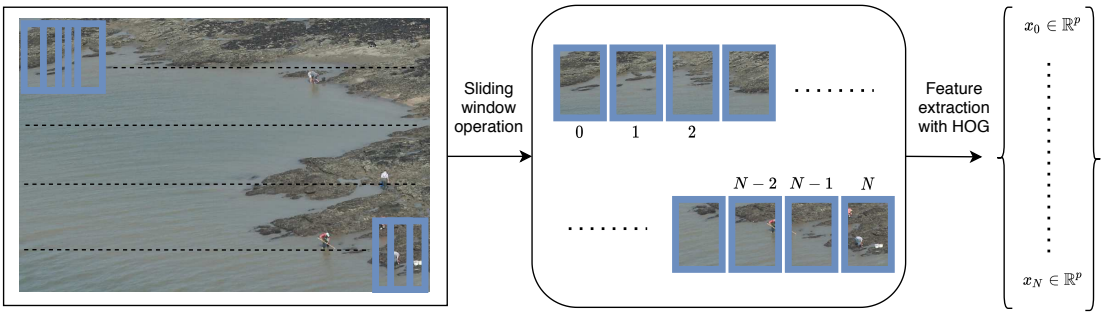


Figure 1: From image to patches and features extraction. The N patches from image k are considered together in the annotation process

Let us denote by $\mathcal{U}_\ell$ (resp $\mathcal{L}_\ell$) the set of unlabelled (resp. labelled) instances at iteration ℓ. At the beginning, $\mathcal{L}_0 = \{x_i^0; y_i^0\}_{i=1\cdots N}$ denotes the labelled data for initialisation and $\mathcal{U}_0$ is the set of remaining data to be labelled.

Algorithm 1: Main steps of the proposed active learning algorithm

Input : Initial training set $\mathcal{L}^0$
Input : Pool of unlabelled candidates $\mathcal{U}^0$
Initiate $\ell = 0$
repeat
  Train a classifier $\mathcal{C}$ with current training set $\mathcal{L}^\ell$
  Predict labels y for all instances in current unlabelled set $\mathcal{U}^\ell$ with $\mathcal{C}$
  for *each image k candidate in $\mathcal{U}^\ell$* **do**
    Compute the score s_k according to (1) and predictions by $\mathcal{C}$
  Rank candidates in $\mathcal{U}^\ell$ according to the scores s_k
  Select the most interesting image $I^* = \text{argmax}_k\{s_k\}$
  Query I^* to the oracle O to receive labels for all its instances: the oracle confirms positive labels and/or correct false negative ones.
  Add labelled instances to the training set: $\mathcal{L}^{\ell+1} = \mathcal{L}^\ell \cup I^*$ and $\mathcal{U}^{\ell+1} = \mathcal{U}^\ell \setminus I^*$.
  $\ell = \ell + 1$
until $\mathcal{S}$ *has been fully annotated*;

In usual algorithms, at each iteration, queried instances are selected one after another on criteria based on their informativeness or their uncertainty. In our context, the observed instances to be labelled are grouped into consistent sets (images) that we propose to process as a whole in the active query. This strategy necessitates therefore to define a global score to rate the relevance of each image to be the next one be queried to get all the labels of its patches. We propose a selective sampling query taking into consideration two main criteria: (i) Certain Instances predicted positive (resp. negative) by the classifier with a probability greater than a fixed threshold t_c (resp. lower than $1 - t_c$); (ii) Uncertain Instances predicted either positive or negative by the classifier with a probability lying in

$[0.5 \pm t_{unc}]$, for a fixed threshold t_{unc}. The global score is defined as the harmonic mean combining the number of positive certain instances PCI and all uncertain instances UI:

$$s = \frac{PCI \times UI}{PCI + UI} \tag{1}$$

The data are strongly imbalanced in our context: consequently, at each step, there are much more positive instances than negatives ones that are labelled. Learning from such data requires appropriate training strategies and metrics to assess the algorithm performances [2]. Besides, selecting the effective training subset at each step is not trivial and must be done very carefully to keep representative instances in the majority class. In the following, we consider several under-sampling strategies, based both on user interactions and classifier confidence: (UC) a balanced subset of uncertain instances; (UC+C) a balanced subset of uncertain and certain instances; (UC+C+EK) a balanced subset of uncertain and certain instances enriched with Extra-Knowledge (all instances corrected by the oracle that would not have been selected with the previous strategies: false positives, false negatives and positives with medium confidence).
In the imbalanced learning case, usual performance evaluation metrics are based on positives detection (precision, recall and F-score). In our context, we also consider a user-oriented criterion, based on the number of interactions. Interactions with the expert for the label annotation process can be either to correct the false positives or to add the missing positives (i.e. correct the false negatives). True positives and true negatives are validated implicitly. The evolution of the total number of interactions over the iterations of the algorithm is considered in the following as an evaluation criterion of the method: the method succeeds if, at the end of the algorithm, it is less than the number of interactions needed to annotate the full dataset (total number of true positives N^+). The gain in interaction G_K through our method is calculated by their difference.

3 Experiments

Settings The experiments are carried out on a set of aerial images of humans gathering shellfish on the seashore in the Natural Park of Morbihan (South Brittany, France) during spring tides [3]. The aim is to evaluate the number of people on the seashore in this period of high attendance and deduce the pressure of their activity on the environment. Table 1 reports the dataset characteristics and 3 images are shown in Figure 2. Instances (patches) are extracted with a sliding window of size 64×32 with a stride of 8 and Histogram of Oriented Gradients (HOG) features are extracted. The classification task at each iteration is performed considering Support Vector Machine (SVM). A 4-fold training-test setting is considered to assess the performance of the algorithm. The procedure may depend upon the image chosen for initialisation, hence averages computed over all possible initialisations are reported, for each criterion. Thresholds for certain and uncertain data are empirically set to $t_c = 0.8$ and $t_{unc} = 0.1$.

Figure 2: Images from shellfish dataset with 15, 3 and 26 shellfish gatherers

Table 1: Dataset statistics (average over the 4 train/test splits)

	Training set	Test set
# images	23	5
# total instances	682,410	148,350
# positive instances	537 (N^+)	114

Main results The proposed method is first classically evaluated regarding the classifier performances *w.r.t.* the retraining strategies (see Table. 2). Adding diversity to the uncertain training (UC) set considering also certain examples (UC+C) allows better detection performances (higher F-score) but taking into account the extra-knowledge (UC+C+EK) decreases the average F-score on the test sets: if the precision increases, the recall sharply decreases. The number of interactions with the user is the number of corrections that have to be made by the user to the classifier predictions (FP, FN) during successive iterations $\ell = 1..K$. The interaction gain G_ℓ is the difference with the number of true positives, that would have been the number of interactions in an annotation task without any selective sample process. A crucial characteristic for an active process is therefore the total number of interaction gain at the end of the process (G_K). According to this criterion, UC+C+EK performs best as it is more robust to the initialisation step. This conversely differs from classification performances, but this user metric evaluation fits the objective of easing the annotation.

Table 2: Evaluation of re-training strategies: classifier performances and user interaction gain - Average (std) over the 4 test sets.

	Classifier perf.			Interaction gain (G_K)	
	F-score	Recall	Precision	mean (std)	min/max
UC	0.22 (0.17)	0.58 (0.30)	0.25 (0.25)	20.5 (26.5)	0/87
UC+C	**0.42** (0.10)	**0.76** (0.07)	0.36 (0.13)	41.8 (38.9)	0/**135**
UC+C+EK	0.08 (0.04)	0.05 (0.03)	**0.50** (0.20)	**77.5** (16.6)	**43**/102

4 Conclusion and discussion

We introduced an active learning annotation process to reduce the annotation cost when creating a ground truth. Usual active learning algorithms perform instances selection from the whole set of input data. In the present work, the selection of the queried instances is constrained by requiring that they belong to a group, i.e. (a part of) an image here, to ease the annotator task as the queried instances are proposed in their comprehensive context. We defined a score to rank the images and identify the one that should be annotated at each iteration, based on both uncertainty and true positives. The main objective is to reduce the number of human interactions on the overall process, starting from a first annotated image, rather than reaching the maximum final accuracy. Therefore, the annotation cost is measured through the gain in interactions (corrections of the classifier decisions by the annotator) with respect to a labelling task from scratch. At each iteration, the classifier is retrained according to a specific subset of data. Several strategies have been compared and their performances regarding the interaction gain have been discussed. We also highlight that initialisation is a crucial step to our design. While out of the scope of this study, it requires further investigation to gain robustness in the process. Improvements can also be brought considering more appropriate features, such as those based on convolutional auto-encoder that can suit better than HOG for small object detection problem in aerial images.

References

[1] Y. Fu, X. Zhu, and B. Li. A survey on instance selection for active learning. *Knowledge and Information Systems*, 35(2):249–283, 2013.

[2] H He and E. A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009.

[3] M. Laroze, L. Courtrai, and S. Lefèvre. Human detection from aerial imagery for automatic counting of shellfish gatherers. In *Inter. Conf. on Computer Vision Theory and Applications (VISAPP)*, 2016.

[4] M. Laroze, R. Dambreville, C. Friguet, E. Kijak, and S. Lefèvre. Active Learning to Assist Annotation of Aerial Images in Environmental Surveys. In *International Conference on Content-Based Multimedia Indexing*, pages 1–6, 2018.

[5] B. Settles. Active learning literature survey. Computer sciences technical report, University of Wisconsin–Madison, 2010.

[6] D. Tuia, M. Volpi, L. Copa, M. Kanevski, and J. Munoz-Mari. A survey of active learning algorithms for supervised remote sensing image classification. *IEEE Journal of Selected Topics in Signal Processing*, 5(3):606–617, 2011.

Inférence efficace des modèles à blocs stochastiques et à blocs latents pour les graphes creux

Gabriel Frisch, Jean-Benoist Leger & Yves Grandvalet

Université de Technologie de Compiègne, CNRS, Heudiasyc UMR 7253, Compiègne France. {gabriel.frisch,jbleger,yves.grandvalet}@hds.utc.fr

Résumé. Nous présentons un algorithme d'inférence variationnelle pour les modèles à blocs stochastiques et à blocs latents pour les graphes creux, qui tire parti de la rareté des arêtes pour passer à l'échelle sur un très grand nombre de nœuds. Cet algorithme est implémenté sous la forme d'un module python, appelé *sparsebm*, qui peut traiter des graphes comportant des millions de nœuds.
Mots-clés. Co-clustering, LBM, SBM, inférence variationnelle, graphe creux, matrice creuse

Abstract. We present a variational inference algorithm for the Stochastic Block Model and the Latent Block Model for sparse graphs, which leverages on the sparsity of edges to scale up to a very large number of nodes. This algorithm is implemented as a python package, called *sparsebm*, which can handle graphs with millions of nodes.
Keywords. Co-clustering, LBM, SBM, variationnal inference, sparse graph, sparse matrix

1 Introduction

Dans un graphe dense, l'adjacence des noeuds est traditionnellement représentée sous la forme d'une matrice d'adjacence. Ce choix est motivé par la simplicité de sa représentation mais aussi par la performance des calculs vectoriels fournis par les outils de programmation. Lorsque le degré moyen des nœuds est faible, les matrices d'adjacences comportent majoritairement des zéros et peuvent être qualifiées de « creuses ». Ces types de graphes apparaissent souvent dans les jeux de données issues de réseaux sociaux ou de systèmes collaboratifs. Par exemple, le jeu de données de LinkedIn [1] compte 3 millions d'utilisateurs avec un degré moyen de 8.3 par utilisateur. Le jeu de données MovieLens-25M [2], référence pour le filtrage collaboratif, peut être modélisé par un réseau bipartite composé de 120 000 nœuds utilisateurs et de 60 000 nœuds films, avec un degré moyen de 112. Dans ces contextes, la taille des matrices d'adjacence devient problématique pour les implémentations existantes des modèles à blocs stochastiques (SBM) ou des modèles à blocs latents (LBM). Nous montrons ici comment, en utilisant une représentation sous

1. Disponible sur https://lfs.aminer.cn/lab-datasets/multi-sns/aminer.tar.gz
2. Disponible sur https://grouplens.org/datasets/movielens/

forme de liste de l'adjacence des nœuds, il est possible de réaliser une inférence efficace du SBM et du LBM dans de très larges graphes creux.

Nous présentons dans un premier temps les modèles à blocs latents (LBM) et les modèles à blocs stochastiques (SBM) tels qu'initialement proposés par leurs auteurs. Nous comparons ensuite les inférences variationnelles originelles et celles utilisant des astuces calculatoires afin de réduire la complexité pour les graphes creux. Nous montrons enfin que, grâce à ces astuces, ces modèles peuvent être utilisés pour analyser des graphes comportant des centaines de milliers de nœuds pour peu qu'ils aient relativement peu d'arêtes. Les implémentations sont disponibles publiquement dans un module python [3] qui utilise l'accélération matérielle des GPUs.

2 Modèles

2.1 Modèle à blocs latents (LBM)

Le modèle à blocs latents (LBM) est un modèle probabiliste de co-clustering [3] qui permet de classifier simultanément les nœeuds d'un graphe bipartite, dont la matrice d'adjacence est notée $\boldsymbol{X}$. Le LBM repose sur l'hypothèse que cette matrice d'adjacence est structurée en blocs homogènes issus d'une double partition des lignes, en k_1 parties, et des colonnes, en k_2 parties, de cette matrice.

Comme introduit pour la première fois par Govaert et Nadif [2], nous considérons ici la version la plus simple du LBM, dans laquelle la matrice d'adjacence $\boldsymbol{X}$, de taille $n_1 \times n_2$, est binaire. Ses lignes correspondent aux n_1 nœuds de type (1) et ses colonnes aux n_2 nœuds de type (2) du graphe bipartite. La variable aléatoire X_{ij} associée à chaque paire de nœuds (i, j), respectivement de type (1) et de type (2), code la présence ou l'absence d'arête entre i et j : $X_{ij} = 1$ si l'arête est présente, et $X_{ij} = 0$ sinon.

On introduit $\boldsymbol{U}$ la matrice $n_1 \times k_1$ indicatrice de l'appartenance aux classes lignes et $\boldsymbol{V}$ la matrice $n_2 \times k_2$ indicatrice de l'appartenance aux classes colonnes. Nous avons $U_{iq} = 1$ si la ligne i appartient à la classe ligne q et $U_{iq} = 0$ autrement. De façon similaire, V_{jl} prend ses valeurs dans $\{0, 1\}^{k_2}$ et indique l'appartenance de la colonne j à la classe l.

Le LBM fait plusieurs hypothèses sur la forme et la dépendance des variables :

— Les appartenances aux classes des nœuds de type (1) et des nœuds de type (2) sont *a priori* indépendantes. Les variables latentes $\boldsymbol{U}$ et $\boldsymbol{V}$ sont *a priori* indépendantes : $p(\boldsymbol{U}, \boldsymbol{V}) = p(\boldsymbol{U})p(\boldsymbol{V})$.

— Les catégories des nœuds de type (1) sont indépendantes et identiquement distribuées selon une loi multinomiale $\mathcal{M}(1; \boldsymbol{\alpha})$ où $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{k_1})$ représente les proportions du mélange des classes en ligne.

3. Module python : sparsebm, disponible sur `https://pypi.org/project/sparsebm/`

— Les catégories des nœuds de type (2) sont indépendantes et identiquement distribuées selon une loi multinomiale $\mathcal{M}(1;\boldsymbol{\beta})$ où $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{k_2})$ représente les proportions du mélange des classes en colonne.
— La présence de l'arête entre le nœud i de type (1) et le nœud j de type (2), représentée par la variable aléatoire X_{ij} est considérée comme dépendant uniquement des classes des nœuds i et j. Autrement dit, le LBM suppose que tous les éléments d'un bloc suivent la même distribution de probabilité. La densité conditionnelle pour une observation X_{ij} du bloc ql s'écrit alors : $\phi(X_{ij};\pi_{ql}) = \pi_{ql}^{X_{ij}}(1-\pi_{ql})^{1-X_{ij}}$

La variable aléatoire $\boldsymbol{X}$ suit une densité mélange :

$$\begin{aligned} p(\boldsymbol{X};\boldsymbol{\theta}) &= \sum_{\boldsymbol{U},\boldsymbol{V}\in I\times J} p(\boldsymbol{U};\boldsymbol{\alpha})\; p(\boldsymbol{V};\boldsymbol{\beta})\; p(\boldsymbol{X}|\boldsymbol{U},\boldsymbol{V};\boldsymbol{\pi}) \\ &= \sum_{\boldsymbol{U},\boldsymbol{V}\in I\times J} \prod_{i,q} {\alpha_q}^{U_{iq}} \prod_{j,l} {\beta_l}^{V_{jl}} \prod_{i,j,q,l} \phi(X_{ij};\pi_{ql})^{U_{iq}V_{jl}} \quad , \end{aligned} \tag{1}$$

où I (resp. J) représente l'ensemble des partitions possibles pour les lignes (resp. colonnes) et $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\pi})$ représente le vecteur de paramètres du modèle.

2.2 Modèle à blocs stochastiques (SBM)

Le modèle SBM [4] peut être vu comme un LBM contraint à une seule partition des nœuds du graphe. Il ne s'applique donc plus aux graphes bipartite, mais il repose toujours sur l'hypothèse d'une structure de la matrice d'adjacence $\boldsymbol{X}$ en blocs homogènes. Le SBM fait plusieurs hypothèses sur la forme et la dépendance des variables :
— Les appartenances aux classes des nœuds sont indépendantes et identiquement distribuées selon une loi multinomiale $\mathcal{M}(1;\boldsymbol{\alpha})$ où $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{k_1})$ représente les proportions du mélange des classes.
— La présence de l'arête entre le nœud i et le nœud i', noté $X_{ii'}$ est considérée comme dépendant uniquement des classes des nœuds i et i'. Autrement dit, le SBM suppose que tous les éléments d'un bloc suivent la même distribution de probabilité. La densité conditionnelle pour une observation $X_{ii'}$ du bloc qq' s'écrit alors : $\phi(X_{ii'};\pi_{qq'}) = \pi_{qq'}^{X_{ii'}}(1-\pi_{qq'})^{1-X_{ii'}}$.

La variable aléatoire $\boldsymbol{X}$ suit une densité mélange :

$$p(\boldsymbol{X};\boldsymbol{\theta}) = \sum_{\boldsymbol{U}\in I} \prod_{i,q} {\alpha_q}^{U_{iq}} \prod_{i',q'} {\alpha_{q'}}^{U_{i'q'}} \prod_{i,q,i',q'} \phi(X_{ii'};\pi_{qq'})^{U_{iq}U_{i'q'}} \quad ,$$

où I représente l'ensemble des partitions possibles et $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\pi})$ représente le vecteur de paramètres du modèle.

3 Inférence

Afin d'inférer les paramètres du modèle pour réaliser le co-clustering, nous maximisons la vraisemblance de chaque modèle. L'estimateur du maximum de vraisemblance doit mener à $\widehat{\theta} = \arg\max_{\theta} p(\boldsymbol{X}; \theta)$. Cependant, la vraisemblance n'est pas calculable car elle implique un nombre de termes exponentiel dans la taille du graphe. Pour pallier cela, une reformulation variationnelle du critère est utilisée dans l'algorithme EM [1].

Inférence du modèle à blocs latents Nous présentons l'inférence variationnelle proposée par Govaert et Nadif [3] que nous adaptons afin d'utiliser la faible connectivité des nœuds pour accélérer les calculs. Le critère variationnel, borne inférieure de la vraisemblance est :

$$\mathcal{J}(q_\gamma, \boldsymbol{\theta}) = \mathcal{H}(q_\gamma) + \mathbb{E}_{q_\gamma}[\log p(\boldsymbol{X}, \boldsymbol{U}, \boldsymbol{V}; \boldsymbol{\theta})] \ , \tag{2}$$

avec q_γ représentant la distribution variationnelle, $\mathcal{H}$ l'entropie différentielle et $\mathbb{E}_{q_\gamma}$ l'espérance sous la distribution variationnelle. La distribution variationnelle est restreinte à un ensemble de distributions factorisables (approximation champ moyen ou *mean field* [5]) de la forme : $q_\gamma = \prod_i \mathcal{M}(1, \tau_i^{(U)}) \ \prod_j \mathcal{M}(1, \tau_j^{(V)})$, où $\gamma = (\tau^{(U)}, \tau^{(V)})$ est le vecteur des paramètres variationnels. En utilisant l'indépendance des variables latentes, le critère se réécrit sous la forme :

$$\mathcal{J}(q_\gamma, \boldsymbol{\theta}) = \mathcal{H}(q_\gamma) + \mathbb{E}_{q_\gamma}[\log p(\boldsymbol{U}; \boldsymbol{\alpha})] + \mathbb{E}_{q_\gamma}[\log p(\boldsymbol{V}; \boldsymbol{\beta})] + \mathbb{E}_{q_\gamma}[\log p(\boldsymbol{X}|\boldsymbol{U}, \boldsymbol{V}; \boldsymbol{\theta})] \ , \tag{3}$$

avec

$$\mathcal{H}(q_\gamma) = -\sum_{iq} \tau_{iq}^{(U)} \log \tau_{iq}^{(U)} - \sum_{jl} \tau_{jl}^{(U)} \log \tau_{jl}^{(V)} \tag{4}$$

$$\mathbb{E}_{q_\gamma}[\log p(\boldsymbol{U}; \boldsymbol{\alpha})] = \sum_{iq} \tau_{iq}^{(U)} \log \alpha_q \tag{5}$$

$$\mathbb{E}_{q_\gamma}[\log p(\boldsymbol{V}; \boldsymbol{\beta})] = \sum_{jl} \tau_{jl}^{(V)} \log \beta_l \tag{6}$$

$$\mathbb{E}_{q_\gamma}[\log p(\boldsymbol{X}|\boldsymbol{U}, \boldsymbol{V}; \boldsymbol{\theta})] = \sum_{ijql} \tau_{iq}^{(U)} \tau_{jl}^{(V)} (X_{ij} \log \pi_{ql} + (1 - X_{ij}) \log(1 - \pi_{ql})) \ . \tag{7}$$

Le calcul coûteux est celui de l'espérance de la log-vraisemblance conditionnelle de $\boldsymbol{X}$ (7), dont la somme implique tous les éléments de la matrice d'adjacence, entraînant une complexité calculatoire en $\mathcal{O}(n_1 n_2 k_1 k_2)$ où n_1, n_2 sont les dimensions de $\boldsymbol{X}$ et k_1, k_2 sont respectivement le nombre de classes des nœuds de type (1) et le nombre de classes des nœuds de type (2).

Il est cependant possible de réécrire l'équation (7) afin d'itérer uniquement sur les éléments non nuls de la matrice entraînant une complexité calculatoire en $\mathcal{O}(\#\{ij : X_{ij} \neq$

$1\}k_1k_2$) où $\#\{ij : X_{ij} = 1\}$ désigne le nombre d'entrées non nulles dans $\boldsymbol{X}$:

$$\mathbb{E}_{q_\gamma}[\log p(\boldsymbol{X}|\boldsymbol{U}, \boldsymbol{V}; \boldsymbol{\theta})] = \sum_{qlij:X_{ij}=1} \tau_{iq}^{(U)} \tau_{jl}^{(V)} (\log \pi_{ql} - \log(1 - \pi_{ql})) + \sum_{ql} \log(1 - \pi_{ql}) \left(\sum_i \tau_{iq}^{(U)} \right) \left(\sum_i \tau_{jl}^{(V)} \right) . \quad (8)$$

En utilisant l'équation (8), le critère variationnel $\mathcal{J}(q_\gamma, \boldsymbol{\theta})$ est calculé de manière efficace pour les graphes creux.

L'algorithme EM se réécrit en une double maximisation alternée du critère $\mathcal{J}(q_\gamma, \boldsymbol{\theta})$ par rapport à q_γ et par rapport à $\boldsymbol{\theta}$. Nous contrastons ci-dessous les mises à jour des paramètres telle que définie dans le LBM original d'une part, et telle que nous la réécrivons pour les graphes creux d'autre part. La complexité mémoire qui était en $\mathcal{O}(n_1 n_2)$ est réduite à $\mathcal{O}(\#\{ij : X_{ij} = 1\})$.

EM - version originale	EM - version creuse
Étape-E	*Étape-E*
Répéter	Répéter
$\tau_{iq}^{(U)} \propto \alpha_q \prod_l \pi_{ql}^{Q_{il}} (1 - \pi_{ql})^{R_l - Q_{il}}$	$\tau_{iq}^{(U)} \propto \alpha_q \prod_{jl} (1 - \pi_{ql})^{\tau_{jl}^{(V)}} \prod_l \frac{\pi_{ql}^{Q_{il}}}{(1-\pi_{ql})^{Q_{il}}}$
avec $Q_{il} = \sum_j \tau_{jl}^{(V)} X_{ij}$ et $R_l = \sum_j \tau_{jl}^{(V)}$	avec $Q_{il} = \sum_{j:X_{ij}=1} \tau_{jl}^{(V)}$
$\tau_{jl}^{(V)} \propto \beta_l \prod_q \pi_{ql}^{S_{jq}} (1 - \pi_{ql})^{T_q - S_{jq}}$	$\tau_{jl}^{(V)} = \beta_l \prod_{iq} (1 - \pi_{ql})^{\tau_{iq}^{(U)}} \prod_q \frac{\pi_{ql}^{S_{jq}}}{(1-\pi_{ql})^{S_{jq}}}$
avec $S_{jq} = \sum_i \tau_{iq}^{(U)} X_{ij}$ et $T_q = \sum_i \tau_{iq}^{(U)}$	avec $S_{jq} = \sum_{i:X_{ij}=1} \tau_{iq}^{(U)}$
jusqu'à convergence	jusqu'à convergence
Étape-M	*Étape-M*
$\alpha_q = \frac{\sum_i \tau_{iq}^{(U)}}{n_1}$ $\quad$ $\beta_l = \frac{\sum_j \tau_{jl}^{(V)}}{n_2}$ $\quad$ $\pi_{ql} = \frac{\sum_{ij} \tau_{iq}^{(U)} \tau_{jl}^{(V)} X_{ij}}{\sum_{ij} \tau_{iq}^{(U)} \tau_{jl}^{(V)}}$	$\alpha_q = \frac{\sum_i \tau_{iq}^{(U)}}{n_1}$ $\quad$ $\beta_l = \frac{\sum_j \tau_{jl}^{(V)}}{n_2}$ $\quad$ $\pi_{ql} = \frac{\sum_{ij:X_{ij}=1} \tau_{iq}^{(U)} \tau_{jl}^{(V)}}{\sum_i \tau_{iq}^{(U)} \sum_j \tau_{jl}^{(V)}}$

Inférence du Modèle à blocs stochastiques L'inférence et les astuces calculatoires sont similaires à celles utilisées pour le LBM. Nous ne donnons donc pas ici le détail du critère ou celui de la mise à jour des paramètres du modèle.

4 Expérimentations

Nous présentons les temps de calcul pour réaliser l'inférence du SBM avec des données simulées et réelles. Notre package *sparsebm* offre la possibilité d'utiliser l'accélération

matérielle fournie par un GPU. Le GPU utilisé pour ces expérimentations est un Tesla V100-SXM2-32GB. Le temps d'exécution de l'algorithme correspond à l'exécution de 100 itérations d'EM sur 100 initialisations aléatoires, suivie d'itérations jusqu'à convergence pour les 10 meilleurs résultats à l'issue de ces 100 itérations. La convergence est établie classiquement par la stagnation du critère optimisé.

TABLE 1 – Temps d'exécution du SBM : haut à gauche, pour un nombre de nœuds variable avec un degré moyen de 10 ; haut à droite, pour un nombre de nœuds fixe (10^4) et à degré moyen variable ; bas, pour des jeux de données réels. Le nombre de classes est toujours fixé à 4.

Nombre de nœuds	Temps d'exécution
$1 \cdot 10^3$	14.1 s
$5 \cdot 10^3$	16.6 s
$1 \cdot 10^4$	23.7 s
$5 \cdot 10^4$	1 min 51.3 s
$1 \cdot 10^5$	3 min 50.0 s
$5 \cdot 10^5$	20 min 30.4 s
$1 \cdot 10^6$	44 min 41.1 s
$5 \cdot 10^6$	266 min 42.0 s
avec un degré moyen de 10	

Degré moyen	Temps d'exécution
10	23.7 s
15	31.8 s
20	46.1 s
50	1 min 55.4 s
100	4 min 53.2 s
150	9 min 2.0 s
200	12 min 48.4 s
avec 10^4 nœuds	

Dataset	Nombre de nœuds	Degré moyen	Temps d'exécution
Flickr-medium [a]	$2 \cdot 10^5$	42.4	52 min 47.7 s
LinkedIn	$3 \cdot 10^6$	8.3	340 min 45.0 s

a. Disponible sur `https://www.aminer.cn/data-sna#Flickr-medium`

Références

[1] A. P. DEMPSTER, N. M. LAIRD et D. B. RUBIN : Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society : Series B (Methodological)*, 39(1):1–22, 1977.

[2] G. GOVAERT et M. NADIF : Block clustering with Bernoulli mixture models : Comparison of different approaches. *Computational Statistics & Data Analysis*, 52(6):3233–3245, February 2008.

[3] G. GOVAERT et M. NADIF : *Co-clustering : models, algorithms and applications*. ISTE Ltd, 2013.

[4] P. W. HOLLAND, K. B. LASKEY et S. LEINHARDT : Stochastic blockmodels : First steps. *Social Networks*, 5(2):109 – 137, 1983.

[5] T. S. JAAKKOLA : Tutorial on variational approximation methods. *In Advanced mean field methods : theory and practice*, p. 129–159. MIT Press, 2000.

Deep Gaussian Mixture Models for mixed type data

Robin Fuchs [1] & Cinzia Viroli [2] & Denys Pommeret [3]

[1] *Institut de Mathématiques de Marseille Campus de Luminy Case 901, 70, RTE Léon Lachamp, 13009 Marseille - robin.fuchs@univ-amu.fr*
[2] *Department of Statistics, University of Bologna, Via Belle Arti 41, Bologna, Italy - cinzia.viroli@unibo.it*
[3] *Institut de Mathématiques de Marseille Campus de Luminy Case 901, 70, RTE Léon Lachamp, 13009 Marseille - denys.pommeret@univ-amu.fr*

Résumé. Les *Deep Gaussian Mixture Models* (DGMM) sont des modèles multicouches popularisés par McLachlan et Viroli (2019). Ils opèrent dans le cadre de l'apprentissage non-supervisé et effectuent simultanément une réduction de la dimensionalité ainsi qu'un *clustering* des données. Toutefois, les DGMMs ne peuvent à l'heure actuelle uniquement traiter que des données continues. Le présent travail propose de définir une couche d'*embedding* apprenant une représentation continue des données mixtes (discrètes et continues) présentes dans le jeu de données et d'utiliser cette représentation pour entraîner le DGMM. Deux méthodes d'intégration différentes sont ici envisagées : utiliser une couche reposant sur les Modèles Linéaires Généralisés à Variables Latentes (GLLVM) ou intégrer le DGMM au sein d'un Auto-encodeur Variationnel. Cette dernière considération a motivé l'écriture d'une version variationnelle du DGMM qui, en elle-même, présente également des avantages importants tels que la stabilisation de l'algorithme ou la possibilité de faire de la sélection de modèle au cours de l'entraînement.

Mots-clés. Apprentissage non-supervisé, Apprentissage Profond, Données mixtes

Abstract. Deep Gaussian Mixture Models (DGMM) are multilayer models popularized by McLachlan and Viroli (2019). They are designed to conduct unsupervised tasks and perform both dimension reduction and clustering at the same time. However, DGMMs could until now only deal with continuous datasets. The current work proposes to define an embbeding layer that learns a continuous representation of both discrete and continuous data in the dataset and uses it to train the DGMM. Two different embedding methods are considered, either to use a Generalized Linear Latent Variable Model (GLLVM) based layer or to integrate the DGMM into a Variational AutoEncoder framework. The latter consideration has required to write a Variational version of the DGMM which in itself present significant advantages such as to stabilize the algorithm or to conduct model selection during training.

Keywords. Unsupervised learning, deep learning, mixed type data

1 Texte long

1.1 Handling mixed type data with a GLLVM layer

Deep Gaussian Mixtures Models (DGMM) [3] are multilayer models in which each layer can be written as a Mixture of Factor Analysers (MFA) [12]. As such they can be seen as a generalisation of both Factor Models and Gaussian Mixtures Models. We denote by y a matrix of observed data of dimension $n \times p$, by i the observations index and we fix L an arbitrary number of layers in the DGMM of dimension $K_1, ..., K_L$, respectively. We consider the following model:

$$\begin{cases} y_i = \eta_{k_1}^{(1)} + \Lambda_{k_1}^{(1)} z_i^{(1)} + u_{i,k_1}^{(1)} \text{ with probability } \pi_{i,k_1}^{(1)}, k_1 = 1, ..., K_1 \\ z_i^{(1)} = \eta_{k_2}^{(2)} + \Lambda_{k_2}^{(2)} z_i^{(2)} + u_{i,k_2}^{(2)} \text{ with probability } \pi_{i,k_2}^{(2)}, k_2 = 1, ..., K_2 \\ ... \\ z_i^{(L-1)} = \eta_{k_L}^{(L)} + \Lambda_{k_L}^{(L)} z_i^{(L)} + u_{i,k_L}^{(L)} \text{ with probability } \pi_{i,k_L}^{(L)}, , k_L = 1, ..., K_L \\ z_i^{(L)} \sim \mathcal{N}(0, I_{r_L}), \end{cases}$$

where for $l \in [1, L]$, $z_i^{(l)}$ are the latent variables, $\Lambda_{k_l}^{(l)}$ are factor loading matrices of dimension $r_{l-1} \times r_l$, i.e. the base change matrices between $z_i^{(l-1)}$ and $z_i^{(l)}$, $\eta_{k_l}^{(l)}$ are vectors of constants of dimension $r_{l-1} \times 1$ and $u_{k_l}^{(l)}$ are centered Gaussian error terms of covariance matrices $\Psi_{k_l}^{(l)}$.
At the lth layer, an observation i can be viewed as a K_l components mixture with associated probability $\pi_{i,k_l}^{(l)}$. Then the distribution of each observation is a mixture with $K = K_1 \times ... \times K_L$ components or paths, which makes it very flexible. By assigning an observation to the component which has the highest probability $\pi_{i,k_l}^{(l)}$, DGMMs appear as natural clustering models.
Figure 1 presents a graphical representation of a DGMM with $L = 2$, $K_1 = 5$ and $K_2 = 4$.

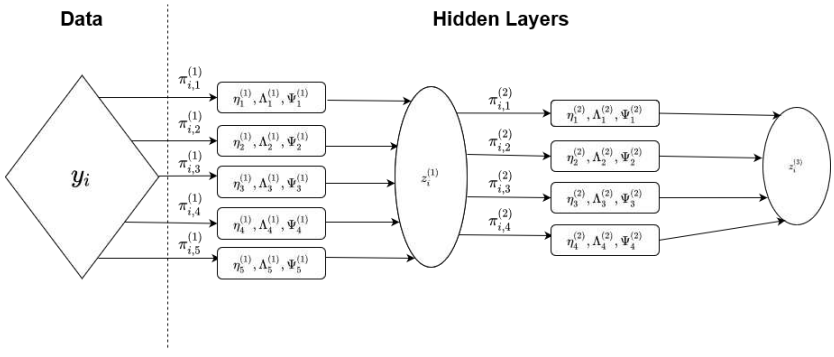


Figure 1: Schematic representation of a Deep Gaussian Mixture Model.

Viroli and McLachlan (2019) [3] have designed the DGMM to handle continuous random variables y. We propose to extend their approach to mixed data (both continuous and discrete) first by finding a continuous representation and then passing it to the DGMM. This issue is very similar to Deep Natural Language processing tasks when one has to find a low dimensional continuous representation for words or sentences in order for them to be fed into a Neural Network. That is why some pre-trained word embeddings have been created for instance by Mikolov, Sutskever and al. (2013) [4].

In order to illustrate what a mixed type dataset is, one can think of the Pima Indians Diabetes dataset which is made of binary variables (e.g. valued 1 if an individual has been tested positive for diabetes and 0 otherwise), count variables (e.g. the number of times an individual has been pregnant) and various continuous variables (e.g. blood pressure or glucose concentration).

In such a case, we propose to introduce a GLLVM layer between the data and the first layer of the DGMM (represented by the dashed line on Figure 1).
GLLVM were developed from the seminal works by Moustaki and Knott (2000) [1] or Skrondal and Rabe-Hesketh (2004) [2]. These models assume the existence of a latent continuous representation of the mixed observed data and try to estimate it.
Hence,

$$f(y) = \int_{R^q} f(y|z^{(1)}) f(z^{(1)}) dz^{(1)} \tag{1}$$

with $z^{(1)}$ the latent continuous variables. GLLVM assume that the variables $y_j \ \forall j \in [1, p]$ composing y, are mutually independent and that the distributions of these variables given the latent ones belong to an exponential family.
Hence, it comes that $f(y|z^{(1)}) = \prod_{j=1}^{p} f(y_j|z^{(1)})$ with $f(y_j|z^{(1)})$ belonging to an exponential family.
To give some examples, for discrete data, if y_j is a binary variable, $f(y_j|z^{(1)})$ can be modeled by a Bernoulli distribution, if it is a categorical variable one can use a multinomial distribution. Note that continuous variables can also "go through" this GLLVM layer by modeling them by continuous distributions of an exponential family.

Concerning $f(z^{(1)})$ in (1), we extend the approach of Viroli and Cagnone (2014) [5] and specify this distribution as being a Mixture of Factor Analysers. As $z^{(1)}$ is a MFA, it then becomes natural to include it as the first layer of the DGMM.

1.2 Handling mixed type data with a Variational DGMM

The quality of the continuous representation of mixed data is of prime importance as the extracted pieces of information are then propagated through the DGMM. Using the GLLVM approach provides a simple and statistically well-funded continuous representation space. However, limiting the links between mixed and continuous spaces to an exponential family can be regarded as a strong hypothesis. As a result, initial information is carried by a relatively low number of parameters compared to the number of parameters of the following DGMM layers, which might act as an information bottleneck. Besides some optimisation steps are needed to perform GLLVM estimation, which could question the scalability of this embedding for large and high dimensional datasets.

Alternatively, Variational AutoEncoders (VaE) may be a more flexible way to learn a continuous representation of mixed type data. VaE have been introduced by Kingma and Welling (2014) [9] and by Rezende and al. (2014) [10].
They are made of two parts: an encoder that aims to compress the signal into a very simple continuous space and a decoder that tries to reconstruct the original signal from the compressed one. Figure 2 gives a schematic representation of the architecture of an AutoEncoder.

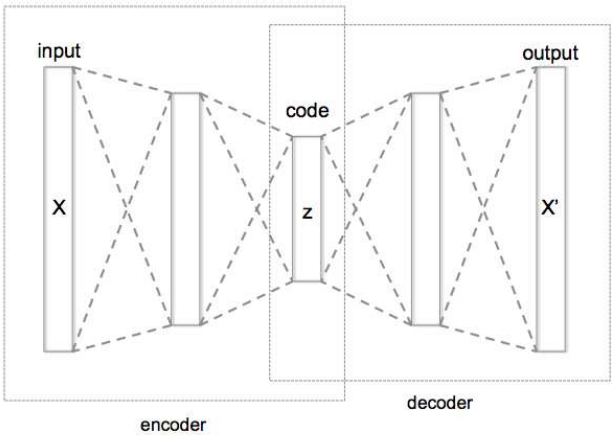


Figure 2: Autoencoder representation[1].

The continuous space hosting the compressed signal is usually chosen to be Gaussian or even to be a Mixture of Gaussians as in Dilokthanakul & al. (2016) [6].

[1]By Chervinskii, under Creative Commons License

We propose here to use a DGMM as a representation space of this compressed signal. After training of the whole model (encoder, DGMM and decoder) the encoder part will provide the link between mixed observed data and their continuous representation.

We have first to rewrite the DGMM using Variational Methods for which Blei and al. (2017) [11] have published a recent state-of-the-art. Hence by assuming that some blocks of parameters are mutually independent at the layer level, it is possible to derive closed form and simple posterior distributions of the parameters.
Nevertheless, one has to change the training method of the DGMM to fully integrate it with the VaE. Currently the training of the DGMM is conducted with a Stochastic Version of the EM algorithm proposed by Celeux and Diebolt (1985) [7], whereas VaE training is performed using backpropagation. As one can exhibit the likelihood (and its lower bound) of the DGMM, it seems possible to train a DGMM with backpropagation which could be undertaken in future works.

This Variational version of the DGMM also presents strong advantages by itself compared to the regular one. First, defining a Bayesian prior on parameters makes it possible to get rid off the well-known singularity occuring in mixture models when one of the groups collapses to one observation. As pointed out in Bishop (2006) [8], in frequentist mixtures when only one observation is associated to a group then the intragroup variance tends to zero and the likelihood tends to infinity: the model becomes degenerated. In Bayesian mixtures, defining a prior acts as a regularizer and remove that singularity making the algorithm more stable.

Furthermore, at each layer all observations have probability $\pi_{i,k_l}^{(l)}$ to be associated to one of the k_l components. It comes that for very wide and deep networks, some of the components are useless as they provide no satisfactory representation of the data. In the Variational Approach $\pi_{i,k_l}^{(l)}$ has a closed form posterior probability. It is then possible to threshold to zero very low probabilities and to remove the associated components during the training. This model selection method is far less compute-intensive than estimating the whole model with different specifications and comparing information criteria.

Finally using Variational Methods, one can force $\Lambda_{k_l}^{(l)T}\Psi_{k_l}^{(l)-1}\Lambda_{k_l}^{(l)}$ to be diagonal which is a necessary condition for model identifiability.

Bibliographie

[1] Moustaki, I. and Knott, M. (2000), Generalized latent trait models, *Psychometrika*, 65(3), Springer, pp. 391-411

[2] Skrondal, A. and Rabe-Hesketh, S. (2004), Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models, *Chapman & Hall*, London

[3] Viroli, C. and McLachlan, G. J. (2019), Deep gaussian mixture models, *Statistics and Computing*, 29(1), pp.43-51

[4] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, pp. 3111-3119.

[5] Cagnone, S., and Viroli, C. (2014). A factor mixture model for analyzing heterogeneity and cognitive structure of dementia. *AStA Advances in Statistical Analysis*, 98(1), pp. 1-20.

[6] Dilokthanakul, N., Mediano, P. A., Garnelo, M., Lee, M. C., Salimbeni, H., Arulkumaran, K., and Shanahan, M. (2016). Deep unsupervised clustering with gaussian mixture variational autoencoders. *arXiv preprint arXiv:1611.02648*.

[7] Celeux, G. and J. Diebolt (1985). The SEM algorithm: a probabilistic teacher algorithm derived from the em algorithm for the mixture problem. *Computational statistics* Q. 2(1), pp. 73–82

[8] Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.

[9] Kingma, D. P. and Welling, M. (2014). Auto-Encoding Variational Bayes. textitInternational Conference on Machine Learning, 1050(1).

[10] Rezende, D. J., Mohamed, S. and Wierstra, D. (2014, January). Stochastic Backpropagation and Approximate Inference in Deep Generative Models. *International Conference on Machine Learning*, pp. 1278-1286.

[11] Blei D. M., Kucukelbir A. and McAuliffe J.D. (2017) Variational Inference: A Review for Statisticians, Journal of the American Statistical Association (112:518), pp. 859-877

[12] Ghahramani Z., Hinton G.E., et al. (2006) The EM algorithm for mixtures of factor analyzers. Technical Report CRG-TR-96-1, University of Toronto.

Comparaison de lois a priori dans des modèles de médiation à réponse binaire

Jean-Michel Galharret [1] & Anne Philippe [2]

[1] *Laboratoire de mathématiques Jean-Leray (2 Rue de la Houssinière, 44322 Nantes), jean-michel.galharret@univ-nantes.fr*

[2] *Laboratoire de mathématiques Jean-Leray (2 Rue de la Houssinière, 44322 Nantes), anne.philippe@univ-nantes.fr*

Résumé. En sciences humaines, l'analyse de médiation consiste à étudier si l'effet d'une variable d'exposition X sur une variable réponse Y peut être décomposé en un effet direct et un effet indirect via une troisième variable M. Dans le cadre plus général de la causalité, l'effet direct naturel et l'effet indirect naturel font partie des paramètres d'intérêt des modèles de médiation. Nous utilisons un cadre bayésien pour estimer ces paramètres. Nous proposons d'inclure une information historique dans la loi a priori pour améliorer la qualité de l'estimation. Les résultats confirment ce fait. L'une des questions sous-jacente à la médiation est de tester l'existence de l'effet direct et de l'effet indirect. Étant donnée l'estimation de la loi a posteriori des paramètres, nous construisons des régions critiques pour un processus de test fréquentiste. En utilisant des simulations, nous comparons cette procédure avec les tests couramment utilisés en analyse de médiation. Enfin, nous appliquons notre approche à des données réelles provenant d'une étude longitudinale sur le bien-être des enfants à l'école.

Mots-clés . Analyse de médiation, estimation bayésienne, effet direct et indirect

Abstract. In the social sciences, mediation analysis consists of studying whether the effect of an exposure variable X on an outcome Y can be decomposed into a direct effect and an indirect effect via a third variable M. In the more general framework of causality, the natural direct effect and the natural indirect effect are among the parameters of interest in mediation models. We use a Bayesian framework to estimate these parameters. We propose to include historical information in the the prior distribution to improve the quality of the estimation. The results confirm this fact. One of the usual issues is to test the existence of the direct effect and the indirect effect. Given the estimation of the a posterior distribution of the parameters, we construct critical regions for a frequentist testing process. Using simulations, we compare this procedure with tests commonly used in mediation analysis. Finally, we apply our approach to real data from a longitudinal study of children's well-being in school.

Keywords. mediation analysis, Bayesian estimation, direct and indirect effect. ...

1 Introduction :

La modélisation proposée a été motivée par une étude visant à comprendre à travers quel processus de bonnes performances scolaires (X) induisent de bonnes performances scolaires l'année suivante (Y) et si une partie de cet effet ne transite pas à travers une augmentation du sentiment d'efficacité personnelle (M) des élèves. Ceci revient à estimer et tester les effets directs et indirects dans un modèle de médiation à réponse binaire. Soient X, Y binaires et $M \in \mathbb{R}$. On considère le modèle de médiation défini par :

$$\begin{cases} M = a_0 + a_1 X + \varepsilon \\ \mathbb{E}(Y|X,M) = \dfrac{1}{1+e^{-(b_0+b_1 M+b_2 X)}} \end{cases} \tag{1}$$

Ce modèle de médiation est donc un système constitué d'un modèle gaussien de régression et d'un modèle de régression logistique.
L'introduction de variables contrefactuelles dans le cadre de l'analyse de la causalité permet de donner un cadre formel à la définition des effets dans le modèle de médiation (voir Pearl, J. et Mackenzie, D. (2018) pour une synthèse récente). Ces effets peuvent être intuitivement définis par :

- L'effet direct est l'effet de X sur Y quand M est bloqué.
- L'effet indirect mesure l'effet de X sur Y uniquement à travers les variations de M.
- L'effet de X sur Y se décomposera comme la somme des deux effets précédents.

Sous certaines hypothèses sur les variables contrefactuelles, Imai et Keele (2010) obtiennent la représentation suivante des effets : pour $x \in \{0,1\}$,

$$\begin{aligned} \mathrm{NDE}(x) &= \int \big[\mathbb{E}(Y|X=1,M=m) - \mathbb{E}(Y|X=0,M=m)\big] \mathrm{d}\mathbb{P}_{M|X=x}(m) \\ \mathrm{NIE}(x) &= \int \mathbb{E}(Y|X=x,M=m) \mathrm{d}\mathbb{P}_{M|X=1}(m) - \int \mathbb{E}(Y|X=x,M=m) \mathrm{d}\mathbb{P}_{M|X=0}(m) \end{aligned} \tag{2}$$

Ces deux quantités seront désignées respectivement par effet direct et indirect dans la suite. L'effet causal ψ de $X \in \{0,1\}$ sur Y se décompose sous la forme

$$\psi = \mathrm{NDE}(0) + \mathrm{NIE}(1) = \mathrm{NDE}(1) + \mathrm{NIE}(0)$$

Dans ce travail, nous proposons une approche bayésienne pour estimer ces effets dans le modèle paramétrique défini en (1) et nous proposons différentes lois a priori sur les paramètres.

2 Le modèle bayésien

2.1 Définition du modèle

Soit $\underline{X} = (X_1, ..., X_N)$, $\underline{M} = (M_1, ..., M_N)$, $\underline{Y} = (Y_1, ..., Y_N)$ un échantillon de N réalisations indépendantes de (X, M, Y). Soient $\alpha := (a_0, a_1), \sigma^2, \beta := (b_0, b_1, b_2)$ les paramètres du modèle de médiation (1), la vraisemblance de ce modèle est

$$f(Y, M|\alpha, \beta, \sigma^2, X) = g(Y|\beta, M, X)\phi(M|\alpha, \sigma^2, X)$$

où

$$\begin{cases} g(Y|\beta, M, X) = \dfrac{\exp(Y(b_0 + b_1 M + b_2 X))}{1 + \exp(b_0 + b_1 M + B_2 X)} \\ \phi \text{ est la densité gaussienne } \mathcal{N}(a_0 + a_1 X, \sigma^2 I_n) \end{cases}$$

Pour compléter la définition du modèle bayésien, il reste à choisir la loi a priori sur le paramètre $\theta = (\alpha, \sigma^2, \beta)$.

Les effets directs et indirects définis en (2) sont des fonctions non explicites de θ, cependant leurs lois a posteriori peuvent être estimées en utilisant un algorithme MCMC. En effet, si on génère une chaine de Markov $(\theta^b)_{b\in\{1,...B\}}$ de loi stationnaire la loi a posteriori de θ, on peut déduire (par une approximation numérique d'intégrale) une chaine de Markov $(\text{NDE}_{\theta^b}(x))_{b\in\{1,...B\}}$ (resp. $(\text{NIE}_{\theta^b}(x))_{b\in\{1,...B\}}$) de loi stationnaire la loi a posteriori de $\text{NDE}_\theta(x)$ (resp. $\text{NIE}_\theta(x)$).

2.2 Choix des lois a priori

La table 1 résume les trois modèles comparés :

- Un modèle informatif proposé par Launay et al. (2015) qui repose sur des données historiques et qui suppose que les liens entre les variables ont peu évolué entre les données historiques et actuelles.
- Deux modèles non informatifs dont la loi a priori sur la partie gaussienne du modèle est un $G-$prior (Zellner 1971). Pour la partie logistique, on utilise une adaptation des $G-$prior (Marin et Robert 2014) et des lois de Cauchy proposées par Gelman et al. (2007).

<table>
<tr><th>model</th><th>σ^2</th><th>$\alpha = (a_0, a_1)$</th><th>$\beta = (b_0, b_1, b_2)$</th></tr>
<tr><td>informative</td><td rowspan="3">improper prior $\mathbb{1}_{\mathbb{R}^+}$</td><td colspan="2">Launay et al. (2015)</td></tr>
<tr><td>$G-$prior, Cauchy</td><td rowspan="2">G-prior</td><td>Gelman et al. (2007)</td></tr>
<tr><td>$G-$prior, $G-$prior</td><td>$G-$prior</td></tr>
</table>

Table 1: Résumé des choix pour les lois a priori du paramètre θ.

Modèle informatif : La loi a priori sur (α, β) définie par Launay et al. (2015) est basée sur des données historiques (X_h, M_h, Y_h) qui ont permis d'estimer la loi a posteriori du paramètre θ (cette estimation est réalisée avec le modèle $G-$prior, $G-$prior décrit dans la suite) et on calcule :

$$\begin{aligned} m &= \mathbb{E}\left(\alpha, \beta | (X_h, M_h, Y_h)\right) \\ V &= \mathbb{V}ar\left(\alpha, \beta | (X_h, M_h, Y_h)\right) \end{aligned} \tag{3}$$

En supposant que les données actuelles ont peu évolué par rapport aux données historiques, la loi a priori sur (α, β) est définie par

$$(\alpha, \beta) \sim \mathcal{N}_J(\mu, \lambda^{-1} V)$$

où

- $\mu = \begin{pmatrix} k_1 & & \\ & \ddots & \\ & & k_5 \end{pmatrix} m$, les coefficients $k_1, ..., k_5$ sont i.i.d. tels que $k_5 \sim \mathcal{N}(1, \tau^2)$. τ^2 est choisi assez grand pour que la loi soit non informative.
- $\lambda \sim \mathcal{G}(a, b)$. Les paramètres (a, b) étant choisis pour que la loi soit également non informative.

Les paramètres $k_1, ..., k_5$ et λ modélisent des éventuelles évolutions entre les deux échantillons.

Modèle $G-$prior, $G-$prior : Les $G-$priors ont été introduits par Zellner (1971) dans le cadre des modèles gaussiens, puis adaptés aux modèles linéaires généralisés (voir Marin et Robert 2014). Un $G-$prior sur l'ensemble des paramètres (α, β) n'est pas envisageable puisque le médiateur M est une variable réponse dans l'équation de régression linéaire et explicative dans l'équation logistique. Cependant, compte tenu de la décomposition de la loi jointe du modèle

$$p(M, Y, \theta | X) = g(Y | \beta, M, X) \pi_1(\beta | X, M) \phi(M | \alpha, \sigma^2, X) \pi_2(\alpha | X) \pi_3(\sigma^2)$$

les lois a priori π_1, π_2 sont alors les $G-$priors suivants :

$$\begin{aligned} (\alpha | X) &\sim \mathcal{N}(0, N(\mathbf{X}_1' \mathbf{X}_1)^{-1}), \qquad \mathbf{X}_1 = [\mathbb{1}, X] \\ (\beta | X, M) &\sim \mathcal{N}(0, 4N(\mathbf{X}_2' \mathbf{X}_2)^{-1}), \quad \mathbf{X}_2 = [\mathbb{1}, X, M] \end{aligned}$$

Modèle $G-$prior, Cauchy : Dans ce modèle la loi a priori sur la partie gaussienne du modèle est le même $G-$prior que précédemment et pour la partie logistique on suit la méthodologie définie par Gelman et al (2007). Les variables explicatives sont standardisées, on pose $\widetilde{X} = X - \bar{X}$ et $\widetilde{M} = \frac{M - \bar{M}}{2 s_M}$. La loi a priori sur le paramètre $\widetilde{\beta} = (\widetilde{b}_0, \widetilde{b}_1, \widetilde{b}_2)$ défini par :

$$\text{logit}(\mathbb{E}(Y | \widetilde{X}, \widetilde{M}, \widetilde{W})) = \widetilde{b}_0 + \widetilde{b}_1 \widetilde{M} + \widetilde{b}_2 \widetilde{X}.$$

est alors définie par $\quad \widetilde{b}_0 \sim \mathcal{C}(0, 10), \quad \widetilde{b}_1 \sim \mathcal{C}(0, 2.5), \quad \widetilde{b}_2 \sim \mathcal{C}(0, 2.5)$

Comparaison par simulations : Les trois modèles précédents sont comparés du point de vue de l'estimation de NDE(0), NIE(1). Comme attendu, la qualité de l'estimation dans modèle 1 est meilleure lorsque les données historiques sont pertinentes et reste comparable à celle des autres modèle même dans le cas le plus défavorable.

3 Test sur l'absence d'effets

Une des questions inhérente à l'analyse de médiation est l'existence des effets directs et indirects. La difficulté principale est le test sur l'effet indirect. MacKinnon et al.(2004) ont proposé une procédure bootstrap pour tester l'effet indirect : un intervalle de confiance BCa (bias-corrected and accelerated) est calculé à un niveau α_c fixé et la règle de décision est :

> On rejette l'absence d'effet indirect au niveau α_c lorsque 0 n'est pas dans cet intervalle.

Cette procédure présente deux défauts principaux : un manque de puissance et un mauvais niveau empirique dans le cas des petits échantillons.

Nous proposons de remplacer l'intervalle de confiance BCa par l'intervalle de crédibilité calculé avec le modèle $G-$prior, $G-$prior. En effet, d'une part, le fait d'utiliser des lois a priori non informatives assure que la couverture fréquentiste des intervalles de crédibilité est asymptotiquement égale à $1-\alpha_c$. D'autre part, sur des données simulées nous montrons que ce modèle possède de bonnes propriétés de couverture à distance finie.

Les simulations réalisées indiquent de meilleures performances pour notre procédure non seulement en terme de niveau mais aussi de puissance. (voir table 2).

4 Application :

Les données ont été recueillies dans la région nantaise lors d'une étude longitudinale diligentée par le ministère de l'Education nationale entre 2014 et 2016. Les performances scolaires ont été évaluées par les enseignants à travers une variable binaire X : les élèves ayant atteint ou dépassé le niveau attendu en mathématiques et en français durant l'année scolaire ($X = 1$) ou élève ayant un niveau inférieur à celui-ci ($X = 0$). La modélisation utilisée ne tient pas compte du caractère longitudinal des données mais utilise le premier pas de temps comme historique dans le modèle informatif (voir Figure 1). Les résultats montrent qu'une partie de l'influence positive des résultats scolaires de l'année antérieure sur ceux de l'année suivante transite effectivement à travers une augmentation du sentiment d'efficacité pour les enfants ayant les meilleurs résultats.

$\alpha\beta$	N	NIE(0)	NIE(1)	boot
$\alpha\beta = 0$	30	0.07	0.07	0.01
	50	0.06	0.06	0.04
	100	0.06	0.06	0.04
$\alpha\beta = 0.1$	30	0.06	0.06	0.01
	50	0.08	0.08	0.04
	100	0.04	0.04	0.03
$\alpha\beta = 1.00$	30	0.13	0.13	0.03
	50	0.20	0.20	0.12
	100	0.28	0.28	0.21
$\alpha\beta = 2.00$	30	0.35	0.35	0.24
	50	0.30	0.30	0.23
	100	0.78	0.78	0.73
$\alpha\beta = 4.00$	30	0.81	0.81	0.51
	50	0.82	0.82	0.73
	100	1.00	1.00	1.00

Table 2: La probabilité empirique de rejeter l'absence d'effet indirect en utilisant notre règle de décision et la procédure usuelle (bootstrap). Le modèle de simulation est défini par $X_i \sim \mathcal{B}(0.4)$, $M_i \sim \mathcal{N}(1 + \alpha X_i, 0.75^2)$, $Y_i \sim \mathcal{B}\left(\frac{1}{1+e^{-(-2+\beta M_i+\gamma X_i)}}\right)$ où $\alpha = 2$.

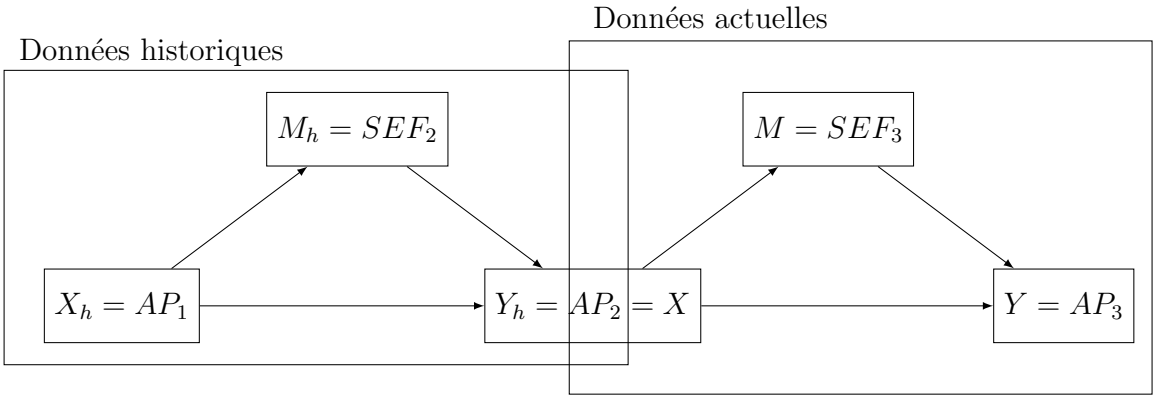


Figure 1: Structural model. AP: Academic Performance, SEF: Self Efficacy Feeling.

Bibliographie

Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y.-S. (2007). A weakly informative default prior distribution for logistic and other regression models. The Annals of Applied Statistics, 2.

Imai, K. and Keele, L. (2010). A general approach to causal mediation analysis. Psychological Methods, 15(4) :309–334.

Launay, T., Philippe, A., and Lamarche, S. (2015). Construction of an informative hierarchical prior for a small sample with the help of historical data and application to electricity load forecasting. TEST, 24(2) :361–385.

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., and Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. Psychological Methods, 7.

MacKinnon, D. P., Lockwood, C. M., and Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. Multivariate Behavioral Research, 39.

Marin, J.-M. and Robert, C. (2014). Bayesian essentials with R. Springer Textbooks in Statistics. Springer Verlag, New York.

Pearl, J. and Mackenzie, D. (2018). The Book of Why : The New Science of Cause and Effect. Basic Books.

Zellner, A. (1971). An Introduction to Bayesian Inference in Econometrics. A Wiley-Interscience publication. Wiley.

La géométrie de l'information appliquée à la robustesse en quantification d'incertitudes

Clément Gauchy [1] & Jérôme Stenger [23]

[1] *DEN-Service d'études mécaniques et thermiques (SEMT), CEA, Université Paris Saclay, F-91191 Gif sur Yvette, France - clement.gauchy@cea.fr*
[2] *EDF R&D Chatou, 6 Quai Watier 78400 Chatou, jerome.stenger@edf.fr*
[3] *Institut de Mathématiques de Toulouse, Université Toulouse*

Résumé. L'analyse de robustesse en quantification d'incertitudes est un champ de recherche récent qui consiste à étudier les quantités statistiques d'intérêt associées aux sorties d'un code de calcul à la perturbation d'une ou de plusieurs distributions de probabilité d'entrée. Ainsi, une méthode pratique d'analyse de robustesse doit se baser sur une bonne perturbation de ces densités. La méthodologie de cet exposé se base sur la distance de Fisher dans la variété des distributions de probabilité, celle ci étant calculée par une méthode numérique inspirée de la mécanique Lagrangienne et consistant à résoudre un système d'équations différentielles ordinaires. Cette définition de la perturbation est ainsi utilisée pour calculer des indices de robustesse basé sur le quantile, les *Perturbed-Law based sensitivity Indices* (PLI).

Mots-clés. Expériences numériques, Perturbation de densité, métrique de Fisher, Echantillonage préférentiel, Quantile, Analyse de sensibilité ...

Abstract. Robustness analysis is an emerging field in the domain of uncertainty quantification. It consists in analysing the statistical quantities of interest of the outputs of a computer model to the perturbation of one or several of its input distributions. Thus, a practical robustness analysis methodology should rely on a coherent definition of a distribution perturbation. A rigorous way of perturbing densities is presented. The proposed methodology is based on the Fisher distance on manifolds of probability distributions. A numerical method to calculate perturbed densities in practice is presented. This method comes from Lagrangian mechanics and consists in solving an ordinary differential equations system. This perturbation definition is then used to compute quantile-oriented robustness indices, the *Perturbed-Law based sensitivity Indices* (PLI).

Keywords. Computer experiments, Density perturbation, Fisher metric, Importance sampling, Quantile, Sensitivity analysis ...

1 Introduction

Dans le domaine de l'industrie, la quantification d'incertitudes consiste à évaluer quantitativement les multiples risques de défaillances des différentes installations (barrages,

centrale nucléaire, avion...) par des méthodes statistiques : on se donne un code de calcul G qui simule un certain phénomène physique (transitoire de température dans une conduite hydraulique, champ de contrainte d'un objet mécanique...). Ce code dépend de paramètres physiques dont on ne connait pas la valeur exacte et que l'on modélise par un vecteur aléatoire $\mathbf{X} = (X_1, ..., X_p)$ de densité de probabilité f, on se placera dans la suite dans le cas indépendant, où chaque X_i a pour densité f_i. Les lois de probabilité sont un *a priori* choisi par l'ingénieur ayant des données et/ou une expertise sur le phénomène physique qu'il étudie. On cherche donc à décrire par des méthodes statistiques l'influence des variables $X_1, ..., X_p$ sur $Y = G(\mathbf{X})$.

Dans la plupart des cas, la densité de $\mathbf{X}$ est considérée représentative des variables impliquées dans le système physique étudié, cette représentativité étant validée par des avis d'experts et des expériences en laboratoire. On cherche à mesurer l'influence du modèle probabiliste ainsi choisi pour une variable d'entrée sur notre quantité d'interêt, ce que l'on appelle l'analyse de robustesse. Dans ce cadre, on introduit une incertitude autour du choix des lois de probabilités en elle mêmes. Ainsi, le PLI (Lemaître et al. [2015], Sueur et al. [2017], Gauchy et al. [2019]) a pour but de mesurer la sensibilité d'une quantité statistique d'intérêt (moyenne, variance, quantile, probabilité de défaillance, ...) émanant de Y par rapport à une perturbation de la densité d'une des variables X_i. Dans notre cas, on cherche à quantifier la variation du quantile d'ordre α de Y (noté q^α) lorsque la densité f_i de la variable X_i devient $f_{i\delta}$ (le quantile devient alors $q^\alpha_{i\delta}$). Le PLI utilisé dans Gauchy et al. [2019] est défini par :

$$S_{i\delta} = \frac{q^\alpha_{i\delta} - q^\alpha}{q^\alpha}.$$

Ce papier s'attache à définir, de façon cohérente, une perturbation de densité de probabilité. Nous nous attardons ici sur la méthodologie developpée dans Gauchy et al. [2019]. Les sections 2 et 3 explicitent la méthode de perturbation utilisant les notions de géométrie de l'information et le calcul de sphères de Fisher La section 4 présente les résultats de la méthode sur une application simple et la section 5 conlue la discussion.

2 Perturbation de densité via la géométrie de l'information

La géométrie de l'information consiste à interpréter un ensemble de lois de probabilités comme une variété Riemannienne ayant une certaine métrique associée. Ainsi, les outils de la géométrie Riemannienne peuvent être appliqués sur des modèles statistiques, ce qui inclut des notions de distance, de géodésiques, etc. Dans le cas des modèles statistiques paramétriques, la métrique Riemannienne n'est rien d'autre que la matrice d'information de Fisher. On considère le modèle paramétrique suivant $\mathcal{S} = \{f_\theta, \theta \in \Theta \subset \mathbb{R}^d\}$, la

métrique associée à la fonction coordonnée θ, appelée métrique de Fisher ou métrique de Fisher-Rao, est définie par :

$$I(\theta) = \mathbb{E}\left[\nabla_\theta \log f_\theta(X)(\nabla_\theta \log f_\theta(X))^T\right] \ . \tag{1}$$

Cette expression est bien connue en statistique, $I(\theta)$ est la matrice d'information de Fisher évaluée au point θ dans le modèle statistique. Cela permet de définir un produit scalaire local dans $\mathcal{S}$:

$$\forall u, v \in \mathbb{R}^d, \ \langle u, v\rangle_\theta = u^T I(\theta) v \ . \tag{2}$$

Cette métrique permet de définir la distance de Fisher. Cette distance entre mesures de probabilité possède une propriété tout à fait remarquable qui est illustrée par la figure 1, qui considère le modèle statistique $\mathcal{S} = \{\mathcal{N}(\mu, \sigma^2), (\mu, \sigma) \in \mathbb{R} \times {\mathbb{R}_+}^*\}$. On remarque que la distance de Fisher entre les Gaussiennes A et B est supérieure à la distance entre les Gaussiennes C et D. En effet, la distance de Fisher permet de mesurer la distinguabilité entre densités de probabilité (voir Amari [1985, p.27] et Gauchy et al. [2019] pour les détails mathématiques).

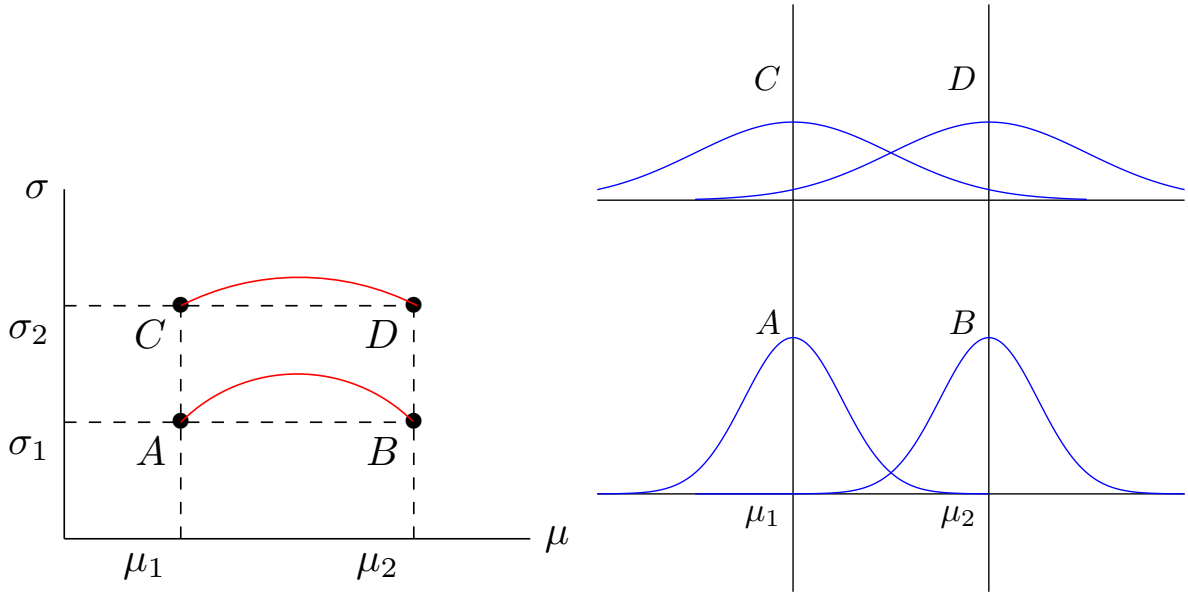


FIGURE 1 – Réprésentations de quatres lois Gaussiennes, les deux courbes rouges sont deux géodésiques dans le plan $(\frac{\mu}{\sqrt{2}}, \sigma)$. La distance entre A et B est plus grande que celle entre C et D. En effet, il est plus difficile de distinguer les paramètres des gaussiennes A et B que C et D.

On définit alors f_δ comme une perturbation de la densité f de niveau δ si la distance de Fisher entre ces deux densités est égale à δ. Ainsi, dans cette méthodologie, on ne

considère pas une seule densité perturbée mais un ensemble de densités. De fait, l'ensemble des densités perturbées de niveau δ sont celles situées sur la sphère de Fisher centrée en f et de rayon δ. Le PLI étant défini pour une densité perturbée spécifique f_δ, cet indice de robustesse sera adapté à notre méthodologie, et le nouvel indice sera le PLI maximal et minimal sur la sphère de Fisher de rayon δ. La question du calcul des sphères de Fisher se pose alors.

3 Calcul de sphères de Fisher

On se place tout d'abord dans un modèle paramétrique $\mathcal{S} = \{f_\theta, \theta \in \Theta \subset \mathbb{R}^d\}$. On peut déterminer une géodésique par la conservation de la quantité suivante, que l'on appelle Hamiltonien :

$$H(p,q) = p\dot{q} - L(t,q,\dot{q}) = \frac{1}{2} p^T I^{-1}(q) p \ , \tag{3}$$

avec la notation $p = \frac{\partial L}{\partial \dot{q}}$. Cette quantité est constante si q est une géodésique comme indiqué dans Gelfand and Fomin [2012]. L'équation (3) est issue de l'équation d'Euler-Lagrange, voir Gelfand and Fomin [2012] pour les détails mathématiques, ce qui implique que (p,q) suit un système d'équations différentielles ordinaires (EDO) appelé équations de Hamilton :

$$\left\{ \begin{array}{rcccl} \dot{q} & = & \frac{\partial H}{\partial p} & = & I^{-1}(q)p \\ \dot{p} & = & -\frac{\partial H}{\partial q} & = & \frac{\partial L(t,q,I^{-1}(q)p)}{\partial q} \end{array} \right. \tag{4}$$

On peut montrer que la conservation de l'Hamiltonien et la condition de la distance de Fisher égale à δ permet de satisfaire le théorème de Cauchy pour le système d'EDO (4) (voir Gauchy et al. [2019]). Comme nous souhaitons calculer les sphères de Fisher centrée en f et de rayon δ, cela se résume donc à la résolution d'un système d'équations différentielles. La figure 2 illustre le résultat obtenu pour différentes méthodes de résolution numérique.

4 Application à un cas simple

Pour illustrer cette nouvelle méthodologie de perturbations de densité, nous nous appuierons sur un exemple industriel simple : un modèle hydraulique simplifié qui permet de calculer la hauteur maximale annuelle d'une rivière soumise à des crues Iooss and Lemaître [2015]. Le code comporte 4 paramètres physiques avec leur distributions associées (cf. Table 1). Le modèle est basé sur une version 1D des équations de Saint Venant :

$$H = \left(\frac{Q}{300 K_s \sqrt{2.10^{-4}(Z_m - Z_v)}} \right)^{0.6} \ . \tag{5}$$

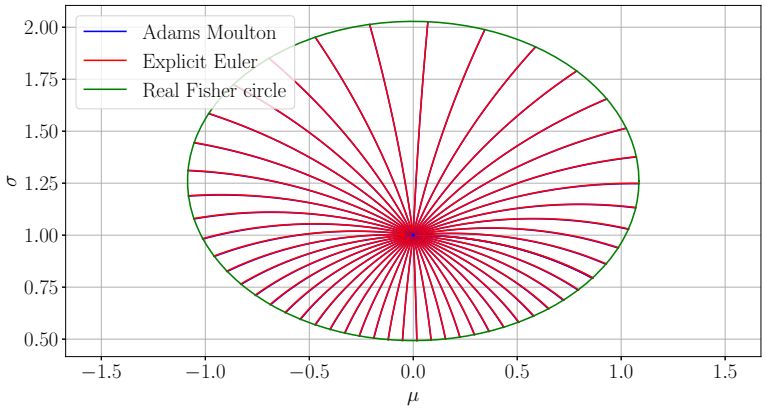


FIGURE 2 – Géodésiques dans le cas gaussien. Le rayon de la sphère de Fisher est $\delta = 1$.

Les résultats sont illustrés dans la Figure 3 comme une fonction du niveau de perturbation δ.

Variable n°	Nom	Description	Distribution de probabilité	Troncature
1	Q	Flot maximal annuel	Gumbel $\mathcal{G}(1013, 558)$	$[500, 3000]$
2	K_s	Coefficient de Strickler	Normale $\mathcal{N}(30, 7.5)$	$[15, +\infty]$
3	Z_v	Niveau aval de la rivière	Triangulaire $\mathcal{T}(50)$	$[49, 51]$
4	Z_m	Niveau amont de la rivière	Triangulaire $\mathcal{T}(55)$	$[54, 56]$

TABLE 1 – Variables d'entrée et lois de probabilités associées.

On remarque dans la figure 3 que les variables Q et K_S ont une variation relative du quantile plus élevée que Z_v et Z_m à δ croissant. On peut donc en conclure que les deux premières variables sont plus sensibles à la perturbation de densité que les deux autres.

5 Conclusion

Basée sur la distance de Fisher, nous avons défini une méthodologie originale de perturbation de densité. L'information de Fisher étant une caractéristique intrinsèque des mesures de probabilités, elle est indépendante de la paramétrisation. Cette méthodologie admet cependant des limites : nous devons nous restreindre à des espaces de dimension finie. Par ailleurs, il est pour le moment difficile d'étendre la méthodologie à des modèles non paramétriques. Cependant, grâce au PLI, cette méthode nous apporte de l'information sur les incertitudes les plus influentes concernant les distributions des variables d'entrées.

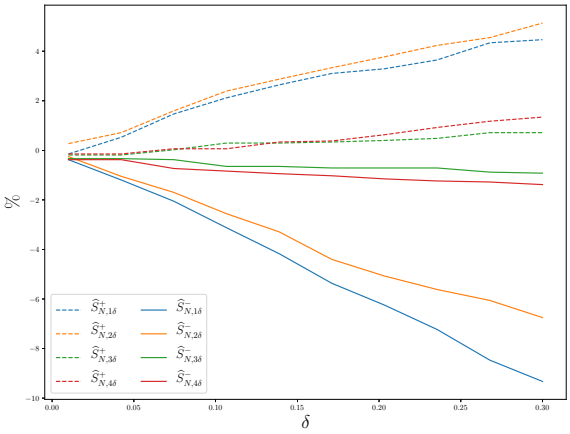


FIGURE 3 – Maximum et minimum du PLI $S_{i\delta}$ pour chaque variable d'entrée du modèle des crues en fonction de δ.

Références

S. Amari. *Differential-Geometrical Methods in Statistics*. Springer New York, New York, NY, 1985.

C. Gauchy, J. Stenger, R. Sueur, and B. Iooss. An information geometry approach for robustness analysis in uncertainty quantification of computer codes. Preprint, December 2019. URL `https://hal.archives-ouvertes.fr/hal-02425477`.

I.M. Gelfand and S.V. Fomin. *Calculus of Variations*. Dover Books on Mathematics. Dover Publications, 2012.

B. Iooss and P. Lemaître. A review on global sensitivity analysis methods. In C. Meloni and G. Dellino, editors, *Uncertainty management in Simulation-Optimization of Complex Systems : Algorithms and Applications*. Springer, 2015.

P. Lemaître, E. Sergienko, A. Arnaud, N. Bousquet, F. Gamboa, and B. Iooss. Density modification based reliability sensitivity analysis. *Journal of Statistical Computation and Simulation*, 85 :1200–1223, 2015.

R. Sueur, B. Iooss, and T. Delage. Sensitivity analysis using perturbed-law based indices for quantiles and application to an industrial case. In *Proceedings of the 10th International Conference on Mathematical Methods in Reliability (MMR 2017)*, Grenoble, France, 2017.

K-BMOM algorithme de clustering robuste

Edouard GENETAY [1] & Adrien SAUMARD [2] & Camille SAUMARD [3]

[1] *ENSAI, Campus de Ker-Lann, 51 Rue Blaise Pascal BP 37203 à 35172 BRUZ Cedex, FRANCE et edouard.genetay@ensai.fr*

[2] *ENSAI, Campus de Ker-Lann, 51 Rue Blaise Pascal BP 37203 à 35172 BRUZ Cedex, FRANCE et adrien.samard@ensai.fr*

[3] *Twice.AI®, 29 Avenue Jean-Janvier 35000 Rennes et csaumard@twice.ai*

Résumé. Les méthodes de clustering classiques telles que K-means souffrent d'un manque de robustesse à la présence de données aberrantes (outliers). Nous proposons un algorithme de clustering robuste basé sur l'utilisation de statistiques de type Median-Of-Means (MOM), une méthode qui s'est déjà avérée efficace en apprentissage supervisé robuste. L'implémentation de l'estimateur que nous proposons (K-BMOM) consiste en la constitution de b sous-échantillons des observations, puis chaque sous-échantillon est clusterisé en affectant chaque point au centre le plus proche pour finalement ne retenir que les centres de clusters qui ont réalisé la médiane de la distorsion K-means parmi les b sous-échantillons. Notre procédure a de meilleures performances que K-means et K-medoids en présence d'outliers ou de distributions à queue lourde. Dans ce contexte, à initialisation donnée, K-BMOM est comparable à K-medians et trimmed-K-means. Enfin, une adaptation de cette procédure fournit une initialisation robuste qui augmente grandement la performance des algorithmes considérés.

Mots-clés. Clustering robuste, Perte empirique K-means, initialisation robuste, Breakdown point

Abstract. Classical clustering methods, such as K-means, suffer from a lack of robustness with respect to outliers. We propose a robust clustering algorithm based on Median-Of-Means statistics, a strategy that has been recently put to emphasis for efficient robust classification. The implementation of our estimator (K-BMOM) begins with the drawing of b subsamples of the observations, after that, each point of the subsamples gets affected to its nearest center and as final step the algorithm outputs the centers of the subsample that lead to the median value of the K-means loss. K-BMOM has better performances than K-means and K-medoids on corrupted or heavy-tailed data while being comparable to K-medians and trimmed-K-means for any given initialisations. Moreover, our procedure supplies also a robust and efficient initialisation method that improves all algorithms' performances.

Keywords. Robust clustering, K-means loss, Robust initialisation, Breakdown point

1 K-BMOM: algorithme de clustering robuste

La sensibilité de la plupart des approches de classification non-supervisée, telles que K-means ou modèles de mélange, à la qualité de l'initialisation d'une part et à la présence d'outliers d'autre part est bien connue. Nous proposons de ce fait d'associer l'estimateur MOM aux méthodes les plus répandues comme K-means puis de mesurer le gain apporté sur les performances de classification non-supervisée.

1.1 Median-of-Means : un estimateur robuste de la moyenne

Grâce aux travaux de Devroye et al. , on sait que Median-of-Means (MOM) est un estimateur de la moyenne μ qui se concentre optimalement autour μ.

Definition 1. Tout d'abord, on note $x_1^n := (x_1 \dots x_n) \in R^n$, S_n l'ensemble des permutations de $\{1, \dots, n\}$ muni de la mesure de probabilité uniforme μ, la médiane d'un ensemble $\mathcal{E} \subset R$, pour une mesure de probabilité λ sur $\mathcal{E}$, est notée med$\mathcal{E}$ et si cette médiane n'est pas unique, on prend la plus petite valeur médiane admissible :

$$\text{med}\mathcal{E} := \inf \left\{ t \in \mathcal{E} \,\middle|\, \lambda\left(\{y \in \mathcal{E} : y \leq t\}\right) \geq \frac{1}{2} \right\}$$

Soit b, t deux entiers tels que $bt = n$ et soit également $\bigcup_{k=1}^{b} B_k$ la partition de $\{1, \dots, n\}$ où $B_k := \{(k-1)t + i \,|1 \leq i \leq t\}$. Median-of-Means (MOM) est alors un estimateur réel randomisé défini de la façon suivante :

$$MOM_{n,b} : \begin{cases} R^n \times S_n & \longrightarrow & R \\ (x_1^n, \sigma) & \longmapsto & \text{med} \quad \left\{ \frac{1}{t} \sum_{i \in B_k} X_{\sigma(i)} : 1 \leq k \leq b \right\} \end{cases}$$

où en pratique, la σ est une permutation aléatoire avec loi uniforme sur S_n.

Theorem 2. *Si $n > 5$ est un entier, $M > 0$, $\delta \in \left[2e^{-n/4}, 1/2\right)$, alors pour n'importe quelle variable aléatoire Y d'espérance μ et de variance* $\text{Var}(Y) < M$*, pour toute précision $\delta \geq e^{1-n/2}$, l'estimateur $MOM_{n,b}$ avec un nombre de blocs $b = \log(1/\delta)$ vérifie*

$$P\left(|MOM_{n,b}(Y_1^n) - \mu| > \sqrt{\frac{96M \log(1/\delta)}{n}} \right) \leq \delta$$

Or, d'après les travaux de Catoni, la meilleure concentration de la moyenne empirique est donnée par l'inégalité de Tchebychev si la seule hypothèse est $\text{Var}(Y) < M$. Cela fait de MOM un bien meilleur estimateur de la moyenne que la moyenne empirique. C'est pourquoi il paraît intéressant de l'injecter dans les estimateurs tels que K-means, là où l'on a recours à une moyenne empirique (voir section 1.3). De plus, MOM octroie une plus grande robustesse aux outliers que la moyenne empirique d'après le survey de Mendelson et Lugosi.

1.2 Le bénéfice d'une version bootstrap de MOM

MOM recourt à une partition des données or clusteriser des données en K classes nécessite d'avoir suffisamment d'observations. Par conséquent, pour éviter une division excessive des données, nous utilisons une version bootstrap de MOM (voir équation 1). Par ailleurs, sur la base d'une définition du breakdown point proche de la définition du livre de Hampel et al :

Definition 3. Soit $\hat{\theta}_.$ un estimateur réel randomisé dont l'aléa est pris par rapport à l'espace probabilisé $(Q, \mathcal{Q}, \mu)$ et dont la loi est invariante par permutation de ses arguments :

$$\forall x_1^n \in R^n, \forall \sigma \in S_n,\ \hat{\theta}_.(x_1 \dots x_n) \overset{\mathcal{D}}{=} \hat{\theta}_.(x_{\sigma(1)} \dots x_{\sigma(n)})$$

On appelle breakdown point probabiliste de $\hat{\theta}_.$, noté $\mathrm{BP}\left(\hat{\theta}_.\right)$, le plus petit nombre d'arguments de $\hat{\theta}_.$ à modifier de sorte que la déviation infligeable à l'estimateur soit infinie avec probabilité au moins $\frac{1}{2}$. On note également $x_1^n y_1^m := (x_1 \dots x_n, y_1, \dots, y_m)$,

$$\mathrm{BP}\left(\hat{\theta}_.\right) := \frac{1}{n} \inf \left\{ q \in N \,\middle|\, \exists x_1^n \in R^n, \mu \left(\sup_{y_1^q \in R^q} |\hat{\theta}_.(x_1^{n-q} y_1^q) - \hat{\theta}_.(x_1^n)| = \infty \right) \geq \frac{1}{2} \right\}$$

Nous avons pu calculer la valeur exacte asymptotique du breakdown point d'une version bootstrap de MOM, définie ci-dessous :

Definition 4. Soit $x_1^n \in R^n$ un n-echantillon et notons $\left(X_j^*\right)_{1 \leq j \leq bt}$ $b \times t$ copies i.i.d. de la variable aléatoire qui vaut x_i avec probabilité $\frac{1}{n}$ pour tout $i \in \{1, \dots, n\}$, toutes définies sur l'espace probabilisé $(\Omega, \mathcal{T}, \mu)$. Median-of-Means bootstrap $(BMOM)$ est alors un estimateur réel randomisé défini comme suit :

$$BMOM_{n,b,t} : \left\{ \begin{array}{ccc} R^n \times \Omega & \longmapsto & R \\ (x_1^n, \omega) & \to & \mathrm{med} \quad \left\{ \frac{1}{t} \sum_{j \in B_k} X_j^*(\omega) : 1 \leq k \leq b \right\} \end{array} \right. \tag{1}$$

où $B_k := \{(k-1)t + i \,|1 \leq i \leq t\}$

Proposition 5. *Pour $n, t, q \in N$, avec n le nombre de données, t la taille des blocs, q le nombre d'outliers, le breakdown point de $MOMB_{n,b,t}$ ne dépend que de la proportion d'outliers et de la taille des blocs puisqu'on peut choisir b de sorte à être le plus robuste possible du fait que* $\lim_{b \to \infty} \mathrm{BP}\left(BMOM_{n,b,t}\right) = 1 - \frac{1}{2^{1/t}} \underset{t \to \infty}{\sim} \frac{\log(2)}{t}$.

1.3 K-BMOM: estimateur robuste de classification non-supervisée

Nous proposons de remplacer la moyenne empirique qui apparaît dans le risque K-means : $\frac{1}{n} \sum_{i=1}^n \min_{1 \leq j \leq K} \|X_i - c_j\|_2^2$ par $BMOM_{n,b,t}$ comme suit :

$$\widehat{c}_{K,n,b,t} \in \underset{\mathbf{c}\in\mathcal{X}^K}{\operatorname{argmin}} \left[BMOM_{n,b,t}\left(\left(\min_{1\le j\le K} \|X_i - c_j\|_2^2 \right)_{1\le i\le n} \right) \right]$$

A l'instar de Lloyd nous proposons une procedure iterative basée sur le risque ci-dessus.

Algorithm 1 K-BMOM

Input: $\{x_1, \ldots, x_n\}$, le nombre de classe K, le nombre de blocs B, leur cardinal n_B (avec la contrainte $n_B > K$) et une partition initiale
Main Loop: jusqu'à avoir fait assez d'itérations

1. Pour tout bloc b, $1 \le b \le B$
   - Tirer au hasard avec remise le bloc b contenant n_B observations, sous contrainte que chaque cluster soit représenté
   - Mettre à jour les centres et calculer la distorsion K-means dans b
2. Calculer la médiane des B distorsions ainsi obtenues et extraire les centres présents dans le bloc réalisant cette médiane
3. Recalculer la partition en affectant chaque point à son centre le plus proche

Output: Les derniers centres obtenus

De plus on propose de rendre robuste l'initialisation K-means++ en une seule itération : tout d'abord, tirer B blocs initialisés par K-means++, puis ne conserver que les centres du bloc realisant le risque médian. on appelle cette initialisation K-MOM++.

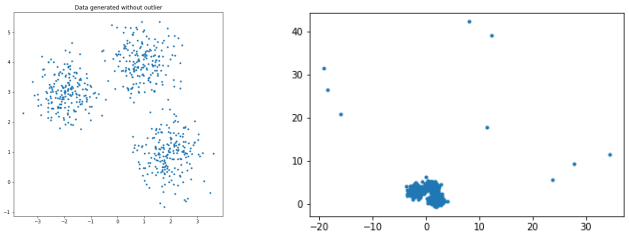
2 Simulations numériques

Afin d'évaluer les performances de l'algorithme K-BMOM par rapport à des déclinaisons robustes de l'algorithme des K-means, le contexte de simulation suivant est considéré: les données sont simulées à partir de $K = 3$ gaussiennes multivariées de dimension $p = 2$. On tire dans chaque cluster 300 points de variance $\sigma^2 = 0.6$ et de moyenne $\mu_1 = [1, 4]$, $\mu_2 = [2, 1]$ and $\mu_3 = [-2, 3]$. Parmi ces $n = 900$ points, on choisit uniformément au hasard $n_{outlier}$ points dont on multiplie les coordonnées par $\beta > 0$. β contrôle la distance de ces outliers à leur cluster d'origine. Nous avons considéré 2 niveaux de corruption $n_{outlier} \in \{9, 27\}$ et 4 niveaux de distance $\beta \in \{5, 10, 20, 40\}$. La figure 1 illustre un jeu de données pour $n_{outlier} = 9$ et $\beta = 10$. Le nombre de clusters est supposé connu, soit $K = 3$.

L'efficacité des méthodes est mesurée d'une part sur la qualité de classification via l'ARI (adjusted Rand Index); d'autre part sur la précision d'estimation des centres des clusters estimés ($\{\hat{c}_1, \ldots, \hat{c}_K\}$) aux moyennes théoriques $\{\mu_1, \ldots, \mu_K\}$. Pour cela, on considère la racine carré de la moyenne des erreurs d'estimation (RMSE) suivante :

$$\text{RMSE}\left(\hat{\mathbf{c}}, \mu\right) = \frac{1}{\sqrt{K}} \min_{\sigma \in S_K} \sqrt{\sum_{k=1}^{K} \left\| \hat{c}_{\sigma(k)} - \mu_k \right\|_2^2}$$

où $\hat{c}_k$ est le centres estimés de la classe k. Afin de rendre le RMSE insensible à la numérotation des classes, on prend le minimum parmi les $\sigma \in S_K$, l'ensemble des permutations d'ordre K et enfin μ_k est le centre théorique de la classe k.



(a) Données avant génération des outliers (b) Données avec outliers

Figure 1: Illustration des données de simulation servant à comparer les algorithmes de clustering en présence d'outliers, dans ce cas $n_{outlier} = 9$ et $\beta = 10$

On compare l'algorithme K-BMOM aux algorithmes des K-means, trimmed-K-means, K-medians et K-medoids; tous initialisés de trois façons : au hasard, par ROBIN et par l'initialisation K-MOM++. Pour K-BMOM, le nombre de blocs b est fixé à 251 et la taille de block $t = 18$; ces hyperparamètres ont été calculés de sorte que la probabilité que le bloc médian contienne un outliers soit inférieure à $0,05$. On répète 100 fois cette expérience.

3 Résultats empiriques

Le tableau 1 présente dans les deux premières colonnes, le RMSE et l'ARI moyen (et leur écart-type associé) pour les 5 algorithmes testés en fonction du type d'initialisation considéré. Dans une troisième colonne, nommée "P[G]" est renseignée la probablité empirique de réaliser l'événement $G := \{RMSE < 2, ARI > 0.75\}$ pour chacun des algorithmes. Les résultats présentés dans le tableau 1 rendent compte du cas ($n_{outlier} = 9$ et $\beta = 40$). On peut remarquer que l'initialisation K-MOM++ est bien meilleure que les autres puisqu'elle

stabilise et améliore les performances de tous les algorithmes. Deuxièmement, l'algorithme K-BMOM est meilleur que tous les algorithmes lorsque l'initialisation est médiocre. Enfin, on peut noter que la petite perte de performance de K-BMOM dans l'estimation des centres est due au fait que cette procédure ne converge pas comme un algorithme EM et que l'on ne garde que les derniers centres rencontrés. Une moyenne sur les dernières itérations donnerait les mêmes résultats que trimmed-K-means bien initialisé.

Table 1: Performances des algorithmes de clustering et de leur initialisation en termes de RMSE dans le cas $n_{outlier} = 9$ et $\beta = 40$

	initialisé au hasard			initialisé avec ROBIN			initialisé avec K-MOM++		
algorithmes	RMSE	ARI	P[G]	RMSE	ARI	P[G]	RMSE	ARI	P[G]
initialisation	0.82 (0.16)	0.93 (0.06)	0.15	0.48 (0.13)	0.97 (0.02)	0.25	**0.37 (0.14)**	**0.97 (0.03)**	**0.99**
K-means	0 (0)	0 (0)	0.0	0 (0)	0 (0)	0.0	0 (0)	0 (0)	0.0
K-medoids	1.05 (0.12)	0.87 (0.06)	0.5	1.03 (0.12)	0.88 (0.06)	0.45	1.07 (0.11)	0.88 (0.06)	0.58
trim-K-means	0.56 (0.29)	0.92 (0.06)	0.53	0.48 (0.33)	0.93 (0.06)	0.51	**0.16 (0.04)**	**0.98 (0.01)**	**0.99**
Kmedian	0.47 (0.19)	0.97 (0.03)	0.97	0.46 (0.19)	0.97 (0.02)	0.99	0.45 (0.17)	0.97 (0.03)	0.99
K-MOM	**0.36 (0.12)**	**0.98 (0.01)**	**0.96**	**0.36 (0.12)**	**0.98 (0.01)**	**0.99**	0.35 (0.1)	0.98 (0.01)	1.0

Bibliographie

Catoni, O. (2012). Challenging the empirical mean and empirical variance: a deviation study, *Annales de l'IHP Probabilités et statistiques*, 48(4), pp. 1148-1185

Lugosi, G. and Mendelson, S. (2019). Mean estimation and regression under heavy-tailed distributions: A survey, *Foundations of Computational Mathematics*, Springer, 19(5), pp. 1145-1190

Hampel, F. R. et Ronchetti, E. M. et Rousseeuw, P. J. and Stahel, W. A. (2011). Robust statistics: the approach based on influence functions, John Wiley & Sons, 196, pp. 98

Lecué, G. et Lerasle, M. (2017). Robust machine learning by median-of-means: theory and practice, *arXiv preprint arXiv:1711.10306*

Devroye, L. et Lerasle, M. et Lugosi, G. et Oliveira, R. I. et al. (2016). Sub-Gaussian mean estimators, *The Annals of Statistics*, Institute of Mathematical Statistics, 44(6), pp. 2695-2725

Al Hasan, M. and Chaoji, V. and Salem, S. and Zaki, M. J. (2009). Robust partitional clustering by outlier and density insensitive seeding, *Pattern Recognition Letters*, Elsevier, 30(11), pp. 994-1002

Explorer l'influence conjointe de prédicteurs fonctionnels sur une réponse réelle via une régression pénalisée

Girault Gnanguenon Guesse [1], Patrice Loisel [2], Bénedicte Fontez [3], Thierry Simonneau [4] & Nadine Hilgert [5]

[1] *MISTEA, Université Montpellier, Institut Agro, INRAE, Montpellier, France.*
girault-bogues.gnanguenon-guesse@inrae.fr
[2] *MISTEA, Université Montpellier, Institut Agro, INRAE, Montpellier, France.*
patrice.loisel@inrae.fr
[3] *MISTEA, Université Montpellier, Institut Agro, INRAE, Montpellier, France.*
benedicte.fontez@supagro.fr
[4] *LEPSE, Université Montpellier, Institut Agro, INRAE, Montpellier, France.*
thierry.simonneau@inrae.fr
[5] *MISTEA, Université Montpellier, Institut Agro, INRAE, Montpellier, France.*
nadine.hilgert@inrae.fr

Résumé. En agronomie, l'avènement de nouveaux capteurs permet d'observer à haute fréquence des dynamiques de variables agro-environnementales affectant la production. Cette nouvelle situation nécessite de faire appel à d'autres outils statistiques ou de les révolutionner afin de tirer de la connaissance de ces données dites fonctionnelles. Dans un contexte où la production est affectée par un effet combiné complexe des différentes dynamiques de variables agro-environnementales, nous proposons une nouvelle approche exploitant des distributions conjointes de variables fonctionnelles pour expliquer une variable réelle (scalaire) représentant un facteur de production. Les simulations effectuées permettent de mettre en exergue une approche exploratoire se rapprochant des techniques de type boosting qui permet d'identifier diverses distributions conjointes associées aux courbes explicatives, d'y associer des coefficients via des régressions linéaires pénalisées et structurées puis de retenir une distribution conjointe optimale expliquant au mieux la variable à prédire. Cette approche a aussi l'avantage de pouvoir intégrer au besoin dans la modélisation des connaissances dites "connaissances d'experts" provenant de la littérature ou autres afin d'améliorer la fiabilité de l'approche statistique proposée. Cette approche qui se veut exploratoire peut être utilisée comme un modèle prédictif sous certaines conditions. Développée à la base pour l' agronomie, cette approche est générique et peut être utilisée pour résoudre des problèmes de type scalar-on function avec comme hypothèse principale l'identification d'effets combinés de variables explicatives fonctionnelles. Une limite de cette approche est un risque de surestimation mais divers critères permettent d'y pallier. L'utilisation de l'approche pour analyser des données réelles permet d'identifier des combinaisons de classes de température - irradiance et de moments de la journée affectant l'accumulation d'anthocyanes et de polyphénols dans la baie de raisin.

Mots-clés. exploration de données fonctionnelles, distribution conjointe, régression linéaire pénalisée, critères d'information, agronomie.

Abstract. In agronomy, the development of new sensors has allowed to observe at high frequency the dynamics of agri-environmental variables affecting production. This new situation

requires using other statistical tools or revolutionizing them in order to learn from this so-called functional data. In a context where production is affected by a complex combined effect of these different dynamics of agri-environmental variables, we propose a new approach using joint distributions of functional variables to explain a real (scalar) variable representing a production factor. Simulations carried out highlight an exploratory approach closed to boosting techniques that identifies various joint distributions associated to the explanatory curves, associates coefficients to each of them via penalized and structured linear regressions and then selects an optimal joint distribution that best explains the variable to be predicted. This approach has the additional advantage of being able to integrate, if necessary, so-called "expert knowledge" from the literature or other sources into the modeling process in order to improve the reliability of the proposed statistical approach or advise on these "expert knowledge". This exploratory approach can be used as a predictive model under certain conditions. Developed initially for agronomy, it is generic and can be used to solve scalar-on-function problems with the main goal of identifying combined effects of functional explanatory variables. One limitation of this approach is the risk of overestimation, but various criteria are available inside the approach to overcome it. Using the approach to analyze real data allows to identify associations of temperature - irradiance and time that affect the accumulation of anthocyanins and polyphenols in grape berries.

Keywords. functional data mining, joint distribution, penalized linear regression, information criteria, agronomy

1 Introduction

De nos jours, plusieurs domaines d'activités sont révolutionnés par l'avènement des données massives. Ces données massives sont considérées de diverses manières parmi lesquelles la grande famille des données fonctionnelles regroupant courbes, spectres, images, etc. Selon Ferraty et Vieu (2006), une variable aléatoire $\mathcal{X}$ est dite fonctionnelle si ses valeurs sont dans un espace de dimension infinie et une observation X de $\mathcal{X}$ est appelée donnée fonctionnelle. En réalité, seulement quelques points discrets du phénomène continu sont observés $\mathcal{X} = \{X(t) \ : \ t \in T\}$.

L'un des axes majeurs de recherche autour de ces données concerne leur implication dans des problèmes de régression. Dans la littérature, ces problèmes sont habituellement classés en 3 catégories en fonction du rôle joué par les données fonctionnelles (Reiss et *al.* (2010) ; Ramsay et Silverman (2005)). On distingue les régressions "scalar-on-function", "function-on scalar" et "function-on-function". Dans cet exposé, nous nous intéresserons au problème de type 'scalar-on-function' où la variable à prédire est un scalaire et le prédicteur, une fonction. Plus précisément, nous nous intéresserons à un problème où les prédicteurs peuvent être deux ou plusieurs variables fonctionnelles. Diverses méthodes existent pour résoudre les régressions de type 'scalar-on-function' et le lecteur pourra se référer à Reiss et *al.* (2017) qui en présente une revue. L'approche proposée nommée SPICEFP (Sparse and Structured Procedure to Identify Combined Effects of Functional Predictors) permet de s'intéresser spécifiquement à l'hypothèse d'influence conjointe des prédicteurs fonctionnels. Nous la présentons brièvement dans la suite de ce texte et présenterons également quelques résultats.

2 L'approche proposée

Considérons deux variables explicatives fonctionnelles que sont $\tau = \{\tau_i(t) : t \in T;\ i = 1, \ldots, n$ et $\mathcal{I} = \{\mathcal{I}_i(t) : t \in T;\ i = 1, \ldots, n\}$, toutes deux des fonctions observées aux mêmes instants t. Considérons d'un autre côté, une variable réponse $y_i \in \mathbb{R},\ i = 1, \ldots, n$ que l'on souhaiterait expliquer par τ et $\mathcal{I}$ en faisant l'hypothèse d'une influence conjointe des deux variables explicatives fonctionnelles sur y. L'approche SPICEFP permet d'atteindre cet objectif. Sa mise en oeuvre nécessite cinq étapes à savoir :

1. transformer (catégoriser) des variables fonctionnelles :
   - *la catégorisation* : pour un individu i fixé, catégorisons en n_τ ($n_\mathcal{I}$) classes l'observation τ_i ($\mathcal{I}_i$) suivant une échelle linaire. Les $n_\tau + 1$ ($n_\mathcal{I} + 1$) bornes de classes nécessaires sont : $l(v),\ v = 1, 2, \ldots, n_\tau + 1$ ($L(w),\ w = 1, 2, \ldots, n_\mathcal{I} + 1$). Leurs valeurs peuvent être obtenues par l'équation (2.1). Les modalités utilisés pour la catégorisation de τ_i ($\mathcal{I}_i$) s'écrivent sous la forme $[l(v), l(v+1)[, v = 1, \ldots, n_\tau$ ($[L(w), L(w+1)[, w = 1, \ldots, n_\mathcal{I}$).
   $$l(v) = \underline{\tau} + (v - 1)\left(\frac{\bar{\tau} - \underline{\tau}}{n_\tau}\right),\ v = 1, \ldots, n_\tau + 1 \tag{2.1}$$
   avec $\underline{\tau} \in \mathbb{R}$ et $\bar{\tau} \in \mathbb{R}$ les valeurs (réelles) minimale et maximale de τ. Précisons que n_τ ($n_\mathcal{I}$) est à fixer afin de calculer $l(v)$ ($L(w)$).
   - *l'obtention d'une distribution conjointe en effectif* : à partir d'un tableau de contingence C_i^u, de dimension $(n_\tau \times n_\mathcal{I})$ dont les valeurs $C^u_{i,(v,w)}$ sont obtenues via (2.2). Les modalités de la distribution conjointe en effectif obtenues sont notées $[l(v), l(v+1)[\_[L(w), L(w+1)[, v = 1, \ldots, n_\tau,\ w = 1, \ldots, n_\mathcal{I}$. Elles seront appelées modalités conjointes et sont au nombre de $n_\tau \times n_\mathcal{I}$.
   $$C^u_{i,(v,w)} = \sum_{t=1}^{T} \mathbb{1}_{\tau_i(t) \in [l(v), l(v+1)[,\ \mathcal{I}_i(t) \in [L(w), L(w+1)[}\ ;\ v = 1, \ldots, n_\tau\ ;\ w = 1, \ldots, n_\mathcal{I} \tag{2.2}$$
   avec $C^u_{i,(v,w)} \in \mathbb{N}$; $C_i^u \in \mathbb{N}^{n_\tau \times n_\mathcal{I}}$; $(\tau_i(t), \mathcal{I}_i(t)) \in \mathbb{R}^2$; $\sum_{v=1}^{n_\tau} \sum_{w=1}^{n_\mathcal{I}} C^u_{i,(v,w)} = Card(T)$ et $u = (n_\tau, n_\mathcal{I}) \in \mathbb{N}^2$
   - *la vectorisation (empilement colonne après colonne) et la transposition du tableau de contingence* C_i^u : elles donnent
   $$X_i^u = {}^\intercal Vect(C_i^u);\ X_i^u \in \mathbb{R}^{(n_\tau n_\mathcal{I})} \tag{2.3}$$
   un vecteur ligne de longueur $n_\tau \times n_\mathcal{I}$ qui représente pour u fixé, le nombre d'instants t au cours desquels un individu i a été observé dans chacune des $n_\tau \times n_\mathcal{I}$ conditions décrites par les modalités conjointes. On obtient ainsi une matrice X^u, dont chaque ligne X_i^u correspond à un individu. A cette matrice de nouvelles variables explicatives X^u, on rajoute l'information relative à la proximité des modalités conjointes crées en utilisant un graphe $G^u(V^u, E^u)$ où V^u représente X^u et E^u l'ensemble des arêtes liant deux modalités conjointes proches. Deux modalités conjointes sont dites proches si les classes suivant la variable τ (indexées par v) ou [1] les classes suivant la variable $\mathcal{I}$

1. ou exclusif

(indexées par w) sont consécutives comme le montre la figure 1. Cette figure présente un exemple de catégorisation d'un couple de variables longitudinales en $n_\tau \times n_\mathcal{I} = 3 \times 3$ modalités conjointes ainsi qu'une une identification par 4 flèches des modalités conjointes voisines de $[v_2, v_3[\_[w_2, w_3[$ indexées par $j : (v = 2, w = 2)$

- *la construction de diverses distributions conjointes en vue d'en identifier une optimale* : ne connaissant pas a priori la valeur optimale du vecteur u que nous noterons u^*, il est proposé de l'identifier dans le cadre de la mise en oeuvre de l'approche. Étant donné qu'à partir d'un vecteur de catégorisation u nous obtenons une matrice X^u et un graphe $G^u(V^u, E^u)$ associé, explorer différents vecteurs u revient donc à explorer différents graphes explicatifs, tous issus des deux variables fonctionnelles τ et $\mathcal{I}$.

2. effectuer un Generalized Fused Lasso (Tibshirani et Taylor (2011)) sur chaque graphe indexé par u à partir de l'équation (2.4) :

$$\beta^{u,GFL} = \underset{\beta^u}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{n} (y_i - X_i^u \beta^u)^2 \right\} + \lambda_p^u \sum_{v=1}^{n_\tau} \sum_{w=1}^{n_\mathcal{I}} |\beta^u| + \lambda_f^u \sum_{(j,k) \in E^u} |\beta_j^u - \beta_k^u| \quad (2.4)$$

avec : $\lambda_p^u \geq 0$ et $\lambda_f^u \geq 0$ les paramètres de régularisation à optimiser. Pour $j = (v, w)$ fixé, les couples (j, k) relatifs à j et contenus dans E sont $(j, k)_1 = ((v, w), (v + 1, w))$ et $(j, k)_2 = ((v, w), (v, w + 1))$.

3. utiliser un critère pour choisir le meilleur graphe de prédicteurs $G^{u^*}(V^{u^*}, E^{u^*})$ et les coefficients estimés $\hat{\beta^{u^*}}$ associés
4. calculer les résidus associés au meilleur modèle $\varepsilon^{u^*} = y_i - X_i^{u^*} \hat{\beta^{u^*}}$
5. vérifier les conditions d'arrêts pour :
   - retourner à l'étape 2 en remplaçant la variable à prédire par les résidus du meilleur modèle ε^{u^*} obtenus à l'étape 4
   - ou arrêter l'approche

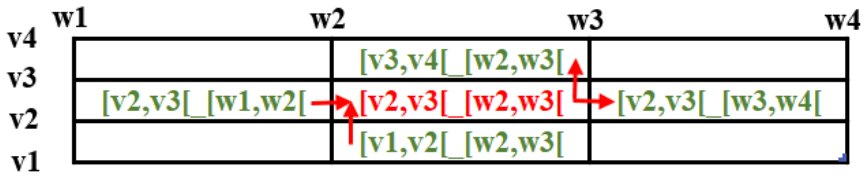


FIGURE 1 – Voisinage pris en compte dans le cadre du Generalized Fused LASSO

3 Quelques résultats

Nous avons tout d'abord illustré notre approche avec des simulations pour bien comprendre ses caractéristiques. Nous nous sommes donnés pour cela des variables explicatives de température et d'irradiance (issues d'expérimentations du projet européen INNOVINE, pour étudier

les effets combinés d'une exposition des baies de raisin plus ou moins forte au soleil et d'une température de l'air normale ou élevée de quelques degrés.) ainsi qu'un vecteur de coefficients parcimonieux. La variable à prédire (variable fictive pour les simulations) a été simulée en associant le tableau de contingence des variables explicatives, le vecteur de coefficients qu'il s'agissait d'estimer et un bruit Gaussien. La figure 2 illustre à gauche un exemple de vecteur de coefficients simulé. Dans cet exemple, on simule une influence positive au coeur d'une zone d'influence négative pour des valeurs faibles de température et d'irradiance sur la variable à prédire. Les cases blanches correspondent à aucune influence (coefficients nuls) des variables explicatives et les cases noires correspondent aux modalités conjointes n'ayant jamais été observées dans le cadre de ces données. L'objectif ici est d'identifier un modèle qui respecte la parcimonie simulée et estime les zones d'influence négative et positive, tout en tenant compte de la continuité des valeurs d'une case à l'autre. L'estimation des coefficients avec notre approche est illustrée avec le graphique de droite. L'estimation des zones d'influence et non influence sont relativement bien reproduites.

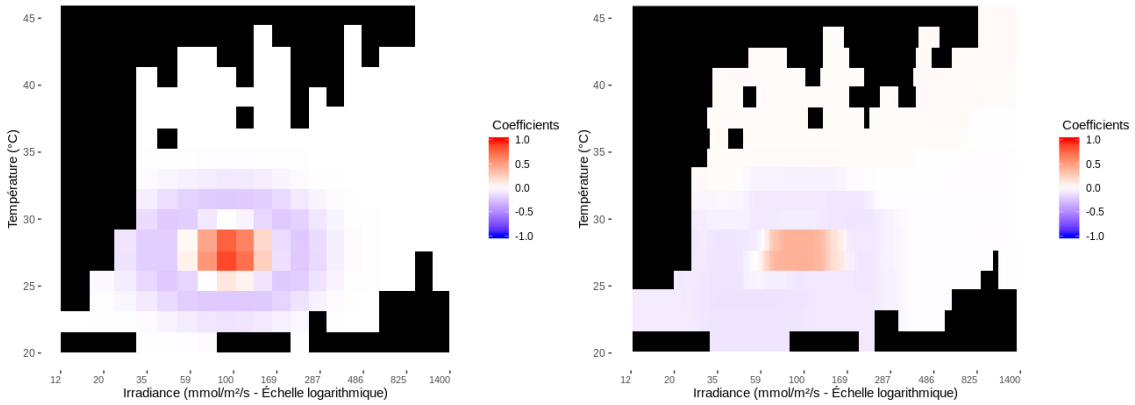


FIGURE 2 – Coefficients simulés (à gauche) et coefficients estimés (à droite).

Cette approche a été appliquée sur les variations hebdomadaires d'indices de Ferari (variables réelles issues des expérimentations d'INNOVINE, qui sont une mesure non destructive de la qualité des baies de raisin). Les résultats seront présentés lors de la conférence.

4 Conclusion

SPICEFP est une approche exploratoire utilisant des outils de la statistique inférentielle. Son but primordial est de fournir des modalités conjointes participant à l'explication d'une variable réponse. Elle sous-entend une influence conjointe des prédicteurs et est conçue pour une exploration dans ce sens. L'un de ses atouts est sa capacité à identifier à une nouvelle itération, une nouvelle distribution conjointe permettant de mieux expliquer la variable à prédire. Ce faisant, on augmente les risques de surestimation. D'où une sensibilité de l'approche à la surestimation lorsque l'erreur de mesure associée à la variable à prédire est élevée. Deux techniques développées lors de la conception de l'approche permettent d'identifier les cas de surestimation.

5 Remerciements

Ce travail a bénéficié d'une aide de l'État gérée par l'Agence Nationale de la Recherche au titre du programme d'Investissements d'Avenir portant la référence ANR-16-CONV-0004 (DigitAg). Les données présentées ont été acquises au cours du projet INNOVINE, financé par le septième programme-cadre de la Communauté européenne (FP7/2007-2013), dans le cadre de la convention de subvention No. FP7-311775.

Bibliographie

Ferraty, F. et Vieu, P. (2006), *Nonparametric Functional Data Analysis : Theory and Practice*, Springer Series in Statistics, Springer-Verlag, New York.

Ramsay J. et Silverman B.W. (2005), *Functional Data Analysis*, Springer Series in Statistics, Springer.

Reiss, P.T., Goldsmith J., Shang H.L. et Ogden R.T. (2017). Methods for scalar-on-function regression, *International Statistical Review*, 85(2), pp. 228–249.

Reiss, P.T., Huang L. et Mennes M. (2010). Fast function-on-scalar regression with penalized basis expansions, *The international journal of biostatistics*, 6(1), article 28.

Tibshirani, R.J. et Taylor, J. (2011). The solution path of the generalized lasso, *Ann. Statist.*, 39(3), pp. 1335-1371.

Tests minimax pour la détection d'une rupture dans un processus de Poisson

Fabrice Grela[1,2] & Magalie Fromont[1,3] & Ronan Le Guével[1,4]

[1] *Univ Rennes, CNRS, IRMAR - UMR 6625, F-35000 Rennes, France*

[2] *fabrice.grela@univ-rennes2.fr*

[3] *magalie.fromont@univ-rennes2.fr*

[4] *ronan.leguevel@univ-rennes2.fr*

Résumé. Nous nous intéressons ici à la question de la détection d'une rupture caractérisée par un saut dans l'intensité d'un processus de Poisson, définie par rapport à une mesure Ldt $(L > 0)$ sur $[0, 1]$. Ce travail, qui peut être vu comme une étape préliminaire à la construction de procédures de localisation de rupture, présente une étude minimax non asymptotique. En considérant la distance usuelle de $\mathbb{L}^2([0, 1])$, nous établissons les vitesses de séparation minimax sur différentes classes d'alternatives, définies selon la connaissance ou non de la position et/ou de la taille du saut. Nous montrons ainsi que la connaissance de la position ou de la taille du saut permet d'obtenir une vitesse de séparation minimax d'ordre $L^{-1/2}$, et que l'adaptation en ces deux paramètres simultanément dégrade la vitesse d'un facteur $(\log \log L)^{1/2}$. Ces résultats, cohérents avec les résultats asymptotiques ou non asymptotiques disponibles en modèle gaussien classique, sont valables à la fois dans le cas où l'intensité de référence est fixée a priori, et dans le cas général, où l'intensité de référence n'est plus fixée. Une démarche progressive est proposée pour la construction de tests minimax et adaptatifs au sens du minimax, allant des tests simples de Neyman-Pearson jusqu'aux tests basés sur des approches d'agrégation. Les propriétés non asymptotiques de ces tests sont établies par le biais d'une astuce de conditionnement dans le cas général pour le risque de première espèce, et de nouvelles inégalités exponentielles pour des suprema de martingales carrées pour le risque de deuxième espèce.

Mots-clés. Détection de rupture, processus de Poisson, test minimax, test adaptatif.

Abstract. We here focus on the question of detecting a single change point characterized by a jump in the intensity of a Poisson process, defined with respect to some measure Ldt $(L > 0)$ over the interval $[0, 1]$. This work, which can be viewed as a preliminary step to the construction of change point localization procedures, presents a nonasymptotic minimax study. By considering the usual distance of $\mathbb{L}^2([0, 1])$, we establish the minimax separation rates over various classes of alternatives, defined according to whether or not the jump position and/or size are known. We thus prove that the knowledge of the position or the size of the jump allows to obtain a minimax separation rate of order $L^{-1/2}$, and that the adaptation to those two parameters simultaneously deteriorates the rate by a $(\log \log L)^{1/2}$ factor. These results, that are consistent with available asymptotic and nonasymptotic results in the classical Gaussian framework, are valid in the case where the reference intensity is fixed a priori, as well as in the general case, where the reference

intensity is not fixed. A progressive methodology is proposed for the construction of minimax and minimax adaptive tests, starting from single Neyman-Pearson tests up to aggregation-based tests. The nonasymptotic properties of those tests are established from a conditioning trick in the general case as for the first kind error rate, and new exponential inequalities for suprema of square martingales as for the second kind error rate.
Keywords. Change point detection, Poisson process, minimax test, adaptive test.

1 Introduction

Les suites d'occurrences d'événements aléatoires observées sur un intervalle de temps sont généralement modélisées par des processus ponctuels. Il semble par exemple désormais admis que les occurrences de certaines tentatives d'intrusion dans des systèmes informatiques peuvent être modélisées par des processus de Poisson ou des processus de Hawkes lorsque l'on considère des phénomènes de contagion (Peng et al. (2016), Baldwin et al. (2017)). La question de la détection de ruptures dans la loi de processus ponctuels peut donc traduire des problèmes concrets de détection de changements de régimes dans les tentatives d'intrusion, qui sont des enjeux-clés de la sécurité numérique.
Nous nous intéressons plus particulièrement à la détection d'une rupture dans la loi P_λ d'un processus de Poisson $N = (N_t)_{t\in[0,1]}$, d'intensité $\lambda \in \mathbb{L}^2([0,1])$ par rapport à une mesure Ldt ($L>0$). Nous envisageons cette question sous l'angle d'un test d'homogénéité du processus de Poisson N et des vitesses de séparation minimax correspondantes sur des classes d'alternatives caractérisées par un saut, de position et de taille connues ou inconnues, de l'intensité λ, pour la distance d associée à la norme usuelle $\|\cdot\|_2$ de $\mathbb{L}^2([0,1])$.
Avant d'aborder le cas général du test d'homogénéité, nous nous penchons sur le cas plus simple du test d'adéquation à un processus de Poisson d'intensité λ_0 constante connue. Pour simplifier les notations, on confondra la fonction λ_0 constante sur $[0,1]$, et sa valeur en n'importe quel point de $[0,1]$. L'ensemble $\mathcal{H}_0$ désignera donc dans une première partie le singleton $\{\lambda_0\}$, puis dans une deuxième partie l'ensemble des intensités constantes. Le sous-ensemble $\mathcal{H}$ de $\mathbb{L}^2([0,1])\backslash\mathcal{H}_0$ désignera une classe d'alternatives pour λ.
Pour $(\alpha,\beta) \in (0,1)^2$ et tout test non randomisé ϕ_α de (H_0):"$\lambda \in \mathcal{H}_0$" contre (H_1):"$\lambda \in \mathcal{H}$", de niveau α (i.e. tel que $\sup_{\lambda\in\mathcal{H}_0} P_\lambda(\phi_\alpha = 1) \leq \alpha$), la vitesse de séparation uniforme de niveau β de ϕ_α sur la classe d'alternative $\mathcal{H}$ (pour d) est définie (c.f. Baraud (2002)) par :

$$\mathrm{SR}(\phi_\alpha, \mathcal{H}, \beta) = \inf\{r \geq 0 : \sup_{\lambda\in\mathcal{H},\ d(\lambda,\mathcal{H}_0)\geq r} P_\lambda(\phi_\alpha = 0) \leq \beta\}.$$

La vitesse de séparation minimax de niveaux α et β sur $\mathcal{H}$ (pour d) est alors définie par :

$$\mathrm{mSR}(\mathcal{H}, \alpha, \beta) = \inf_{\phi_\alpha,\ \sup_{\lambda\in\mathcal{H}_0} P_\lambda(\phi_\alpha=1)\leq\alpha} \mathrm{SR}(\phi_\alpha, \mathcal{H}, \beta).$$

Remarquons que les définitions ci-dessus s'étendent naturellement aux tests randomisés.

2 Détection de rupture dans un processus de Poisson d'intensité de référence λ_0 fixée

Dans cette partie, l'objectif est de tester l'hypothèse nulle selon laquelle le processus de Poisson N est homogène, d'intensité constante connue λ_0 contre l'hypothèse alternative selon laquelle N est d'intensité λ définie par $\lambda(t) = \lambda_0 + \delta \mathbb{1}_{(\tau,1]}(t)$ pour tout $t \in (0,1)$ avec $\delta \in (-\lambda_0, +\infty) \setminus \{0\}$ et $\tau \in (0,1)$, en distinguant différents cas selon la connaissance ou non de la position τ et/ou de la taille δ du saut d'intensité. Afin de formaliser les hypothèses des différents problèmes de test envisagés, nous introduisons donc $\mathcal{H}_0 = \{\lambda_0\}$, et pour $(\delta^*, \tau^*, R) \in ((-\lambda_0, +\infty) \setminus \{0\}) \times (0,1) \times (0, +\infty)$ les classes d'alternatives suivantes :

$$\begin{array}{lcl}
\mathcal{H}_{\lambda_0,\delta^*,\tau^*} & = & \{\lambda : \lambda(t) = \lambda_0 + \delta^* \mathbb{1}_{(\tau^*,1]}(t)\}, \\
\mathcal{H}_{\lambda_0,\tau^*} & = & \{\lambda : \exists \delta \in (-\lambda_0, +\infty) \setminus \{0\},\ \lambda(t) = \lambda_0 + \delta \mathbb{1}_{(\tau^*,1]}(t)\}, \\
\mathcal{H}_{\lambda_0,\delta^*} & = & \{\lambda : \exists \tau \in (0,1),\ \lambda(t) = \lambda_0 + \delta^* \mathbb{1}_{(\tau,1]}(t)\}, \\
\mathcal{H}_{\lambda_0}(R) & = & \{\lambda : \exists \delta \in (-\lambda_0, R - \lambda_0) \setminus \{0\}, \exists\ \tau \in (0,1),\ \lambda(t) = \lambda_0 + \delta \mathbb{1}_{(\tau,1]}(t)\}.
\end{array}$$

Pour les problèmes de test de (H_0):"$\lambda \in \mathcal{H}_0$" contre (H_1):"$\lambda \in \mathcal{H}_{\lambda_0,\delta^*,\tau^*}$" ou (H_1):"$\lambda \in \mathcal{H}_{\lambda_0,\tau^*}$", des tests randomisés, uniformément plus puissants ou uniformément plus puissants parmi les tests sans biais, peuvent être construits à partir des rapports de vraisemblances. Nous montrons que ces tests, basés sur la statistique de comptage $N((\tau^*, 1])$, atteignent la vitesse de séparation minimax, d'ordre $L^{-1/2}$ dans le deuxième problème. Pour les problèmes de test liés aux classes d'alternatives $\mathcal{H}_{\lambda_0,\delta^*}$ et $\mathcal{H}_{\lambda_0}(R)$, traitant donc de l'adaptation par rapport à la position de saut, les statistiques de test sont plus complexes, basées sur des approches d'agrégation, étroitement liées aux tests multiples.
Notons au préalable que pour $\lambda = \lambda_0 + \delta \mathbb{1}_{(\tau,1]}$ et $\mathcal{H}_0 = \{\lambda_0\}$, $d(\lambda, \mathcal{H}_0) = |\delta|\sqrt{1-\tau}$.

2.1 Adaptation à la position de saut

Considérant le problème de test de (H_0):"$\lambda \in \mathcal{H}_0$" contre (H_1):"$\lambda \in \mathcal{H}_{\lambda_0,\delta^*}$" pour $\delta^* \in (-\lambda_0, +\infty) \setminus \{0\}$ (connu), on obtient pour la vitesse de séparation minimax sur $\mathcal{H}_{\lambda_0,\delta^*}$ la borne inférieure suivante.

Proposition 1 (Borne inférieure)**.** *Soit* $(\alpha, \beta) \in (0,1)^2$ *vérifiant* $\alpha + \beta < 1$ *et* $\delta^* \in (-\lambda_0, +\infty) \setminus \{0\}$. *Pour* $C_1(\alpha, \beta) = 1 + 4(1 - \alpha - \beta)^2$, *et* $L \geq \lambda_0 \log C_1(\alpha, \beta)/(\delta^*)^2$, *on a :*

$$\mathrm{mSR}(\mathcal{H}_{\lambda_0,\delta^*}, \alpha, \beta) \geq \sqrt{\lambda_0 \log C_1(\alpha, \beta)/L}.$$

Les arguments de preuve sont adaptés de la théorie bayésienne développée notamment par Ingster (1993) et Baraud (2002) dans des cadres asymptotique et non asymptotique.

Pour montrer que cette borne inférieure est d'ordre optimal, nous construisons un test de niveau α, basé sur l'agrégation de statistiques de comptage driftées. Pour tout $\alpha \in (0,1)$ et $L \geq 1$, on définit :

$$\phi_\alpha^{(1)} = \mathbb{1}_{\left\{S_{\lambda_0,\delta^*}(N) > s_{\lambda_0,1-\alpha}\right\}},$$

où $S_{\lambda_0,\delta^*}(N) = \sup_{t\in[0,1]} \left(\mathrm{sgn}(\delta^*)(N((t,1]) - \lambda_0 L(1-t)) - |\delta^*|L(1-t)/2\right)$ et $s_{\lambda_0,1-\alpha}$ est le $(1-\alpha)$-quantile de $S_{\lambda_0,\delta^*}(N)$ sous l'hypothèse (H_0). Le résultat ci-dessous, qui découle des résultats sur les suprema et infima de processus de Poisson de Pyke (1959) et Takács (1965), montre que la vitesse de séparation uniforme du test $\phi_\alpha^{(1)}$ sur $\mathcal{H}_{\lambda_0,\delta^*}$ atteint la vitesse minimax à une constante multiplicative près. Ainsi, l'adaptation à la position de saut, quand la taille de saut est connue, peut se faire sans perte pour la vitesse.

Proposition 2 (Borne supérieure)**.** *Soit* $(\alpha,\beta) \in (0,1)^2$ *et* $\delta^* \in (-\lambda_0, +\infty) \setminus \{0\}$. *Il existe* $C(\alpha,\beta,\delta^*) > 0$ *telle que pour tout* $L \geq 1$, $\mathrm{SR}(\phi_\alpha^{(1)}, \mathcal{H}_{\lambda_0,\delta^*}, \beta) \leq C(\alpha,\beta,\delta^*)/\sqrt{L}$ *et par conséquent* $\mathrm{mSR}(\mathcal{H}_{\lambda_0,\delta^*}, \alpha, \beta) \leq C(\alpha,\beta,\delta^*)/\sqrt{L}$.

2.2 Adaptation à la position et la taille de saut simultanément

Considérant le problème de test de (H_0):"$\lambda \in \mathcal{H}_0$" contre (H_1):"$\lambda \in \mathcal{H}_{\lambda_0}(R)$" $(R > 0)$, on obtient pour la vitesse de séparation minimax sur $\mathcal{H}_{\lambda_0}(R)$ la borne inférieure suivante.

Proposition 3 (Borne inférieure)**.** *Soit* $(\alpha,\beta) \in (0,1)^2$ *vérifiant* $\alpha+\beta < 1-(\log(2)/3)^{1/2}$ *et* $R > 0$. *Il existe* $L(\alpha,\beta,\lambda_0,R) > 0$ *telle que pour tout* $L \geq L(\alpha,\beta,\lambda_0,R)$, $\mathrm{mSR}(\mathcal{H}_{\lambda_0}(R),\alpha,\beta) \geq \sqrt{\lambda_0 \log\log L/L}$.

Pour $L \geq 3$, soit $\Theta_L = \left\{1 - 2^{-k};\ k \in \{1,...,2\lfloor \log_2 L \rfloor\}\right\}$. Pour tout θ de Θ_L, on considère V_θ le sous-espace vectoriel de $\mathbb{L}^2([0,1])$ engendré par $\varphi_\theta = (1-\theta)^{-1/2}\mathbb{1}_{(\theta,1]}$. Notant Π_{V_θ} la projection orthogonale sur V_θ dans $\mathbb{L}^2([0,1])$, un estimateur sans biais de $\|\Pi_{V_\theta}(\lambda-\lambda_0)\|_2^2$ est donné par $T_{\lambda_0,\theta}(N) = \frac{1}{L^2(1-\theta)}\left(N((\theta,1])^2 - N((\theta,1])\right) - \frac{2\lambda_0}{L}N((\theta,1]) + \lambda_0^2(1-\theta)$.
Si $t_{\lambda_0,\theta,u}$ désigne le u-quantile de la loi de $T_{\lambda_0,\theta}(N)$ sous (H_0), on considère le test :

$$\phi_\alpha^{(2)} = \mathbb{1}_{\max_{\theta\in\Theta_L}\left(T_{\lambda_0,\theta}(N) - t_{\lambda_0,\theta,1-\alpha/(\,2\lfloor\log_2(L)\rfloor\,)}\right)>0}.$$

Theorem 4 (Borne supérieure)**.** *Soit* $(\alpha,\beta) \in (0,1)^2$. *Il existe* $L_0(\alpha,\beta,\lambda_0,R) > 0$ *et* $C(\alpha,\beta,\lambda_0,R) > 0$ *telles que pour tout* $L \geq L_0(\alpha,\beta,\lambda_0,R)$, $\mathrm{SR}(\phi_\alpha^{(2)}, \mathcal{H}_{\lambda_0}(R), \beta) \leq C(\alpha,\beta,\lambda_0,R)\sqrt{\log\log L/L}$ *et* $\mathrm{mSR}(\mathcal{H}_{\lambda_0}(R),\alpha,\beta) \leq C(\alpha,\beta,\lambda_0,R)\sqrt{\log\log L/L}$.

Cette borne supérieure, basée sur une nouvelle inégalité de concentration pour les suprema de martingales carrées (c.f. Le Guével (2020)), avec la borne inférieure obtenue dans la Proposition 3, mettent en évidence une perte inévitable de l'ordre d'un facteur $(\log\log L)^{1/2}$ dans la vitesse de séparation minimax sur $\mathcal{H}_{\lambda_0}(R)$. Une telle perte apparaît fréquemment dans la littérature traitant de problèmes d'adaptation au sens du minimax (c.f. Spokoiny (1996) ou Baraud (2002) dans des cadres gaussiens, Fromont et al. (2011) dans un cadre poissonien inhomogène). Ce résultat vient donc confirmer l'intuition que nous pouvions avoir à la lecture de ces références. De plus, il est semblable à la borne inférieure obtenue (comme conséquence de résultats plus généraux) par Gao, Han and Zhang (2019) pour le problème de détection de rupture dans l'espérance de variables i.i.d. de loi gaussienne.

3 Détection de rupture dans un processus de Poisson d'intensité de référence inconnue

Nous considérons à nouveau le problème de test de l'hypothèse nulle selon laquelle le processus de Poisson N est homogène contre l'hypothèse alternative selon laquelle son intensité présente un saut de taille δ en τ, en distinguant différents cas selon la connaissance ou non de δ et/ou τ, mais en ne supposant plus l'intensité du processus connue sous (H_0). $\mathcal{H}_0 = \mathcal{H}_0(R)$ désigne l'ensemble des fonctions constantes positives sur $[0,1]$ et bornées par $R > 0$, et on introduit pour $\delta^* \in (-\infty, R) \setminus \{0\}$ et $\tau^* \in (0,1)$ les classes :

$$\begin{aligned}
\mathcal{H}_{\delta^*,\tau^*}(R) &= \{\lambda : \exists \lambda_0 \in ((-\delta^*) \vee 0, (R-\delta^*) \wedge R), \lambda(t) = \lambda_0 + \delta^* \mathbb{1}_{(\tau^*,1]}(t)\},\\
\mathcal{H}_{\tau^*}(R) &= \{\lambda : \exists \lambda_0 \in (0,R), \delta \in (-\lambda_0, R-\lambda_0) \setminus \{0\}, \lambda(t) = \lambda_0 + \delta \mathbb{1}_{(\tau^*,1]}(t)\},\\
\mathcal{H}_{\delta^*}(R) &= \{\lambda : \exists \lambda_0 \in ((-\delta^*) \vee 0, (R-\delta^*) \wedge R), \tau \in (0,1), \lambda(t) = \lambda_0 + \delta^* \mathbb{1}_{(\tau,1]}(t)\},\\
\mathcal{H}(R) &= \{\lambda : \exists \lambda_0 \in (0,R), \tau \in (0,1), \delta \in (-\lambda_0, R-\lambda_0) \setminus \{0\}, \lambda(t) = \lambda_0 + \delta \mathbb{1}_{(\tau,1]}(t)\}.
\end{aligned}$$

Comme ci-dessus, pour les problèmes de test de (H_0):"$\lambda \in \mathcal{H}_0(R)$" contre (H_1):"$\lambda \in \mathcal{H}_{\delta^*,\tau^*}(R)$" ou (H_1):"$\lambda \in \mathcal{H}_{\tau^*}(R)$", on peut construire des tests uniformément plus puissants ou uniformément plus puissants parmi les tests sans biais, basés sur la statistique $N((\tau^*,1])$, et atteignant la vitesse de séparation minimax, d'ordre $L^{-1/2}$ dans le deuxième problème. La loi de $N((\tau^*,1])$ n'étant pas libre de l'intensité du processus sous (H_0), l'utilisation de quantiles conditionnels, inspirés de Fromont et al. (2011), vient éviter le recours à des méthodes de rééchantillonnage de type bootstrap ou permutation pour pallier ce problème. Nous montrons pour les classes d'alternatives $\mathcal{H}_{\delta^*}(R)$ et $\mathcal{H}(R)$ des résultats similaires à ceux obtenus dans les propositions 1 et 3, à savoir une borne inférieure pour la vitesse de séparation minimax sur $\mathcal{H}_{\delta^*}(R)$ d'ordre $L^{-1/2}$ et une borne inférieure pour la vitesse de séparation minimax sur $\mathcal{H}(R)$ d'ordre $(\log\log L/L)^{1/2}$. Pour ces problèmes, nous construisons des tests basés comme précédemment sur des approches d'agrégation, mais combinées ici à une astuce de conditionnement pour le choix des quantiles.

3.1 Adaptation à la position de saut

Considérant le problème de test de (H_0):"$\lambda \in \mathcal{H}_0(R)$" contre (H_1):"$\lambda \in \mathcal{H}_{\delta^*}(R)$" pour $R > 0$ et $\delta^* \in (-\infty, R) \setminus \{0\}$ (connu), nous définissons :

$$\phi_\alpha^{(3)} = \mathbb{1}_{\left\{S_{\delta^*}^{(1)}(N) > s_{1-\alpha/2}^{(1)}(N_1)\right\}} \vee \mathbb{1}_{\left\{S_{\delta^*}^{(2)}(N) > s_{1-\alpha/2}^{(2)}(N_1)\right\}},$$

où $S_{\delta^*}^{(1)}(N) = \sup_{t\in[0,1]} \left(\mathrm{sgn}(1/2-t)(N((0,t]) - tN((0,1])) - |\delta^*|Lt(1-t)/2\right)$, $S_{\delta^*}^{(2)}(N) = \sup_{t\in[0,1]} \left(\mathrm{sgn}(t-1/2)(N((0,t]) - tN((0,1])) - |\delta^*|Lt(1-t)/2\right)$ et $s_{1-\alpha/2}^{(1)}(n)$ (respectivement $s_{1-\alpha/2}^{(2)}(n)$) est le $(1-\alpha/2)$-quantile conditionnel de $S_{\delta^*}^{(1)}(N)$ (respectivement de $S_{\delta^*}^{(2)}(N)$) sachant $N_1 = n$ pour tout $n \in \mathbb{N}$ sous (H_0). Nous montrons que ce test atteint, à une constante multiplicative près, la vitesse de séparation minimax sur la classe d'alternatives $\mathcal{H}_{\delta^*}(R)$ considérée, qui est d'ordre $L^{-1/2}$.

3.2 Adaptation à la position et la taille de saut simultanément

Dans le cas du problème de test général de (H_0):"$\lambda \in \mathcal{H}_0(R)$" contre (H_1):"$\lambda \in \mathcal{H}(R)$", nous considérons, pour $L \geq 4$, $\Theta_L = \{2^{-k};\ k \in \{2, ..., \lceil \log_2(L) \rceil\}\} \cup \{1 - 2^{-k};\ k \in \{1, ..., \lfloor \log_2(L) \rfloor\}\}$, et pour $\theta \in \Theta_L$, le sous-espace vectoriel W_θ de $\mathbb{L}^2([0,1])$, engendré par $\psi_0 = \mathbb{1}_{[0,1]}$, et $\psi_\theta = -\sqrt{1-\theta/\theta}\mathbb{1}_{(0,\theta]} + \sqrt{\theta/(1-\theta)}\mathbb{1}_{(\theta,1]}$. Un estimateur sans biais de $\|\Pi_{W_\theta}(\lambda) - \Pi_{\psi_0}(\lambda)\|_2^2$ (Π_{W_θ} et Π_{ψ_0} désignant respectivement les projections orthogonales dans $\mathbb{L}^2([0,1])$ sur W_θ et $\mathrm{Vect}(\psi_0)$) est donné par la statistique $T_\theta(N) = \frac{\theta}{1-\theta}\frac{1}{L^2}\left(N((\theta,1])^2 - N((\theta,1])\right) + \frac{1-\theta}{\theta}\frac{1}{L^2}\left(N((0,\theta])^2 - N((0,\theta])\right) - \frac{2}{L^2}N((0,\theta])N((\theta,1])$.
Pour tout $n \in \mathbb{N}$, on note $t_{\theta,u}(n)$ le u-quantile de la loi de $T_\theta(N)$ sachant $N_1 = n$ sous (H_0). Alors le test défini par

$$\phi_\alpha^{(4)} = \mathbb{1}_{\left\{\max_{\theta \in \Theta_L}\left(T_\theta(N) - t_{\theta,1-\alpha/(2\lfloor \log_2(L) \rfloor - 1)}(N_1)\right) > 0\right\}},$$

atteint la vitesse de séparation minimax sur $\mathcal{H}(R)$, d'ordre $(\log\log L/L)^{1/2}$. Notons que ce test, lié aux tests multiples de type Bonferroni, peut être rendu moins conservatif en remplaçant le niveau ajusté $\alpha/(2\lfloor \log_2(L) - 1)$ par un niveau dépendant de N, inspiré des tests multiples de type *minp* (c.f. Fromont et al. (2011), Fromont et al. (2016)).

Note : Ce travail a été réalisé avec le soutien de la D.G.A. et de la région Bretagne.

Bibliographie

Baldwin, A., Gheyas, I., Ioannidis, C., Pym, D., Williams, J. (2017). Contagion in cybersecurity attacks, *J. Oper. Res. Soc.*, 68(780).
Baraud, Y. (2002). Non-asymptotic minimax rates of testing in signal detection, *Bernoulli*, 8(5), 577-606.
Fromont, M., Laurent, B., Reynaud-Bouret, P. (2011), *Ann. I.H.P. P&S*, 47(1), 176-213.
Fromont, M., Lerasle, M., Reynaud-Bouret, P. (2016), Family-Wise Separation Rates for multiple testing, *Ann. Statist.*, 44(6), 2533-2563.
Gao, C., Han, F., Zhang, C.-H. (2019). On Estimation of Isotonic Piecewise Constant Signals, *arXiv:1705.06386*.
Ingster, Yu. I. (1993). Asymptotically minimax testing for nonparametric alternatives I-II-III. *Math. Meth. Stat.*, 2, 85–114, 171–189, 249–268.
Le Guével, R. (2020). Exponential inequalities for the supremum of some counting processes and their square martingales, *https://hal.archives-ouvertes.fr/hal-02275583*.
Peng, C., Xu, M., Xu, S., and Hu, T. (2016). Modeling and predicting extreme cyber attack rates via marked point processes, *Journal of Applied Statistics*, 44(14).
Pyke, R. (1959). The supremum and infimum of the Poisson process, *A. Math. Stat.*, 30.
Spokoiny, V. G. (1996). Adaptive hypothesis testing using wavelets, *Ann. Statist.*, 24(6).
Takács, L. (1965). On the distribution of the supremum for stochastic processes with interchangeable increments, *Trans. Amer. Math. Soc.*, 119, 367–379.

Explaining the Explainer: A First Theoretical Analysis of LIME

Damien Garreau [1] & Ulrike von Luxburg [2]

[1] *Université Côte d'Azur, Inria, CNRS, LJAD*
Parc Valrose, 06108 Nice CEDEX 2, France
damien.garreau@unice.fr
[2] *University of Tübingen, Department of Computer Science*
Maria von Linden Straße 6, 72076 Tübingen, Germany
ulrike.luxburg@uni-tuebingen.de

Résumé. Les algorithmes d'apprentissage automatique sont de plus en plus souvent utilisés dans des situations pratiques, remplaçant parfois l'humain dans des processus de décision complexes. Ainsi, l'interprétabilité de ces algorithmes est un besoin pressant. Une méthode populaire pour obtenir l'explicabilité d'un algorithme est LIME (Local Interpretable Model-Agnostic Explanation). Dans ce travail, nous proposons la première analyse théorique de LIME. En particulier, nous obtenons l'expression exacte des coefficients du modèle interprétable lorsque la fonction à expliquer est linéaire. Nous montrons que dans ce cas les explications fournies par LIME sont fondées. Cependant, notre analyse révèle également que certains choix de largeur de bande peuvent mener à l'oubli de paramètres importants.

Mots-clés. apprentissage automatique, intelligence artificielle explicable.

Abstract. Machine learning is used more and more often for sensitive applications, sometimes replacing humans in critical decision-making processes. As such, interpretability of these algorithms is a pressing need. One popular algorithm to provide interpretability is LIME (Local Interpretable Model-Agnostic Explanation). In this paper, we provide the first theoretical analysis of LIME. We derive closed-form expressions for the coefficients of the interpretable model when the function to explain is linear. We show that in this case LIME indeed discovers meaningful features. However, our analysis also reveals that poor choices of parameters can lead LIME to miss important features.

Keywords. machine learning, explainable AI.

1 Introduction

The recent advance of machine learning methods is partly due to the widespread use of very complicated models, for instance deep neural networks. As an example, the Inception Network (Szegedy et al., 2015) depends on approximately 23 million parameters. While

these models achieve and sometimes surpass human-level performance on certain tasks (image classification being one of the most famous), they are often perceived as *black boxes*, with little understanding of how they make individual predictions.

This lack of understanding is a problem for several reasons. First, it can be a source of catastrophic errors when these models are deployed *in the wild*. For instance, for any safety system recognizing cars in images, we want to be absolutely certain that the algorithm is using features related to cars, and not exploiting some artifacts of the images. Second, this opacity prevents these models from being *socially accepted*. It is important to get a basic understanding of the decision making process to accept it.

Model-agnostic explanation techniques aim to solve this interpretability problem by providing qualitative or quantitative help to understand how black-box algorithms make decisions. Since the global complexity of the black-box models is hard to understand, they often rely on a *local* point of view, and produce an interpretation for a specific instance. In this article, we focus on such an interpretability technique: **Local Interpretable Model-Agnostic Explanations** (LIME, Ribeiro et al. (2016)).

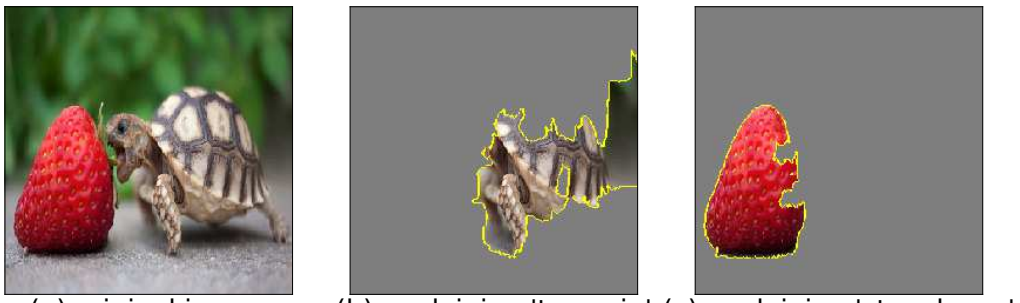


Figure 1: LIME in the context of object identification in images. We used Inception (Szegedy et al., 2015) as a black-box model. Terrapin, a sort of turtle, is the top label predicted for the image in panel (a). Panel (b) shows the results of LIME, explaining how this prediction was made. The highlighted parts of the image are the superpixels with the top 5 coefficients in the surrogate linear model. We ran the same experiment for the 'strawberry' label in panel (c).

2 LIME: Outline and notation

From now on, we will consider a particular model encoded as a function $f : \mathbb{R}^d \to \mathbb{R}$ and a particular instance $\xi \in \mathbb{R}^d$ to explain. We make no assumptions on this function, *e.g.*, how it might have been learned. We simply consider f as a black-box model giving us predictions for all points of the input space. Our goal will be to explain the decision $f(\xi)$ that this model makes for one particular instance ξ.

As soon as f is too complicated, it is hopeless to try and fit an interpretable model globally, since the interpretable model will be too simple to capture all the complexity of f. Thus a reasonable course of action is to consider a *local* point of view, and to explain f in the neighborhood of some fixed instance ξ. This is the main idea behind LIME: To explain a decision for some fixed input ξ, sample other examples around ξ, use these samples to build a simple interpretable model in the neighborhood of ξ, and use this surrogate model to explain the decision for ξ.

One additional idea that makes a huge difference with other existing methods is to use *discretized* features of smaller dimension d' to build the local model. These new categorical features are easier to interpret, since they are categorical (one can think of a single letter for the size, S-M-L, instead of a number). In the case of images, they are built by using a split of the image ξ into superpixels (homogeneous patches of the image, Ren and Malik, 2003). See Figure 1 for an example of LIME output in the case of image classification. In this situation, the surrogate model highlights the superpixels of the image that are the most "active" in predicting a given label.

Whereas LIME is most famous for its results on images, it is easier to understand how it operates and to analyze theoretically on **tabular data** (that is, multivariate data without specific structure). In this setting, LIME works essentially in the same way, with a main difference: `TabularLIME` requires a train set, and each feature is discretized according to the empirical quantiles of this training set. Note that the discretization process in this case does not group features together. A succinct description of the algorithm is given in Algorithm 1.

Algorithm 1 `TabularLIME` for regression

Require: Model f, # of new samples n, instance ξ, bandwidth ν, # of bins p, mean μ, variance σ^2

$q \leftarrow$ `GetQuantiles`(p,μ,σ)
$t \leftarrow$ `Discretize`(ξ,q)
for $i = 1$ to n **do**
  for $j = 1$ to d **do**
    $y_{i,j} \leftarrow$ `SampleUniform`$(\{1,\dots,p\})$
    $(q_\ell, q_u) \leftarrow (q_{j,y_{ij}}, q_{j,y_{ij}+1})$
    $x_{i,j} \leftarrow$ `SampleTruncGaussian`$(q_\ell, q_u, \mu, \sigma)$
    $z_{i,j} \leftarrow \mathbf{1}_{t_j = y_{i,j}}$
  end for
  $\pi_i \leftarrow \exp\left(\frac{-\|x_i - \xi\|^2}{2\nu^2}\right)$
end for
$\widehat{\beta} \leftarrow$`WeightedLeastSquares`$(z, f(x), \pi)$
return $\widehat{\beta}$

Let us note that `TabularLIME` can be used both for regression and classification. Here we focus on the *regression* mode: the outputs of the model are real numbers, and not discrete elements. In some sense, this is a more general setting than the classification case, since the classification mode operates as `TabularLIME` for regression, but with f chosen as the function that gives the likelihood of belonging to a certain class according to the model.

Given a class of simple, interpretable models G, `TabularLIME` selects the best of these

models by solving

$$\arg\min_{g\in G}\left\{L_n(f,g,\pi_\xi)+\Omega\left(g\right)\right\}, \tag{2.1}$$

where L_n is a local loss function evaluated on the new examples $x_1,\ldots,x_n$, and $\Omega:\mathbb{R}^d\to\mathbb{R}$ is a regularizer function. For instance, a natural choice for the local loss function is the weighted squared loss

$$L_n(f,g,\pi):=\frac{1}{n}\sum_{i=1}^{n}\pi_i\left(f(x_i)-g(z_i)\right)^2\,. \tag{2.2}$$

In this work, we will focus exclusively on the linear models, in our opinion the easiest models to interpret. Namely, we set $g(z_i)=\beta^\top z_i+\beta_0$, with $\beta\in\mathbb{R}^d$ and $\beta_0\in\mathbb{R}$. To get rid of the intercept β_0, we now use the standard approach to introduce a phantom coordinate 0, and $z,\beta\in\mathbb{R}^{d+1}$ with $z_0=1$. We also stack the z_is together to obtain $Z\in\{0,1\}^{n\times(d+1)}$.

The regularization term $\Omega(g)$ is added to insure further interpretability of the model by reducing the number of non-zero coefficients in the linear model given by `TabularLIME`. Typically, one uses L^2 regularization (ridge regression is the default setting of LIME) or L^1 regularization (the Lasso). To simplify the analysis, we will set $\Omega=0$ in the following. We believe that many of our results stay true in a regularized setting, especially the switch-off phenomenon that we are going to describe below: coefficients are even more likely to be set to zero when $\Omega\neq 0$.

In other words, in our case `TabularLIME` performs *weighted linear regression* on the interpretable features z_is, and outputs a vector $\widehat{\beta}\in\mathbb{R}^{d+1}$ such that

$$\widehat{\beta}\in\arg\min_{\beta\in\mathbb{R}^{d+1}}\left\{\frac{1}{n}\sum_{i=1}^{n}\pi_i(y_i-\beta^\top z_i)^2\right\}. \tag{2.3}$$

3 Main results

We are now ready to state our main result. Let us denote by $\widehat{\beta}$ the coefficients of the linear surrogate model obtained by `TabularLIME`. In a nutshell, when the underlying model f is linear, we can derive the average value β of the $\widehat{\beta}$ coefficients. The exact form of the proportionality coefficients is given in the formal statement below, it essentially depends on the scaling parameters

$$\tilde{\mu}:=\frac{\nu^2\mu+\sigma^2\xi}{\nu^2+\sigma^2}\in\mathbb{R}^d \text{ and } \tilde{\sigma}:=\frac{\nu^2\sigma^2}{\nu^2+\sigma^2}>0\,,$$

and the $q_{j\pm}$s, the quantiles left and right of the ξ_js.

Theorem 3.1 (Coefficients of the surrogate model, theoretical values). *Assume that f is of the form $x \mapsto a^\top x + b$, and set*

$$\beta_j := \begin{cases} f(\tilde{\mu}) + \sum_{j=1}^{d} \frac{a_j \theta_j}{1-\alpha_j} & \text{if } j = 0 \\ \frac{-a_j \theta_j}{\alpha_j(1-\alpha_j)} & \text{otherwise,} \end{cases} \tag{3.1}$$

where, for any $1 \leqslant j \leqslant d$, we defined

$$\alpha_j := \left[\frac{1}{2}\mathrm{erf}\left(\frac{x-\tilde{\mu}_j}{\tilde{\sigma}\sqrt{2}}\right)\right]_{q_j-}^{q_j+} \quad \text{and} \quad \theta_j := \left[\frac{\tilde{\sigma}}{\sqrt{2\pi}}\exp\left(\frac{-(x-\tilde{\mu}_j)^2}{2\tilde{\sigma}^2}\right)\right]_{q_j-}^{q_j+}.$$

Let $\eta \in (0,1)$. Then, with high probability greater than $1-\eta$, it holds that $\left\|\widehat{\beta} - \beta\right\| \lesssim \sqrt{\frac{\log 1/\eta}{n}}$, with constants depending on f, μ, and σ.

Figure 2 shows how our theoretical predictions match empirical results. We now discuss the consequences of Theorem 3.1.

Dependency in the partial derivatives. A first consequence of Theorem 3.1 is that, in the linear case, the coefficients of the linear model given by `TabularLIME` are approximately **proportional to the partial derivatives of** f at ξ, with constant depending on our assumptions. An interesting follow-up is that, if f depends only on a few features, then the partial derivatives in the other coordinates are zero, and the coefficients given by `TabularLIME` for these coordinates will be 0 as well. In a simple setting, we thus showed that `TabularLIME` does not produce interpretations with additional erroneous feature dependencies. Indeed, when the number of samples is high, the coordinates which do not influence the prediction will have a coefficient close to the theoretical value 0 in the surrogate linear model. We believe that this forgetting of irrelevant features holds for more general functions.

Influence of the bandwidth. Unfortunately, Theorem 3.1 does not provide directly a founded way to pick ν, which would for instance minimize the variance for a given level of noise. The quest for a founded heuristic is still open. However, we gain some interesting insights on the role of ν. Namely, for fixed ξ, μ, and σ, the multiplicative constants $\theta_j/(\alpha_j(1-\alpha_j))$ appearing in Eq. (3.1) depend essentially on ν.

It is possible to artificially (or by accident) put θ_j to zero, therefore **forgetting** about feature j in the explanation, whereas it could play an important role in the prediction. An interesting take is that ν not only decides at which scale the explanation is made, but also the magnitude of the coefficients in the interpretable model, even for small changes of ν.

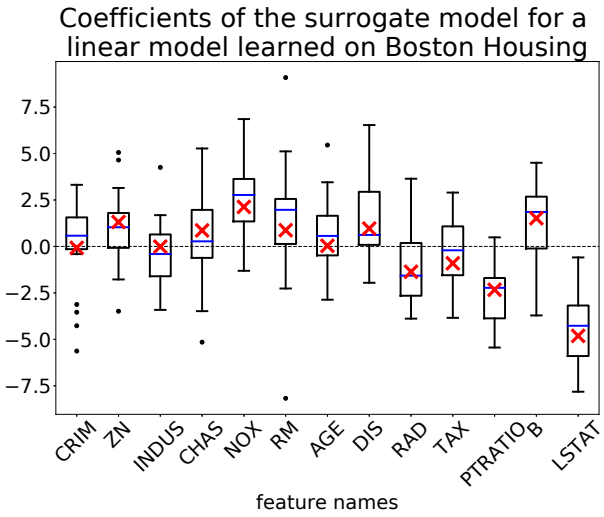


Figure 2: Values of the coefficients obtained by `TabularLIME` on each coordinate in dimension $d = 13$ for a linear model trained on the Boston housing dataset (Harrison Jr. and Rubinfeld, 1978). The β_js are concentrated around the red crosses, which denote the β_js, the theoretical values predicted by Theorem 3.1. To produce the figure, we ran 20 experiments with $n = 10^3$ new samples generated for each run and we set $\nu = 1$.

4 Conclusion

In this paper we provide the first theoretical analysis of LIME, with some good news (LIME discovers interesting features) and bad news (LIME might forget some important features). We invite the interested reader to read Garreau and von Luxburg (2020), a much more complete version of this work accepted to AISTATS 2020.

Bibliography

D. Garreau and U. von Luxburg. Explaining the Explainer: A first Theoretical Analysis of LIME. *AISTATS*, 2020.

D. Harrison Jr. and D. L. Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of environmental economics and management*, 5(1):81–102, 1978.

X. Ren and J. Malik. Learning a classification model for segmentation. In *ICCV*, 2003.

M. T. Ribeiro, S. Singh, and C. Guestrin. Why should I trust you? explaining the predictions of any classifier. In *SIGKDD*, 2016.

C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

Estimation de fonction de répartition conditionnelle pour l'analyse de données RNA-seq en cellule unique

Marine Gauthier [1,3], Rodolphe Thiébaut [1,2,3], Véronique Godot [3,4] & Boris P. Hejblum [1,3]

[1] *Université de Bordeaux, INSERM Bordeaux Population Health Research Center, INRIA SISTM, 33000 Bordeaux* [2] *CHU Bordeaux, 33000, Bordeaux* [3] *Vaccine Research Institute (VRI), Créteil, 94000, France* [4] *Inserm, U955, Team 16, Université Paris-Est, Créteil, Faculté de médecine*

marine.gauthier@u-bordeaux.fr, rodolphe.thiebaut@u-bordeaux.fr, veronique.godot@gmail.com, boris.hejblum@u-bordeaux.fr

Résumé. Contrairement à l'analyse de données RNA-sequencing en masse (*bulk RNA-seq*) mesurant l'expression génique moyenne dans une population cellulaire, l'analyse de données RNA-seq par cellule unique (*scRNA-seq*) permet à présent d'étudier des phénomènes biologiques à l'échelle uni-cellulaire. L'analyse de l'expression différentielle consiste à tester l'association de l'expression d'un gène avec un ou plusieurs facteurs. Alors que la complexité grandissante des études appelle à une plus grande polyvalence des approches, la majorité des méthodes existantes se limite à une comparaison entre deux conditions uniquement. Ces approchent s'appuient sur des hypothèses distributionnelles fortes qu'il est difficile de vérifier en pratique. De plus, les données RNA-seq en cellule unique sont caractérisées par une proportion de zéros très importante, ce qui complexifie leur modélisation. Nous proposons donc une nouvelle approche permettant de tester l'association de l'expression d'un gène, quelque soit sa distribution, avec une ou plusieurs variables explicatives d'intérêt (continues ou discrètes), potentiellement ajustées sur des covariables additionnelles. Nous estimons la fonction de répartition de l'expression d'un gène dans une cellule par une méthode de régression et comparons l'estimation conditionnelle aux variables d'intérêt à l'estimation marginale, à l'aide d'une distance. Entre autres, le choix de la norme 1, la norme 2 ou la norme infinie sera discuté. Grâce à un test par permutations, nous pouvons ainsi détecter les gènes différentiellement exprimés selon les facteurs d'intérêt. Notre méthode se veut particulièrement flexible de par son adaptabilité à tout design expérimental qui peut s'exprimer dans l'équation de régression. Nous présenterons une application sur un jeu de données réelles étudiant des sous-populations de cellules dendritiques, qui a motivé ce développement méthodologique, ainsi que les premiers résultats de nos simulations numériques.

Mots-clés. RNA-seq par cellule unique, expression génique, fonction de répartition conditionnelle, test par permutations, cellules dendritiques

Abstract. Unlike bulk RNA-seq data analysis (using the average gene expression in a cell population), the analysis of single-cell RNA-seq (scRNA-seq) data makes it possible to study biological mechanisms at the cellular level. Differential Expression Analysis

(DEA) consists in testing the association of a gene expression with one or more factors. State-of-the-art methods for scRNA-seq DEA face methodological issues, as they rely on strong distributional assumptions that are difficult to test in practice, questioning the validity of their results given the high rate of zeros in scRNA-seq data. While the increasing complexity of clinical and biological studies calls for greater tools versatility, the majority of existing methods only tackles the comparison between two conditions. We propose a new distribution-free approach to test the association of gene expression to one or several variables of interest (continuous or discrete) potentially adjusted to additional covariates. We estimate the cumulative distribution function of gene expression through regression method and compare the conditional estimation to the marginal estimation using a distance (the choice of L1 norm, L2 norm or the infinite norm will be discussed). Through a permutation test, we can thus detect genes that are differentially expressed according to the factors of interest. Our method is particularly flexible because of its adaptability to any experimental design that can be expressed in the regression equation. We will present an application on a real data set studying subpopulations of blood dendritic cells, which motivated this methodological development, as well as the first results of our numerical simulations.

Keywords. single-cell RNA-seq, gene expression, conditional cumulative function, permutation test, dendritic cells

1 Limites des méthodes scRNA-seq existantes

La technologie single-cell RNA-seq (*scRNA-seq*) rend possible la mesure simultanée de l'expression génique à l'échelle unicellulaire, dans des milliers de cellules (contrairement à la technologie RNA-seq en masse) permettant ainsi d'étudier de nouvelles questions biologiques comme l'identification des différents types de cellules ou l'hétérogénéité des réponses cellulaires. Notre objectif est de réaliser une analyse différentielle de l'expression génique, c'est-à-dire trouver quels gènes sont différentiellement exprimés en fonction de certaines variables d'intérêt.

L'expression d'un gène étant mesurée dans des centaines voire milliers de cellules, on a accès à la répartition des différentes expressions de ce gène dans celles-ci. D'une part, la particularité des données single-cell est leur extrême parcimonie : l'expression des gènes n'est pas mesurée dans la plupart des cellules conduisant à l'imputation de zéros pour de nombreux gènes. D'autre part, certains bruits techniques ainsi que les différences biologiques des cellules présentes dans un même échantillon peuvent génèrer des distributions multimodales et hétérogènes. Par conséquent, caractériser la distribution de l'expression génique par cellule de manière paramétrique reste alors délicat; des modèles faisant appel à une Binomiale Négative enflée en zéro (ZINB), à une distribution de Poisson enflée en zéro (ZIP) ou bien à des modèles de mélange de loi Gamma et de loi

Normale [1, 2, 3, 4] ont été proposés. Or, la variété des distributions possibles pour chacun des gènes rend ces modèles difficilement généralisables. Dans cette direction, Wang et Nabavi ont proposé `SigEMD` [5], une méthode non-paramétrique basée sur la distance de Wassertein mais ne donnant la possibilité de comparer que deux densités, et Delmans et Hemberg ont présenté `D3E` [6], une autre méthode non-paramétrique qui propose d'utiliser soit le test de Cramer-Von Mises soit Kolmogorov-Smirnov afin de comparer la distribution d'un gène entre deux conditions.

2 Motivation

Un nombre important d'expériences biologiques générant des données scRNA-seq reste dans le cadre classique de l'analyse différentielle et se limitent à l'étude de 2 conditions (par exemple, vaccin/placebo). Il est pourtant envisageable de vouloir tester l'association de l'expression d'un gène avec une variable discrète à plus de 2 classes (sous-populations cellulaires ou différentes doses vaccinales), ou bien tester l'association avec une variable continue, comme l'expression d'un autre gène ou des marqueurs biologiques. Le jeu de données réelles qui a motivé notre travail est composé de 18 513 gènes pour 2 914 cellules dendritiques (DC). Auparavant, 4 sous-populations ont été annotées: DC1, DC2 & DC3, pDC et preDC. On cherche à identifier quels gènes s'expriment différentiellement dans les 4 sous-populations cellulaires et à tester l'association entre profils transcriptomiques. Les méthodes actuelles ne permettant pas de prendre en compte ce schéma d'étude, nous proposons une nouvelle méthode d'analyse différentielle pour données scRNA-seq. Cette dernière se passe d'hypothèse distributionnelle et peut tester l'association de l'expression génique avec une ou plusieurs variables d'intérêt, qu'elles soient continues et/ou discrètes, en ajustant sur de potentielles covariables. Notre méthode se voudra être la plus flexible possible pour s'adapter à n'importe quel schéma expérimental.

3 Méthode

3.1 Estimation de la fonction de distribution conditonnelle

Soit un n-échantillon (X_1, Y_1, Z_1), ...,(X_n, Y_n, Z_n) i.i.d de même loi que (X, Y, Z). On considère $Y \in \mathbb{R}^n$ l'expression d'un gène, $X \in \mathbb{R}^{n \times p}$ avec $p \geq 1$ les variables à tester (continues et/ou discrètes) et $Z \in \mathbb{R}^{n \times q}$ avec $q \geq 1$ les covariables (continues et/ou discrètes). Afin de caractériser la distribution de l'expression génique sachant une ou plusieurs variables de façon non-paramétrique, nous allons utiliser la fonction de répartition conditionnelle.

On définit la fonction de répartition conditionnelle de Y sachant X et Z par :

$$F_{Y|X,Z}(y \mid x, z) = \mathbb{P}(Y \leq y \mid X = x, Z = z)$$

On cherche à estimer F. Or, $F_{Y|X,Z}(y \mid x, z) = \mathbb{E}(\mathbb{1}_{\{Y \leq y\}} \mid X = x, Z = z)$. On peut donc estimer F par une méthode de régression. On propose d'utiliser une régression logistique. Le recours à un arbre CART ou à une forêt aléatoire dans le cas d'éventuelles relations non-linéaires représente une alternative potentielle. Par ailleurs, l'estimation de F se fait par un ensemble de régressions, une pour chaque observation Y_i.

3.2 Test d'hypothèse

3.2.1 Sans covariables Z

Nous cherchons à tester l'indépendance entre Y et X sans ajustement sur des covariables :

$$H_0 : \text{“}Y \perp\!\!\!\perp X\text{”}$$

ce qui est équivalent à comparer la fonction de répartition conditionnelle à la fonction de répartition marginale :

$$H_0 : \text{“}F_{Y|X}(y, x) = F_Y(y)\text{” contre } H_1 : \text{“}F_{Y|X}(y, x) \neq F_Y(y)\text{”}$$

$F_Y(y)$ est estimée par l'estimateur empirique $\widehat{F}_Y(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{y_i \leq y\}}$.
On compare donc la fonction de répartition conditionnelle de Y sachant X à la fonction de répartition marginale de Y. On définit alors la statistique de test suivante : $D = \textit{distance}\left(F_{Y|X}(y, x), F_Y(y)\right)$. On peut considérer la distance L^1, L^2 ou L^∞. Le choix de celle-ci sera discuté.

3.2.2 Avec covariables Z

Nous cherchons à tester l'indépendance conditionnelle entre Y et X sachant Z. On teste alors :

$$H_0 : \text{“}Y \perp\!\!\!\perp X \mid Z\text{”}$$

ce qui est équivalent à :

$$H_0 : \text{“}F_{Y|X,Z}(y, x, z) = F_{Y|Z}(y, z)\text{” contre } H_1 : \text{“}F_{Y|X,Z}(y, x, z) \neq F_{Y|Z}(y, z)\text{”}$$

On compare alors la fonction de répartition conditionnelle de Y sachant X et Z à la fonction de répartition conditionnelle de Y sachant Z pour constater l'influence de X.
On définit alors la statistique de test suivante : $D = \textit{distance}\left(F_{Y|X,Z}(y, x, z), F_{Y|Z}(y, z)\right)$.

3.3 Test par permutations

La distribution asymptotique de la statistique de test D n'étant pas déterminée, nous proposons d'utiliser des permutations afin d'estimer la distribution empirique de $\widehat{D}$ sous l'hypothèse nulle.

3.3.1 $\mathbf{Y} \perp\!\!\!\perp X$

Dans le premier cas, sous H_0, Y et X sont indépendantes, donc les observations de X sont échangeables. Nous effectuons alors B permutations des observations de X afin d'induire l'indépendance :

- $\forall i\ X_i^* = X_{\sigma(i)}$ with $\sigma \in Perm\{1, ..., n\}$
- On estime $F_{Y|X^*}(Y \mid X^*)$
- On calcule $D_b^* = distance(F_{Y|X^*}(Y \mid X^*), F_Y(Y)), \forall b = 1, ..., B$

On obtient donc B statistiques de test sous H_0 : $\mathcal{D} = \{D_1^*, ..., D_B^*\}$. On calcule la p-valeur associée comme suit : $\widehat{p} = \dfrac{1}{1+B}\left(1 + \displaystyle\sum_{b=1}^{B} \mathbb{1}\{\widehat{D} \leq D_b^*\}\right)$. Enfin, on applique la correction de Benjamini-Hochberg pour la multiplicité des tests.

3.3.2 $\mathbf{Y} \perp\!\!\!\perp X \mid Z$

Si nous devons ajuster sur des covariables Z, alors la procédure de permutation n'est pas directe. En effet, lorsque nous permutons les observations de X, nous cassons non seulement le lien entre X et Y comme désiré, mais également le lien entre X et Z. Il va donc s'agir de préserver la dépendance qui existe entre X et Z. Si Z est une variable catégorielle, il est possible de regrouper les observations de X en fonction de la valeur de Z associée, pour ensuite permuter X au sein des groupes. En revanche, cette stratégie ne peut être appliquée directement dans le cas où Z est continue. On discutera donc d'une stratégie de permutation conditionnelle basée sur la distance entre les prédictions des observations de X.

4 Simulation numérique

On simule un n-échantillon $((X_1, Y_1), ...(X_n, Y_n)) \in \{1, 2, 3, 4\} \times \mathbb{R}$ avec $n = 400$ tels que : $Y_i \sim \mathcal{N}(0, 1)$ si $X_i = 0$, $Y_i \sim \mathcal{N}(\beta, 1)$ si $X_i = 1$, $Y_i \sim \mathcal{N}(2\beta, 1)$ si $X_i = 2$ et $Y_i \sim \mathcal{N}(3\beta, 1)$ si $X_i = 3$. On teste si $Y \perp\!\!\!\perp X$ en comparant $F_{Y|X}(y, x)$ à $F_Y(y)$. Si $\beta = 0$, nous sommes donc sous l'hypothèse nulle. D'après la figure 1, pour $\beta = 0$, l'estimation de Monte-Carlo sur 1000 simulations de l'erreur de Type 1 est bien controlée à 5% (pour un seuil fixé arbitrairement à 5%) pour les 3 normes choisies. A mesure que β augmente, les 4 distributions sont de plus en plus distinctes et la proportion estimée d'hypothèses nulles rejetées est de plus en plus grande, illustrant ainsi la bonne puissance statistique de la méthode pour un schéma de simulation simple.

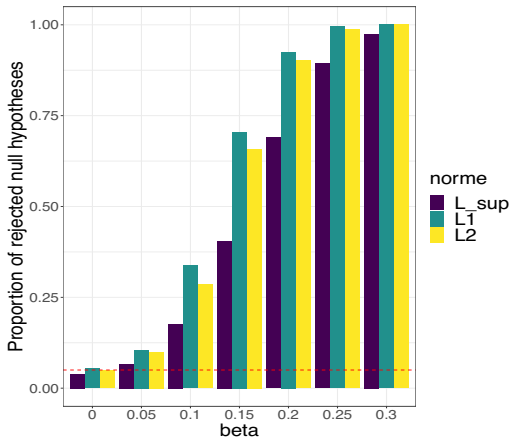


Figure 1: Estimation de Monte-Carlo sur 1000 simulations de la proportion d'hypothèses nulles rejetées en fonction de β pour la norme 1, la norme 2 et la norme infinie. La ligne rouge en pointillés représente le seuil arbitraire de 5%.

References

[1] Korthauer KD, Chu LF, Newton MA, Li Y, Thomson J, Stewart R, et al. A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. Genome biology. 2016;17(1):222.

[2] Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome biology. 2015;16(1):278.

[3] Kharchenko PV, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. Nature methods. 2014;11(7):740.

[4] Risso D, Perraudeau F, Gribkova S, Dudoit S, Vert JP. A general and flexible method for signal extraction from single-cell RNA-seq data. Nature communications. 2018;9(1):284.

[5] Wang T, Nabavi S. SigEMD: a powerful method for differential gene expression analysis in single-cell RNA sequencing data. Methods. 2018;145:25–32.

[6] Delmans M, Hemberg M. Discrete distributional differential expression (D 3 E)-a tool for gene expression analysis of single-cell RNA-seq data. BMC bioinformatics. 2016;17(1):110.

Extension de la régression linéaire généralisée sur composantes supervisées à la modélisation jointe des réponses

Julien GIBAUD[1], Xavier BRY[1] et Catherine TROTTIER[1,2]

[1] Institut Montpelliérain Alexander Grothendieck, CNRS, Univ. Montpellier, France.
[2] Univ. Paul-Valéry Montpellier 3, F34000, Montpellier, France.

Contact : julien.gibaud@umontpellier.fr, xavier.bry@umontpellier.fr et catherine.trottier@univ-montp3.fr.

Résumé

Dans ce travail, nous proposons d'étendre la méthode SCGLR, pour la rendre capable d'identifier des groupes de réponses expliquées par des composantes communes. À l'origine, SCGLR vise la construction de composantes explicatives dans un grand nombre de covariables, éventuellement fortement redondantes. Ces composantes sont supervisées conjointement par l'ensemble des réponses. Désormais, nous cherchons à identifier des groupes de réponses partageant les mêmes dimensions explicatives. Dans un cadre écologique par exemple, des communautés d'espèces devraient pouvoir être modélisées par des composantes propres à chaque communauté. Un algorithme est proposé afin d'estimer le modèle.

Mots clefs : SCGLR, mélange de réponses, algorithme EM, classification

Abstract

In this work, we propose to extend the SCGLR methodology, enabling it to identify clusters of responses sharing explanatory components. Originally, SCGLR was designed to find explanatory components in a large set of possibly highly redundant covariates, something much needed in a high-dimensional framework. These components are jointly supervised by all the responses. Henceforth, we aim at identifying clusters of responses sharing the same explanatory dimensions. In an ecological framework for instance, communities of species should be modeled by components which are characteristic of each community. An algorithm is proposed in order to estimate the model.

Keywords : SCGLR, response mixture, EM algorithm, clustering

1 Contexte

Les changements climatiques entraînent certains dérèglements des écosystèmes pouvant causer des extinctions d'espèces animales ou végétales. Dans ce contexte, le développement de modèles permettant de prédire le futur de la biodiversité est devenu un enjeu crucial. Récemment, de nombreuses avancées ont été faites dans ce domaine, en particulier par l'extension des modèles de distribution des espèces (Species Distribution Models, SDM), qui traitent les espèces séparément, à des modèles de distribution jointe (Joint Species Distribution Models, JSDM). Les JSDM permettent de formaliser l'interdépendance des espèces et de mieux comprendre son impact sur la composition des communautés. Par ailleurs, modéliser l'abondance des espèces requiert de prendre en compte un grand nombre de covariables explicatives souvent corrélées, ce qui impose une réduction de

dimension et la régularisation des modèles.

Dans leur article, Bry et *al.* [1] proposent une méthode - la régression linéaire généralisée sur composantes supervisées (Supervised Component-based Generalized Linear Regression, SCGLR) - combinant le modèle linéaire généralisé multivarié avec les méthodes à composantes permettant la réduction de dimension. SCGLR optimise un critère compromis entre la qualité d'ajustement (Goodness-of-Fit, GoF) et la proximité à des dimensions d'intérêt (Structural Relevance, SR) [2]. Cette technique ne trouve pas seulement des directions fortes et interprétables, elle produit aussi des prédicteurs régularisés, ce qui permet le traitement de données de grande dimension. Cependant, SCGLR suppose que l'ensemble des réponses dépend des mêmes dimensions explicatives. Pour nous affranchir de cette hypothèse, nous proposons d'étendre cette méthode aux mélanges sur les réponses. L'objectif est de trouver des classes de réponses (espèces) telles que toutes les réponses d'une classe soient modélisables par les mêmes dimensions explicatives.

2 Modélisation

Dans cette section, nous présentons la méthode SCGLR, puis son extension aux mélanges sur les réponses.

2.1 SCGLR

n individus sont décrits par K réponses y_k, $k = 1, \ldots, K$, ainsi que des covariables explicatives séparées en deux groupes : un groupe X de covariables *a priori* nombreuses et possiblement redondantes, et un autre A de covariables additionnelles peu nombreuses et faiblement, voire non-redondantes. On notera X et A les matrices correspondantes. Chaque réponse y_k fait l'objet d'un modèle linéaire généralisé (Generalized Linear Model, GLM) [6]. Pour la partie explicative du modèle, seule la matrice X requiert réduction de dimension et régularisation. À cette fin, SCGLR cherche dans X des composantes communes à l'ensemble des réponses. Une composante $f \in \mathbb{R}^n$ est donnée par $f = Xu$ où $u \in \mathbb{R}^p$ est un vecteur de coefficients. Le prédicteur linéaire associé à la réponse y_k est donné par

$$\eta_k = (Xu)\,\gamma_k + A\delta_k,$$

où γ_k et δ_k sont les paramètres de régression. La composante f est commune à l'ensemble des réponses y_k et pour assurer l'identifiabilité, nous imposons $u^T M^{-1} u = 1$, où $M \in \mathbb{R}^{p\times p}$ est une matrice symétrique définie positive. Nous supposons que les réponses sont indépendantes conditionnellement aux variables explicatives.

À cause du produit $u\gamma_k$, le modèle "linéarisé" à chaque étape de l'algorithme des scores de Fisher (Fisher Scoring Algorithm, FSA) pour l'estimation du GLM, n'est pas linéaire et doit être estimé de façon alternée sur u et sur $\{\gamma_k, \delta_k\}$. Soient w_k, la pseudo-réponse (ou variable de travail) associée à chaque étape du FSA, et W_k^{-1} sa matrice de variance-covariance. L'estimateur des moindres carrés de u est solution des programmes équivalents suivants :

$$\min_{u,u^T M^{-1}u=1} \sum_{k=1}^{K} \left\| w_k - \Pi^{W_k}_{\mathrm{vect}(f,A)} w_k \right\|^2_{W_k} \Leftrightarrow \max_{u,u^T M^{-1}u=1} \sum_{k=1}^{K} \left\| \Pi^{W_k}_{\mathrm{vect}(f,A)} w_k \right\|^2_{W_k} \Leftrightarrow \max_{u,u^T M^{-1}u=1} \psi_A(u),$$

avec $\psi_A(u) = \sum_{k=1}^{K} \|w_k\|^2_{W_k} \cos^2_{W_k}\left(w_k, \Pi^{W_k}_{\mathrm{vect}(f,A)} w_k\right)$. La quantité ψ_A est une mesure de GoF. Pour trouver des composantes fortes et interprétables, le GoF ne suffit pas. Il faut le combiner avec une

mesure de pertinence structurelle.

Soient un ensemble $N = \{N_1, \ldots, N_J\}$ de matrices symétriques semi-définies positives, un ensemble $\omega = \{\omega_1, \ldots, \omega_J\}$ de poids et un scalaire $l \geq 1$. La mesure de pertinence structurelle (SR) ϕ associée est donnée par

$$\phi(u) = \left(\sum_{j=1}^{J} \omega_j \left(u^T N_j u \right)^l \right)^{1/l}.$$

Les matrices N_j sont telles que les formes quadratiques $u^T N_j u$ mesurent la proximité du vecteur u à des structures de référence.

Dans ce travail nous utilisons une mesure particulière de SR : la variance de la composante. On appelle W la matrice des poids *a priori* des observations (typiquement, $W = \frac{1}{n} I_n$). On prend X centrée en colonne. Nous voulons trouver une direction vect(u) captant une inertie suffisante des observations. Pour cela, on pose : $N = \{X^T W X\}$, $\omega = \{1\}$ et $l = 1$. Ainsi la SR devient

$$\phi(u) = u^T X^T W X u = \|Xu\|_W^2 = \mathbb{V}(Xu).$$

Nous reconnaissons le critère maximisé, sous la contrainte $u^T M^{-1} u = 1$, par l'ACP de X avec la métrique M et la matrice des poids W. De façon générale, quelle que soit la SR choisie, la métrique M de la contrainte $u^T M^{-1} u = 1$ est choisie de la forme $M^{-1} = \tau I_n + (1 - \tau) X^T W X$, où $\tau \in [0, 1]$ est un paramètre de régularisation de type ridge [4].

Pour construire un compromis entre le GoF et la SR, SCGLR introduit un réel $s \in [0, 1]$ et considère le programme de maximisation suivant :

$$\max_{u, u^T M^{-1} u = 1} \phi(u)^s \psi_A(u)^{1-s} \Leftrightarrow \max_{u, u^T M^{-1} u = 1} s \ln(\phi(u)) + (1 - s) \ln(\psi_A(u)).$$

2.2 MixRep-SCGLR

Dans cette section, nous proposons d'étendre SCGLR aux modèles de mélange sur les réponses. Nous considérons désormais que l'ensemble des réponses n'est pas modélisé par les mêmes dimensions explicatives. Nous cherchons à classifier les réponses dans des groupes tels que toutes les réponses d'un même groupe soient expliquées par les mêmes dimensions explicatives. Nous étudions le cas où toutes les réponses sont gaussiennes et nous supposons ici qu'il n'existe qu'une direction explicative par groupe.

Soit $\mathbf{Y} = [y_1, \ldots, y_K]$, l'ensemble des réponses. On note G le nombre de classes *a priori* du modèle et p_g la probabilité d'appartenance de chaque réponse à la classe g. Si y_k appartient au groupe g, alors y_k suit une loi normale multivariée $N_n(\mu_{kg}, \Sigma_{kg})$ avec $\mu_{kg} = (Xu_g)\gamma_{kg} + A\delta_{kg}$ et $\Sigma_{kg} = \sigma_{kg}^2 I_n$. La densité de y_k est donc :

$$f_Y(y_k; \Theta_k) = \sum_{g=1}^{G} \frac{p_g}{(2\pi\sigma_{kg}^2)^{n/2}} \exp\left(-\frac{1}{2\sigma_{kg}^2} \|y_k - (Xu_g)\gamma_{kg} - A\delta_{kg}\|^2 \right).$$

Conditionnellement aux variables explicatives, les réponses sont indépendantes. Ainsi,

$$f_{\mathbf{Y}}(\mathbf{y}; \Theta) = \prod_{k=1}^{K} f_Y(y_k; \Theta_k),$$

où l'ensemble des paramètres à estimer est $\Theta = \{p_1, \ldots, p_G, \gamma_{11}, \ldots, \gamma_{KG}, \delta_{11}, \ldots, \delta_{KG}, \sigma^2_{11}, \ldots, \sigma^2_{KG}\}$. Cette densité sert de mesure de GoF : $\psi_A(u, \Theta)$. Cependant, concernant Θ, la log-vraisemblance correspondante étant difficile à optimiser, nous utilisons l'algorithme EM [3] pour estimer les paramètres du modèle.

Soient z_{kg} la variable indicatrice latente valant 1 si la réponse y_k est dans le groupe g, le vecteur $Z_k = (z_{kg}\,;\, g = 1, \ldots, G)$ et la matrice $\mathbf{Z} = [Z_k\,;\, k = 1, \ldots, K]$. Conditionnellement à $z_{kg} = 1$, la réponse y_k suit une loi normale multivariée $N_n(\mu_{kg}, \Sigma_{kg})$. La log-vraisemblance complétée est donnée par

$$l(\Theta; \mathbf{Y}, \mathbf{Z}) = \ln\left(f_{\mathbf{YZ}}(\mathbf{y}, \mathbf{z}; \Theta)\right) = \ln\left(\prod_{k=1}^{K} f_{YZ}(y_k, z_k; \Theta_k)\right).$$

L'**étape M** de EM consiste à maximiser l'espérance conditionnelle de la log-vraisemblance complétée $\mathbb{E}\left[l(\Theta; \mathbf{Y}, \mathbf{Z})|\mathbf{Y}; \Theta'\right]$. Cette espérance conditionnelle est mise à jour dans l'**étape E**. Par ailleurs, nous prenons ici la variance de la composante comme SR. Ainsi, le critère à maximiser à l'**étape M** courante devient :

$$c(U, \gamma, \delta, \sigma^2) = s\sum_{g=1}^{G} \ln\left(\|Xu_g\|^2_W\right) + (1-s)\mathbb{E}\left[l(\Theta; \mathbf{Y}, \mathbf{Z})|\mathbf{Y}; \Theta'\right],$$

où $U = \{u_1, \ldots, u_G\}$, Xu_g étant la composante du groupe g, et $\gamma = \{\gamma_{11}, \ldots, \gamma_{KG}\}$, $\delta = \{\delta_{11}, \ldots, \delta_{KG}\}$ et $\sigma^2 = \{\sigma^2_{11}, \ldots, \sigma^2_{KG}\}$ sont les ensembles des paramètres à estimer.

3 Algorithme

Étape E (Espérance conditionnelle)

Pour réaliser l'**étape E** de l'algorithme, nous devons calculer explicitement l'espérance conditionnelle de la log-vraisemblance complétée. Tout calcul fait :

$$\begin{aligned}\mathbb{E}\left[l(\Theta; \mathbf{Y}, \mathbf{Z})|\mathbf{Y}; \Theta'\right] &= -\frac{nK}{2}\ln(2\pi) + \sum_{k=1}^{K}\sum_{g=1}^{G}\alpha_{kg}\ln(p_g) \\ &\quad -\frac{1}{2}\sum_{k=1}^{K}\sum_{g=1}^{G}\alpha_{kg}\left(n\ln(\sigma^2_{kg}) + \frac{1}{\sigma^2_{kg}}\|y_k - (Xu_g)\,\gamma_{kg} - A\delta_{kg}\|^2\right),\end{aligned}$$

avec $\alpha_{kg} = \mathbb{P}(Z_k = g|y_k; \Theta'_k) = \dfrac{p_g N_n(\mu_{kg}, \Sigma_{kg})}{\sum_{g'=1}^{G} p_{g'} N_n(\mu_{kg'}, \Sigma_{kg'})}$, où $Z_k = g$ est un raccourci pour signifier que 1 est à la g-ième place sur l'indicatrice.

Étape M (Maximisation)

L'**étape M** maximise sur Θ : $\mathbb{E}\left[l(\Theta; \mathbf{Y}, \mathbf{Z})|\mathbf{Y}; \Theta'\right]$, sous la contrainte $\sum_{g=1}^{G} p_g = 1$. Ainsi nous devons annuler le gradient en Θ du Lagrangien suivant :

$$L(\Theta, \lambda) = \mathbb{E}\left[l(\Theta; \mathbf{Y}, \mathbf{Z})|\mathbf{Y}; \Theta'\right] - \lambda\left(\sum_{g=1}^{G} p_g - 1\right).$$

On obtient,

$$\nabla_{p_g} L(\Theta, \lambda) = 0 \Leftrightarrow \hat{p}_g = \frac{1}{K}\sum_{k=1}^{K}\alpha_{kg}.$$

Pour estimer γ_{kg} et δ_{kg}, nous posons $T_g = [Xu_g, A]$ et $\theta_{kg} = (\gamma_{kg}, \delta_{kg})^T$. La contrainte ne dépendant pas de θ_{kg}, il suffit d'annuler le gradient de $\mathbb{E}\left[l(\Theta; \mathbf{Y}, \mathbf{Z})|\mathbf{Y}; \Theta'\right]$. On obtient,

$$\nabla_{\theta_{kg}} \mathbb{E}\left[l(\Theta; \mathbf{Y}, \mathbf{Z})|\mathbf{Y}; \Theta'\right] = 0 \Leftrightarrow \hat{\theta}_{kg} = \left(T_g^T W T_g\right)^{-1} \left(T_g^T W y_k\right).$$

On estime de même σ_{kg}^2 :

$$\nabla_{\sigma_{kg}^2} \mathbb{E}\left[l(\Theta; \mathbf{Y}, \mathbf{Z})|\mathbf{Y}; \Theta'\right] = 0 \Leftrightarrow \hat{\sigma}_{kg}^2 = \left\|y_k - T_g \hat{\theta}_{kg}\right\|_W^2.$$

Trouver la composante de chaque classe

Pour trouver les vecteurs de U, nous allons utiliser l'algorithme du gradient normé projeté itéré (Projected Iterated Normed Gradient, PING) [5] dans chaque classe. En effet, le critère global étant une somme de sous-critères relatifs aux classes, il suffit de maximiser isolément chaque sous-critère de classe. Ainsi, pour la classe g, le sous-critère est :

$$\begin{aligned} c(u_g, \gamma, \delta, \sigma^2) &= s \ln\left(\|Xu_g\|_W^2\right) + (1-s)\left[-\frac{nK}{2G}\ln(2\pi) + \ln(p_g)\sum_{k=1}^{K}\alpha_{kg}\right. \\ &\quad \left. -\frac{1}{2}\sum_{k=1}^{K}\alpha_{kg}\left(n\ln(\sigma_{kg}^2) + \frac{1}{\sigma_{kg}^2}\|y_k - (Xu_g)\gamma_{kg} - A\delta_{kg}\|^2\right)\right]. \end{aligned}$$

Algorithme

Algorithme MixRep-SCGLR

On initialise l'algorithme avec les valeurs initiales $u^{(0)}$, $\gamma^{(0)}$, $\delta^{(0)}$, $p_g^{(0)}$ et $t = 0$.
À l'itération $t+1$:

1. On estime les paramètres (hors U) par l'algorithme EM.
   À l'itération $m+1$:

   (a) **Étape E** :
   Pour $k = 1, \ldots, K$:
   Pour $g = 1, \ldots, G$:
   $$\alpha_{kg}^{(m+1)} = \frac{p_g^{(m)} N_n(\mu_{kg}, \Sigma_{kg})}{\sum_{g'=1}^{G} p_{g'}^{(m)} N_n(\mu_{kg'}, \Sigma_{kg'})}$$

   (b) **Étape M** :

   i. Pour $g = 1, \ldots, G$:
   $p_g^{(m+1)} = \frac{1}{K}\sum_{k=1}^{K}\alpha_{kg}^{(m+1)}$

   ii. Pour $k = 1, \ldots, K$:
   Pour $g = 1, \ldots, G$:
   $$\theta_{kg}^{(m+1)} = \left(T_g^{(t)T} W T_g^{(t)}\right)^{-1} T_g^{(t)T} W y_k$$
   $$\sigma_{kg}^{2(m+1)} = \left\|y_k - T_g^{(t)}\theta_{kg}^{(m+1)}\right\|_W^2$$

   Les paramètres $\gamma^{(t+1)}$, $\delta^{(t+1)}$ et $\sigma^{2(t+1)}$ sont alors respectivement égaux à $\gamma^{(m_{\max})}$, $\delta^{(m_{\max})}$ et $\sigma^{2(m_{\max})}$ obtenus à la convergence de l'algorithme EM.

2. On calcule $U = (u_g)_g$ à l'aide de l'algorithme PING.
   Pour $g = 1, \ldots, G$:
   $u_g^{(t+1)} = \underset{u, u^T M^{-1} u = 1}{\operatorname{argmax}} \; c(u_g^{(t)}, \gamma^{(t+1)}, \delta^{(t+1)}, \sigma^{2(t+1)})$
   $T_g^{(t+1)} = [X u_g^{(t+1)}, A]$

Lorsque l'algorithme a convergé, on peut classer les réponses parmi les groupes à l'aide des probabilités *a posteriori* d'inclusion. Une réponse y_k est dans le groupe g si

$$\alpha_{kg}^{(t_{\max})} > \alpha_{kg'}^{(t_{\max})}$$

pour tout $g' \neq g$.

L'algorithme fut testé avec succès sur différents jeux de données simulées.

Remerciements

Cette recherche a été soutenue par le projet GAMBAS financé par l'Agence National de la Recherche (ANR-18-CE02-0025).

Références

[1] Xavier Bry, Catherine Trottier, Thomas Verron, and Frédéric Mortier. Supervised component generalized linear regression using a PLS-extension of the Fisher scoring algorithm. *Journal of Multivariate Analysis*, 119 : 47–60, 2013.

[2] Xavier Bry and Thomas Verron. THEME : THEmatic Model Exploration through multiple co-structure maximization. *Journal of Chemometrics*, 29(12) : 637–647, 2015.

[3] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society : Series B (Methodological)*, 39(1) : 1–22, 1977.

[4] Arthur E. Hoerl and Robert W. Kennard. Ridge regression : applications to nonorthogonal problems. *Technometrics*, 12(1) : 69–82, 1970.

[5] Alston S. Householder. *The theory of matrices in numerical analysis*. Courier Corporation, 2013.

[6] P. McCullagh and J.A. Nelder. 1989, Generalized Linear Models, Chapman and Hall, New York, NY.

Apprentissage d'un classifieur Minimax-Regret pour données hétérogènes et probabilités a priori incertaines

Cyprien Gilet [1] & Susana Barbosa [2] & Lionel Fillatre [1*]

[1] *Université Côte d'Azur, CNRS, I3S, France*
Email: fillatre@i3s.unice.fr, gilet@i3s.unice.fr
[2] *Université Côte d'Azur, CNRS, IPMC, France*
Email: sudocarmo@gmail.com

Résumé. Cet article s'intéresse à l'apprentissage d'un classifieur minimax-regret pour résoudre un problème de classification supervisé entre plusieurs classes. Le regret correspond à la perte de performance d'un classifieur par rapport à la performance optimale atteignable par le classifieur de Bayes lorsque les probabilités a priori de chaque classe sont fixées. Le classifieur minimax-regret minimise le regret maximum d'un classifieur quelles que soient les probabilités a priori. Ce critère est pertinent lorsque les proportions par classe de la base de test diffèrent de celles de la base d'apprentissage. Afin de développer une méthode numérique adaptée aux données hétérogènes, nous discrétisons les données afin d'apprendre le classifieur dans un cas discret.

Mots-clés. Apprentissage supervisé, variables discrètes, classifieur minimax-regret.

Abstract. This paper deals with learning a minimax-regret classifier to solve a supervised classification problem between several classes. The regret corresponds to the loss of performance of a classifier compared to the optimal performance attainable by the Bayes classifier when the prior probabilities of each class are fixed. The minimax-regret classifier minimizes the maximum regret of a classifier regardless of the prior probabilities. This criterion is particularly suitable when the class proportions of the test dataset differ from those of the learning dataset. In order to develop a numerical method adapted to heterogeneous data, we discretize the data to learn the classifier in a discrete case.

Keywords. Supervised learning, discrete features, minimax-regret classifier.

1 Introduction

Contexte : Cet article s'intéresse au problème de classification supervisée qui consiste à calculer, à partir d'un ensemble fini d'observations étiquetées, le classifieur qui minimise le regret empirique maximum. Définissons $K \geq 2$ le nombre de classes, $\mathcal{Y} := \{1, \ldots, K\}$ l'ensemble des classes observées, $\mathcal{X}$ l'espace sur lequel l'ensemble des variables observées sont définies, et n le nombre d'observations dans la base d'apprentissage. On note Y_i la

*Les auteurs remercient la région Provence-Alpes-Côte d'Azur pour son soutien financier.

variable aléatoire caractérisant la classe de l'observation i, et $X_i = [X_{i1}, \ldots, X_{id}]$ le vecteur aléatoire regroupant l'ensemble des d variables descriptives associées à l'observation i. Définissons $\Delta = \{\delta : \mathcal{X} \to \mathcal{Y}\}$ l'ensemble des classifieurs et considérons une règle de décision $\delta \in \Delta$. On définit enfin une fonction de perte $L : \mathcal{Y} \times \mathcal{Y} \to [0, +\infty)$ qui mesure la perte $L(k, l) = L_{kl}$ lorsque le classifieur décide la classe l alors que la vraie classe est k. D'après Ferguson (1967), le risque empirique $\hat{r}_L(\delta)$ de δ, associé à la perte L, s'écrit

$$\hat{r}_L(\delta) = \frac{1}{n}\sum_{i=1}^{n} L(Y_i, \delta(X_i)) = \sum_{k\in\mathcal{Y}} \hat{\pi}_k \hat{R}_{k,L}(\delta), \quad \text{où} \quad \hat{R}_{k,L}(\delta) = \sum_{l\in\hat{\mathcal{Y}}} L_{kl}\, \hat{\mathbb{P}}(\delta(X_i) = l | Y_i = k), \quad (1)$$

$\hat{\pi}_k = \hat{n}_k/n$ correspond à la proportion d'observations appartenant à la classe k, $\hat{n}_k = \sum_{i=1}^{n} \mathbb{1}_{\{Y_i=k\}}$ est le nombre d'observations de la classe k, $\hat{\mathbb{P}}(\delta(X_i) = l \mid Y_i = k) = \frac{1}{\hat{n}_k}\sum_{i:Y_i=k} \mathbb{1}_{\{\delta(X_i)=l\}}$ et $\mathbb{1}_{\{\cdot\}}$ est la fonction indicatrice.

Introduction du classifieur minimax-regret : Notons $\delta_{\hat{\pi}}$ le classifieur $\delta \in \Delta$ calibré à partir des observations $\mathcal{S}_n = (Y_i, X_i)_{i=1,\ldots,n}$ de la base d'apprentissage dont les proportions par classe sont $\hat{\pi} = [\hat{\pi}_1, \ldots, \hat{\pi}_K]$. Ce classifieur est ensuite utilisé pour prédire la classe de nouvelles observations de test $\mathcal{S}'_{n'} = (Y_i, X_i)_{i=n+1,\ldots,n+n'}$. Supposons que l'ensemble des données de test vérifie les proportions $\pi' = [\pi'_1, \ldots, \pi'_K]$, le risque d'erreurs associé au classifieur $\delta_{\hat{\pi}}$ et aux proportions π' est alors noté et défini par

$$\hat{r}_L(\pi', \delta_{\hat{\pi}}) = \sum_{k\in\mathcal{Y}} \pi'_k \hat{R}_{k,L}(\delta_{\hat{\pi}}). \quad (2)$$

Comme illustré sur la Figure 1, ce risque évolue linéairement lorsque les proportions π' diffèrent de $\hat{\pi}$. Ceci peut donc poser problème lorsque les proportions de la base d'apprentissage sont incertaines. Le classifieur minimax, étudié dans Gilet et al. (2019), est une solution légitime car elle tend à fournir un classifieur robuste face aux problèmes de distributions a priori incertaines. Néanmoins, il est bien connu qu'un classifieur minimax est pessimiste par nature: il maximise le risque d'erreurs moyen. Comme décrit dans Berger (1985), une solution moins pessimiste pour rendre une règle de décision robuste à des différences de proportions entre $\mathcal{S}_n$ et $\mathcal{S}'_{n'}$ consiste à calculer le classifieur

$$\tilde{\delta} \;=\; \underset{\delta\in\Delta}{\operatorname{argmin}} \max_{\pi\in\mathbb{S}} D_L(\pi, \delta), \quad (3)$$

où $\mathbb{S}$ est le simplexe probabiliste de dimension K et le regret $D_L(\pi, \delta)$ est défini par

$$D_L(\pi, \delta) \;:=\; \hat{r}_L(\pi, \delta) - \min_{\hat{\delta}\in\Delta} \hat{r}_L(\pi, \hat{\delta}) \;=\; \hat{r}_L(\pi, \delta) - V_L(\pi). \quad (4)$$

Le regret défini en (4) décrit l'écart entre le risque du classifieur δ et le risque minimum réalisable $V_L(\pi) = \min_{\hat{\delta}\in\Delta} \hat{r}_L(\pi, \hat{\delta})$ pour une distribution des proportions par classe π fixée.

État de l'art : Généralement, le problème minimax-regret est résolu de façon analytique Berger (1985) pour des problèmes de classification spécifiques. Il n'existe pas de méthode générale pour calculer le classifieur minimax-regret dans le cas général. Alaíz-Rodríguez et al. (2007) se sont intéressés au problème (3) et ont proposé un algorithme assez générique mais qui ne fonctionne que lorsque le problème d'optimisation implique des fonctions différentiables au point optimum recherché. Notre approche propose un algorithme très différent qui résoud le problème sans condition de différentiabilité lorsque la base d'apprentissage est discrétisée.

Contributions : Cet article s'intéresse au cas où les variables observées sont discrètes ($\mathcal{X} \subset \mathbb{N}^d$) ou discrétisées, et $K \geq 2$ classes sont à prédire. Nous montrons que le calcul du classifieur minimax-regret (3) pour une fonction de perte L est équivalent au calcul d'un classifieur minimax pour une fonction de perte modifiée déduite de L. Nous proposons un algorithme itératif permettant d'estimer le classifieur minimax-regret. Nous illustrons la robustesse du classifieur sur des données réelles comportant 130 variables numériques.

2 Classifieur minimax-regret pour variables discrètes

Le fait de travailler sur des bases de données hétérogènes contenant à la fois des variables descriptives catégorielles et numériques est difficile. Une approche raisonnable est de discrétiser les variables numériques pour se ramener à ne traiter seulement des variables discrètes, ce qui conduit à des résultats analytiques intéressants comme soulignés dans Devroye et al. (1996); Braga-Neto and Dougherty (2005); Gilet et al. (2019). Lorsque l'ensemble des d variables descriptives sont discrètes ou discrétisées, il existe un nombre fini T de "profiles" possibles permettant de caractériser chaque combinaison des d variables. Nous avons donc $\mathcal{X} = \{x_1, \ldots, x_T\}$ où $x_t \in \mathbb{N}^d$.

Le théorème suivant nous permet de simplifier l'expression du regret défini en (4).

Théorème 1. *Le regret $D_L(\pi, \delta)$ défini en (4) vérifie*

$$\min_{\delta \in \Delta} \max_{\pi \in \mathbb{S}} D_L(\pi, \delta) = \min_{\delta \in \Delta} \max_{\pi \in \mathbb{S}} \tilde{D}_L(\pi, \delta) = \max_{\pi \in \mathbb{S}} \min_{\delta \in \Delta} \tilde{D}_L(\pi, \delta) \tag{5}$$

avec $\tilde{D}_L(\pi, \delta) = \hat{r}_L(\pi, \delta) - \sum_{k \in \mathcal{Y}} \pi_k L_{kk}$.

Le théorème 1 établit deux résultats importants très utiles en pratique. D'une part, il n'est pas utile de calculer $\min_{\hat{\delta} \in \Delta} \hat{r}_L(\pi, \hat{\delta})$ puisqu'il suffit de calculer $\tilde{D}_L(\pi, \delta)$ au lieu de $D_L(\pi, \delta)$. D'autre part, le problème de minimisation-maximisation est équivalent à un problème de maximisation-minimisation. Ce second résultat a une grande utilité car le problème $\min_{\delta \in \Delta} \tilde{D}_L(\pi, \delta)$ peut être résolu de façon analytique grâce à la proposition 1.

Proposition 1. *L'écart $\tilde{D}_L(\pi, \delta)$ est égal au risque $\hat{r}_{\tilde{L}}(\pi, \delta)$ donné en (2) associé à la fonction de perte $\tilde{L}$ définie par $\tilde{L}_{kl} = L_{kl} - L_{kk}$ pour $1 \leq k, l \leq K$: $\tilde{D}_L(\pi, \delta) = \hat{r}_{\tilde{L}}(\pi, \delta)$.*

La proposition 1 montre donc que le classifieur minimax-regret pour la perte L revient en pratique à calculer un classifieur minimax pour le risque $\hat{r}_{\tilde{L}}(\pi, \delta)$ associé à la perte $\tilde{L}$. Dans le cas discret, la valeur $\min_{\delta \in \Delta} \hat{r}_{\tilde{L}}(\pi, \delta)$ peut être calculée de façon analytique. Elle est égale au risque $V_{\tilde{L}}(\pi) = \hat{r}_{\tilde{L}}\left(\pi, \delta^B_{\pi,\tilde{L}}\right)$ du classifieur de Bayes empirique

$$\delta^B_{\pi,\tilde{L}} : X_i \mapsto \arg\min_{l \in \hat{\mathcal{Y}}} \sum_{t \in \mathcal{T}} \sum_{k \in \mathcal{Y}} \tilde{L}_{kl}\, \pi_k\, \hat{p}_{kt}\, \mathbb{1}_{\{X_i = x_t\}}, \text{ avec } \hat{p}_{kt} := \frac{1}{\hat{n}_k} \sum_{i:Y_i=k} \mathbb{1}_{\{X_i = x_t\}}, \quad (6)$$

où $\hat{p}_{kt}$ est l'estimation (sur la base d'apprentissage) de la probabilité d'observer le profil x_t dans la classe k. La notation $\delta^B_{\pi,L}$ souligne le fait que, de façon générale, le classifieur de Bayes dépend des proportions π et d'une fonction de perte L. De ce fait, d'après le Théorème 1 et la Proposition 1, le classifieur minimax-regret $\tilde{\delta}$ correspond au classifieur Bayésien $\delta^B_{\tilde{\pi},L}$ associé à la distribution $\tilde{\pi}$, et donc au classifieur Bayésien $\delta^B_{\tilde{\pi},\tilde{L}}$, qui maximise le risque de Bayes minimum $V_{\tilde{L}}(\pi)$ par rapport à π. Dans Gilet et al. (2019), les auteurs ont montré que, dans le cas discret, la maximisation de $V_{\tilde{L}}(\pi)$ conduit à un problème d'optimisation non-différentiable. Les auteurs utilisent une méthode itérative de sous-gradient projeté. Leur algorithme converge vers la proportion optimale $\tilde{\pi}$ avec une erreur de convergence bien contrôlée. L'algorithme est détaillé dans Gilet et al. (2019).

3 Expériences numériques

Base de données APS Scania Trucks : La base de données réelle et publique Scania (2016) s'intéresse au système de pression d'air (APS), utilisé pour diverses fonctions dans les camions Scania, telles que le freinage et les changements de vitesse. Des mesures d'un composant spécifique de l'APS ont été collectées sur plusieurs camions Scania. L'objectif est de prédire une défaillance potentielle de ce composant. On considère donc $K = 2$ classes, où la classe 1 correspond à l'absence de panne, et la classe 2 caractérise les APS présentant une défaillance du composant. On considère la matrice de coûts L telle que $L_{11} = 0$, $L_{12} = 40$, $L_{21} = 500$ et $L_{22} = 30$, de sorte que le fait de bien prédire qu'il n'y aura pas de pannes ne coûte rien, alors que la prédiction d'une faille non-existante est de 40$ (le prix de la révision). De plus, manquer une panne est grave et coûte 500$, alors que bien prédire une panne coûte seulement le prix du remplacement du composant 30$. Les experts ont mis a disposition un échantillon d'apprentissage et un échantillon de test contenant respectivement 54731 et 14578 observations. Chaque observation est décrite par 130 variables descriptives numériques. Les proportions par classe sont très déséquilibrées, avec 98.91% d'observations de la classe 1 et 1.09% d'observations de la classe 2.

Apprentissage sur les données d'entraînement : Pour calibrer notre classifieur minimax-regret à partir de l'échantillon d'apprentissage, nous avons dans un premier temps discrétisé les variables numériques à l'aide d'une quantification Kmeans, comme décrit

dans Gilet et al. (2019). Ensuite, comme expliqué à la fin de la Section 2, nous utilisons l'algorithme établi dans Gilet et al. (2019) pour calculer les proportions $\tilde{\pi} = (\tilde{\pi}_1, 1 - \tilde{\pi}_1)$ qui maximisent $V_{\tilde{L}}$ sur $\mathbb{S}$. Cette étape est illustrée sur la Figure 1, droite, qui montre la courbe $V_{\tilde{L}}$. La Figure 1, gauche, compare le test de Bayes $\delta^B_{\hat{\pi},L}$, le test minimax $\delta^B_{\tilde{\pi},L}$ calculé à partir de l'algorithme Gilet et al. (2019), le test minimax-regret $\delta^B_{\tilde{\pi},L}$ et le risque minimum $V_L(\pi)$. La comparaison des courbes V_L et $V_{\tilde{L}}$ montre l'impact de la modification de la fonction de perte sur le risque minimum.

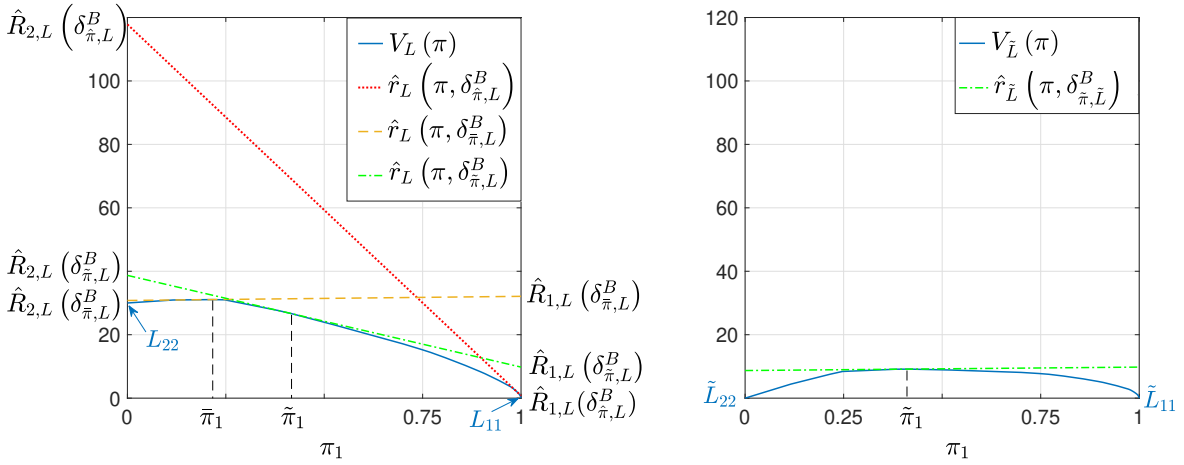


Figure 1: Gauche. Risques associés au classifieur de Bayes discret $\delta^B_{\hat{\pi},L}$, au classifieur minimax $\delta^B_{\tilde{\pi},L}$, et au classifieur minimax-regret $\delta^B_{\tilde{\pi},L}$, lorsque π' diffère des proportions d'apprentissage $\hat{\pi} = [0.9891, 0.0109]$. Puisque $K = 2$ et $\pi' \in \mathbb{S}$, pour chaque classifieur δ, le risque peut se réécrire comme $\hat{r}_L(\pi', \delta) = \pi'_1[\hat{R}_{1,L}(\delta) - \hat{R}_{2,L}(\delta)] + \hat{R}_{2,L}(\delta)$. **Droite.** Calcul des proportions $\tilde{\pi}$ qui maximisent $V_{\tilde{L}}$ en utilisant l'algorithme minimax établi dans Gilet et al. (2019).

Évaluation sur les données de test : Pour évaluer la robustesse des classifieurs lorsque les proportions par classe sont incertaines, nous avons généré $m = 1000$ distributions $\pi^{(s)}$, $s \in \{1, \ldots, m\}$, uniformément réparties sur le simplexe $\mathbb{S}$. Ensuite, à partir de la base de test, nous avons généré aléatoirement m sous-échantillons de test contenant chacun 220 observations et satisfaisant l'une des distributions $\pi^{(s)}$. Enfin, pour chaque $\pi^{(s)}$, nous avons calculé la déviation (4) de chaque classifieur présenté dans la Figure 1. Pour comparer les résultats de chaque classifieur, nous considérons les deux critères suivants:

$$D_{\text{MAX}}(\delta) = \max_{s \in \{1,\ldots,m\}} \left| D_L\left(\pi^{(s)}, \delta\right) \right|, \quad \text{et} \quad D_{\text{MOY}}(\delta) = \frac{1}{m} \sum_{s=1}^{m} \left| D_L\left(\pi^{(s)}, \delta\right) \right|.$$

Ainsi, plus les valeurs de $D_{\text{MAX}}(\delta)$ et $D_{\text{MOY}}(\delta)$ sont faibles, plus le classifieur δ reste proche du classifieur Bayes empirique discret sur le simplexe $\mathbb{S}$, et donc plus δ est robuste face à différents changements des proportions par classe. Les résultats de cette expérience sont

décrits dans le Tableau 1, et montrent que le classifieur minimax-regret a bien les plus petites déviations maximum et moyenne.

Table 1: Résultats moyennés de $D_{\text{MAX}}(\delta)$ et $D_{\text{MOY}}(\delta)$ pour chaque classifieur δ après 10 répétitions de l'expérience. Les résultats sont présentés comme [moyenne $\pm$ écart-type]. Les résultats en gris correspondent à ce que l'on obtient analytiquement, sur la Figure 1, gauche, pour chaque classifieur lors de l'étape d'apprentissage. Le biais entre les résultats de l'apprentissage et ceux des tests s'expliquent par l'erreur de généralisation.

	Figure 1			Échantillons tests		
Classifieurs	$\delta^B_{\tilde{\pi},L}$	$\delta^B_{\tilde{\pi},L}$	$\delta^B_{\tilde{\pi},L}$	$\delta^B_{\tilde{\pi},L}$	$\delta^B_{\tilde{\pi},L}$	$\delta^B_{\tilde{\pi},L}$
$D_{\text{MAX}}(\delta)$	87.26	30.87	8.69	128.02 ± 0.06	33.09 ± 0.83	23.46 ± 0.07
$D_{\text{MOY}}(\delta)$	38.38	9.27	2.39	58.59 ± 0.16	18.60 ± 0.09	9.86 ± 0.06

4 Perspectives

En nous basant sur Gilet et al. (2019), nous souhaitons étendre notre approche pour calculer un classifieur Γ-minimax-regret où les proportions par classe seraient bornées individuellement. Ces bornes seraient définies par des experts du domaine d'application en capacité d'estimer l'incertitude sur les proportions de chaque classe.

References

Alaíz-Rodríguez, R., Guerrero-Curieses, A., and Cid-Sueiro, J. (2007). Minimax regret classifier for imprecise class distributions. *JMLR*, 8:103–130.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis; 2nd ed.* Springer Series in Statistics. Springer, New York.

Braga-Neto, U. and Dougherty, E. R. (2005). Exact performance of error estimators for discrete classifiers. *Elsevier Pattern Recognition*, 38(11):1799–1814.

Devroye, L., Györfi, L., and Lugosi, G. (1996). *A Probablistic Theory of Pattern Recognition*, volume 31. Springer.

Ferguson, T. (1967). *Mathematical Statistics: A Decision Theoretic Approach.* Academic Press.

Gilet, C., Barbosa, S., and Fillatre, L. (2019). Minimax classifier with box constraint on the priors. In *Machine Learning for Health (ML4H) at NeurIPS 2019*. Proceedings of Machine Learning Research.

Scania, C. A. (2016). *APS Failure at Scania Trucks Data Set.* UCI Machine Learning Repository.

Rank-R Multiway Logistic Regression

Fabien Girka [1], Pierrick Chevaillier [1], Arnaud Gloaguen [2,3], Giulia Gennari [4], Ghislaine Dehaene-Lambertz [4], Laurent Le Brusquet [2] & Arthur Tenenhaus[2]

[1] *CentraleSupélec, 91190, Gif-sur-Yvette, France,*
[2] *Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des signaux et systèmes, 91190, Gif-sur-Yvette, France.*
[3] *UNATI, Neurospin, Institut Joliot, CEA, Université Paris-Saclay, France,*
[4] *INSERM, UMR992, Neurospin, Institut Joliot, CEA, Université Paris-Saclay, France*
fabien.girka@supelec.fr, pierrick.chevaillier@supelec.fr, arnaud.gloaguen@centralesupelec.fr

Résumé. Beaucoup de données ont une structure intrinsèque tensorielle lorsque, par exemple, plusieurs modalités de la même variable ont été observées sur chaque individu. Les approches multivoie deviennent alors un choix naturel pour analyser ces données. Les versions standards de ces approches consistent à imposer au vecteur de poids d'être une décomposition de PARAFAC de rang 1. Pour certaines applications, les données peuvent cependant s'avérer trop complexes pour que cette décomposition soit valide. Ce papier présente une version de la régression logistique multivoie associée à une décomposition de rang R, l'objectif étant de proposer une méthode de classification applicable aux situations où la contrainte de rang 1 serait trop restrictive. Un algorithme de directions alternées est proposé pour la régression logistique multivoie de rang R. Les performances de cette méthodes sont évaluées sur des données d'électroencéphalogrammes (EEG).

Mots-clés. Analyse multivoie, régression logistique, EEG

Abstract. Data often has an inherent tensor structure (e.g. data where the same set of variables is collected at different occasions). To deal with such data, multiway models become a natural choice. Standard multiway models impose weight vectors to be rank-1 PARAFAC decomposition. However in some applications, this constraint appears to be too restrictive. This paper presents a more general version of multiway logistic regression (MLR) associated to a rank-R decomposition. The objective of such an approach is to propose a classification model that can cope with situations where rank-1 constraint may be too restrictive. An alternating direction algorithm is proposed for rank-R MLR and its performances are evaluated on electroencephalogram (EEG) data.

Keywords. Multiway analysis, logistic regression, EEG

1 Introduction

Multiway data appears in many research fields as neuroscience, chemometrics or social networks to name a few. It occurs when the same set of variables is collected through

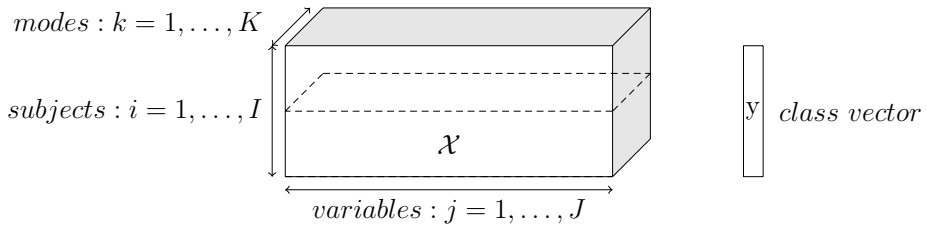


Figure 1: Three-way data, each sample is represented by K vectors of length J

different modes. For example, this is the case for spatio-temporal data (several images collected at different time steps) or when a set of measures is acquired through multiple sensors. See [1] for an overview of methods for the analysis of multiway data.

Let $\mathcal{X} = \{\mathrm{X}_{ijk}\}_{1\leq i\leq I, 1\leq j\leq J, 1\leq k, \leq K}$ be a third order tensor of dimension $I \times J \times K$ where I is the number of subjects, J the number of variables and K the number of modes (see Figure 1). The roles of variables and modes are symmetrical and thus can be interchanged.

A first non-multiway option to deal with such tensor data is to unfold the tensor by concatenating its frontal slices $\mathrm{X}_{..k}$ next to each other, leading to $\mathrm{X} = \begin{bmatrix}\mathrm{X}_{..1} & \dots & \mathrm{X}_{..K}\end{bmatrix}$ which is a matrix of size $I \times JK$ on which classic statistical methods can be applied. However, such an approach leads to issues: (i) A problem of size JK could be computationally impractical for standard computers, (ii) the higher the dimension/order of the tensor, the more the number of variables obtained by flattening, the higher the risk of overfitting is, (iii) the results are not easy to interpret as the considered model does not permit a separate interpretation of the influence of the variables and modes.

A second approach largely used in the multiway literature when the model to construct uses the tensor variables through a linear form is to impose a Kronecker constraint on the weight vector: $\boldsymbol{\beta} = \boldsymbol{\beta}^J \otimes \boldsymbol{\beta}^K$. Taking into account the multiway structure of the data with a Kronecker constraint reduces the degree of freedom from JK to $J + K$, which may limit overfitting effects and computation time. Moreover, the study of the contributions of the variables and modes is made easier by analysing each vector $\boldsymbol{\beta}^J$ and $\boldsymbol{\beta}^K$ separately.

This rank-1 constraint was used for example in Multiway Logistic Regression (MLR) [2]. However, such a constraint can be too restrictive when effect of variables and modes are not strictly parallel. Therefore a higher rank decomposition for $\boldsymbol{\beta}$ can be considered :

$$\boldsymbol{\beta} = \sum_{r=1}^{R} \boldsymbol{\beta}_r^K \otimes \boldsymbol{\beta}_r^J \tag{1}$$

Similar rank R approaches has been studied in [3] for Generalized Linear Model and [4] for Support Vectors Machines. In this paper, we propose a rank-R Multiway Logistic Regression.

This paper is organized as follows: Section 2 presents a short reminder of (regularized) logistic regression. The rank-R logistic regression is described in Section 3. Section 4 presents an application on EEG data.

2 Regularized Logistic Regression

Logistic regression can be directly used on tensor data by unfolding the tensor in a matrix. Let x_i be the vector of the KJ observed variables of subject i (x_i is then the i^{th} line of the unfolded tensor X), y_i its class (either 0 or 1). Logistic regression relies on maximising the conditional log-likelihood $\sum_{i=1}^{n} \log \mathbb{P}(y_i \mid \mathrm{x}_i)$ under the assumption that the conditional probabilities log-ratio is linear:

$$\log\left(\frac{\mathbb{P}(y=1 \mid \mathrm{x}_i)}{1-\mathbb{P}(y=1 \mid \mathrm{x}_i)}\right) = \beta_0 + \boldsymbol{\beta}^{\mathsf{T}}\mathrm{x}_i$$

with β_0 and $\boldsymbol{\beta}$ parameters of the model. From this model it comes the following expression of the regularized log-likelihood:

$$\mathcal{C}(\beta_0, \boldsymbol{\beta}, \mathrm{X}, \mathrm{y}, \lambda) = \sum_{i=1}^{n} y_i(\beta_0 + \boldsymbol{\beta}^{\mathsf{T}}\mathrm{x}_i) - \log(1+\exp(\beta_0 + \boldsymbol{\beta}^{\mathsf{T}}\mathrm{x}_i)) - \lambda g(\boldsymbol{\beta}) \tag{2}$$

where $\lambda > 0$ is a regularization parameter that can be tuned and $g(\beta)$ is a penalty term.

3 Multiway Logistic Regression

In this section, rank-R multiway logistic regression (R-MLR) is presented. R-MLR is defined as the following optimization problem:

$$\max_{\beta_0,\ \boldsymbol{\beta}^K,\ \boldsymbol{\beta}^J} \ \mathcal{C}(\beta_0, \boldsymbol{\beta}, \mathrm{X}, \mathrm{y}, \lambda) \quad \text{s.t.} \quad \boldsymbol{\beta} = \sum_{r=1}^{R} \boldsymbol{\beta}_r^K \otimes \boldsymbol{\beta}_r^J \tag{3}$$

where $\boldsymbol{\beta}^J = \left[(\boldsymbol{\beta}_1^J)^{\mathsf{T}} \ \ldots \ (\boldsymbol{\beta}_R^J)^{\mathsf{T}}\right]^{\mathsf{T}}$ and $\boldsymbol{\beta}^K = \left[(\boldsymbol{\beta}_1^K)^{\mathsf{T}} \ \ldots \ (\boldsymbol{\beta}_R^K)^{\mathsf{T}}\right]^{\mathsf{T}}$. An alternating direction algorithm that monotonically converges is proposed to solve the optimization problem (3). First, we can note that $\boldsymbol{\beta}^{\mathsf{T}}\mathrm{x}_i$ can be expressed as:

$$\left(\sum_{r=1}^{R} \boldsymbol{\beta}_r^J \otimes \boldsymbol{\beta}_r^K\right)^{\mathsf{T}} \mathrm{x}_i = \sum_{r=1}^{R} (\boldsymbol{\beta}_r^J)^{\mathsf{T}} \left((\boldsymbol{\beta}_r^K)^{\mathsf{T}} \otimes \mathrm{I}_J\right) \mathrm{x}_i \doteq \sum_{r=1}^{R} (\boldsymbol{\beta}_r^J)^{\mathsf{T}} \mathrm{z}_{r,i}^J \tag{4}$$

In addition, two types of regularisation are considered in this paper. First an ℓ_2 penalty with $g(\boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{\beta}\|_2^2$ that can be expressed in terms of $\boldsymbol{\beta}^J$ and $\boldsymbol{\beta}^K$ as follows:

$$\boldsymbol{\beta}^\mathsf{T}\boldsymbol{\beta} = \left(\sum_{r=1}^{R} \boldsymbol{\beta}_r^J \otimes \boldsymbol{\beta}_r^K\right)^\mathsf{T}\left(\sum_{r=1}^{R} \boldsymbol{\beta}_r^J \otimes \boldsymbol{\beta}_r^K\right) = \sum_{r=1}^{R} \|\boldsymbol{\beta}_r^J\|_2^2\|\boldsymbol{\beta}_r^K\|_2^2 + 2\sum_{i=1}^{R}\sum_{j=i+1}^{R} (\boldsymbol{\beta}_i^J)^\mathsf{T}\boldsymbol{\beta}_j^J(\boldsymbol{\beta}_i^K)^\mathsf{T}\boldsymbol{\beta}_j^K$$
$$= (\boldsymbol{\beta}^J)^\mathsf{T}\mathrm{R}^J\boldsymbol{\beta}^J$$

where $\mathrm{R}^J = \left((\mathrm{B}^K)^\mathsf{T}\mathrm{B}^K\right) \otimes \mathrm{I}_J$, with $\mathrm{B}^K = \begin{bmatrix}\boldsymbol{\beta}_1^K & \dots & \boldsymbol{\beta}_R^K\end{bmatrix}$. As long as the columns of B^K are not colinear, R^J is symmetric positive definite and $\mathrm{Q}_{\ell_2}^J = (\mathrm{R}^J)^{-\frac{1}{2}} = \left((\mathrm{B}^K)^\mathsf{T}\mathrm{B}^K\right)^{-\frac{1}{2}} \otimes \mathrm{I}_J$ is well defined.

Furthermore, by setting $g(\boldsymbol{\beta}) = \sum_{r=1}^{R} \|\boldsymbol{\beta}_r^J \otimes \boldsymbol{\beta}_r^K\|_1$, R-MLR with variable selection can also be defined. The interest of this structured sparsity-inducing norms is that, with a single parameter λ, the sparsity will spread (driven by the data) among ranks, variables and modes. This penalty term can be expressed as a function of $\boldsymbol{\beta}^J$ and $\boldsymbol{\beta}^K$:

$$\sum_{r=1}^{R} \|\boldsymbol{\beta}_r^J \otimes \boldsymbol{\beta}_r^K\|_1 = \|\mathrm{R}^J\boldsymbol{\beta}^J\|_1 \tag{5}$$

with $\mathrm{R}^J = \left(\|\boldsymbol{\beta}_r^K\|_1 \mathrm{I}_J\right)_{r\in\{1\dots R\}}$ a block diagonal matrix. As before, $\mathrm{Q}_{\ell_1}^J = (\mathrm{R}^J)^{-1} = \left(\|\boldsymbol{\beta}_r^K\|_1^{-1}\mathrm{I}_J\right)_{r\in\{1\dots R\}}$ is well defined.

As a consequence, the objective function of the optimization problem (3) can be expressed in terms $\boldsymbol{\beta}^J$ and $\boldsymbol{\beta}^K$ and therefore can be maximized with respect to β_0, $\boldsymbol{\beta}^J$ and $\boldsymbol{\beta}^K$ using an alternating direction algorithm. Indeed, optimising w.r.t. $(\beta_0, (\mathrm{Q}^J)^{-1}\boldsymbol{\beta}^J)$ can be seen as applying logistic regression to maximise criterion $\mathcal{C}^J = \mathcal{C}(\beta_0, (\mathrm{Q}^J)^{-1}\boldsymbol{\beta}^J, Q^J Z^J, \mathrm{y}, \lambda)$ with Z^J a matrix of size $I \times J$ and $\mathrm{z}_i^J = \begin{bmatrix}(\mathrm{z}_{1,i}^J)^\mathsf{T} & \dots & (\mathrm{z}_{R,i}^J)^\mathsf{T}\end{bmatrix}^\mathsf{T}$. As variables and modes play symmetric roles, the same can be done with $\boldsymbol{\beta}^K$. We can now derive the R-MLR algorithm presented in Algorithm 1.

4 Application on EEG data and discussion

The objective of this study was to identify whether the infant's brain encodes the phonetic features used by linguists to describe speech. 24 different consonant-vowel syllables were presented to 25 infants in a randomized order every 1000 ms during one-hour-long experimental sessions. Brain responses were recorded at 500 Hz with a high-density EEG net comprising 252 channels. After pre-processing, this EEG experiment yields 25 tensors of size 24 syllables $\times$ 500 time steps $\times$ 252 channels each. The consonants varied along the manner of articulation separating the 24 syllables in 2 classes that we want to predict.

Algorithm 1: Rank R Multiway Logistic Regression

Inputs: $\epsilon > 0,\ \lambda,\ R,\ \boldsymbol{\beta}^{K(0)}$, penalty

$q \longleftarrow 0$

repeat

$\mathrm{Z}_r^J = \sum_{k=1}^{K} \left(\boldsymbol{\beta}_r^{K(q)}\right)_k \mathrm{X}_{..k}$ for $r \in \{1, \ldots, R\}$

$\left(\mathrm{Z}^J, \boldsymbol{\beta}^{K(q)}\right) \longleftarrow \left(\left[(\mathrm{Z}_1^J)^\mathsf{T} \quad \ldots \quad (\mathrm{Z}_R^J)^\mathsf{T}\right]^\mathsf{T},\ \left[(\boldsymbol{\beta}_1^{K(q)})^\mathsf{T} \quad \ldots \quad (\boldsymbol{\beta}_R^{K(q)})^\mathsf{T}\right]^\mathsf{T}\right)$

if ℓ_1 penalty $Q^J \longleftarrow Q_{\ell_1}^J$ **else** $Q^J \longleftarrow Q_{\ell_2}^J$

$(\beta_0^{(q)}, (Q^J)^{-1}\boldsymbol{\beta}^{J(q)}) \longleftarrow \mathrm{argmax}_{\beta_0, \boldsymbol{\beta}}\, \mathcal{C}\left(\beta_0, (Q^J)^{-1}\boldsymbol{\beta}, Q^J\mathrm{Z}^J, \mathrm{y}, \lambda\right)$

$\mathrm{Z}_r^K = \sum_{j=1}^{J} \left(\boldsymbol{\beta}_r^{J(q)}\right)_j \mathrm{X}_{.j.}$ for $r \in \{1, \ldots, R\}$

$\left(\mathrm{Z}^K, \boldsymbol{\beta}^{J(q)}\right) \longleftarrow \left(\left[(\mathrm{Z}_1^K)^\mathsf{T} \quad \ldots \quad (\mathrm{Z}_R^K)^\mathsf{T}\right]^\mathsf{T},\ \left[(\boldsymbol{\beta}_1^{J(q)})^\mathsf{T} \quad \ldots \quad (\boldsymbol{\beta}_R^{J(q)})^\mathsf{T}\right]^\mathsf{T}\right)$

if ℓ_1 penalty $Q^K \longleftarrow Q_{\ell_1}^K$ **else** $Q^K \longleftarrow Q_{\ell_2}^K$

$(\beta_0^{(q)}, (Q^K)^{-1}\boldsymbol{\beta}^{K(q+1)}) \longleftarrow \mathrm{argmax}_{\beta_0, \boldsymbol{\beta}}\, \mathcal{C}\left(\beta_0, (Q^K)^{-1}\boldsymbol{\beta}, Q^K\mathrm{Z}^K, \mathrm{y}, \lambda\right)$

$q \longleftarrow q + 1$

until $|\mathcal{C}^K - \mathcal{C}^J| < \epsilon\ \mathcal{C}^J$;

return $(\boldsymbol{\beta}^{K(q)}, \boldsymbol{\beta}^{J(q)}, \beta_0^{(q)})$

Regularized logistic regression and R-MLR with R from 1 to 3 are evaluated and compared using Leave-One-Out cross validation: the model is trained on all infants except one and tested on the remaining one. The Area Under the ROC Curve (AUC), the regularisation parameter λ and the computational time used to run the 25 folds are reported in Table 1.

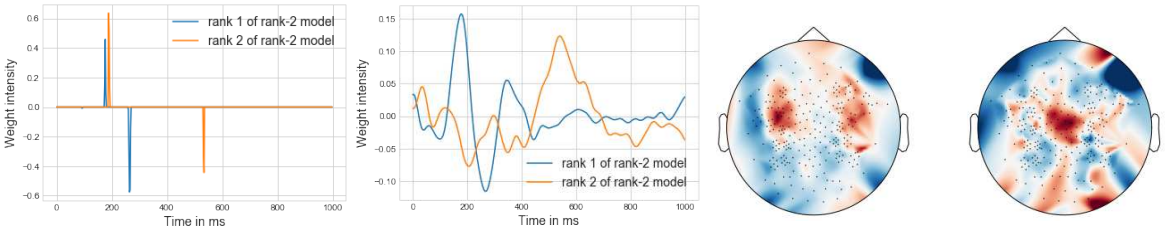
From Table 1, we show that, for complex data such as EEG data, higher rank MLR enable to outperform both rank-1 MLR and regularized logistic regression. For ℓ_1 penalty, finding the best regularization parameter for logistic regression is really challenging given the required computation time. MLR enables to cut down drastically this computation time and yields better results for rank 3.

Model	AUC	λ	Time (in s)
LR	0.822 ± 0.19	1000	892
1-MLR	0.815 ± 0.23	6000	282
2-MLR	0.852 ± 0.17	17500	340
3-MLR	0.857 ± 0.18	17500	498

Model	AUC	λ	Time (in s)
LR	0.830 ± 0.19	20	117000
1-MLR	0.816 ± 0.22	6.5	257
2-MLR	0.826 ± 0.18	10	409
3-MLR	0.853 ± 0.19	15	557

Table 1: Cross validation results by Leave-One-Out for ℓ_2 (left) and ℓ_1 (right) penalties

In addition, multiway models enable the graphical display of the weights and get insights into the importance of variables and modes separately (see Figure 2).



(a) ℓ_1 penalty (b) ℓ_2 penalty (c) ℓ_2 penalty

Figure 2: Weight visualisation for 2-MLR trained on all 25 infants: (a) and (b) show time step weights and (c) shows the topomap of electrode weights.

Topomaps associated with the ℓ_1 penalty (not displayed here due to lack of space) are consistent with the ℓ_2 penalty results. However, similarly to the time weights of the ℓ_1 penalty, they are very sparse with few (less than 5%) isolated selected channels. This variable selection leads spatial/time resolution that is lower than the usual phenomenon observed and described by neuroscientists. Hence, future works include to combine ℓ_1 and ℓ_2 penalties in order to try to catch smoother effects in time and space.

5 Conclusion

Rank-R Multiway Logistic Regression is presented in this paper and shows promising results in EEG application. While R-MLR is presented for third order tensors, it can be generalized to any higher order.

Bibliography

[1] Bro, R. (2000), Multi-way Analysis in the Food Industry - Models, Algorithms, and Applications ICSLP Proceedings.
[2] Le Brusquet L., Lechuga G., Tenenhaus A. (2014), Régression Logistique Multivoie, 46ème Journée de Statistique.
[3] Zhou, Hua; Li, Lexin; Zhu, Hongtu (2013), Tensor Regression with Applications in Neuroimaging Data Analysis, Journal of the American Statistical Association, vol.108.
[4] T. Lyu, E. F. Lock, and L. E. Eberly (2017), Discriminating sample groups with multi-way data, Biostatistics (Oxford, England)

Lissage particulaire en ligne pour une large classe de processus de diffusion partiellement observ.

Marie-Pierre Etienne[1], Pierre Gloaguen[2], Sylvain Le Corff[3], Jimmy Olsson[4]

[1] *Agrocampus-Ouest, marie-pierre.etienne@agrocampus-ouest.fr*
[2] *Agroparistech, pierre.gloaguen@agroparistech.fr*
[3] *Telecom Sud-Paris, sylvain.le_corff@telecom-sudparis.eu*
[4] *KTH Royal institute of technology, jimmyol@kth.se*

Résumé. Cet article propose un nouvel algorithme de Monte Carlo squentiel pour effectuer une estimation du maximum de vraisemblance dans les processus de diffusion partiellement observs. L'apprentissage de tels modles gnratifs et l'obtention d'estimateurs faible variance des distributions postrieures des tats latents conditionnellement aux observations est un dfi car les densits de transition des tats latents ne peuvent pas tre values de manire ponctuelle. Dans cet article, une tape d'chantillonnage d'importance rtrospective est introduite pour estimer de telles distributions postrieures au lieu de l'approche habituelle d'acceptation-rejet. Dans le contexte des diffusions partiellement observes, cela permet d'tendre largement la classe des modles pour lesquels des algorithmes non biaiss existaient.

Mots-clés. Processus de diffusion, Modle de Markov cach, Echantillonnage prfrentiel, filtre particulaire, Monte Carlo squentiel.

Abstract. This paper proposes a new Sequential Monte Carlo algorithm to perform maximum likelihood estimation in partially observed diffusion processes. Training such generative models and obtaining low variance estimators of the posterior distributions of the latent states given the observations is challenging as the transition densities of the latent states cannot be evaluated pointwise. In this paper, a backward importance sampling step is introduced to estimate such posterior distributions instead of the usual acceptance-rejection approach. This allows to use unbiased estimates of the unknown transition densities available under mild assumptions for multivariate stochastic differential equations while acceptance-rejection based methods require strong conditions to obtain upper-bounded estimators.

Keywords. Diffusion processes, Hidden Markov models, Importance sampling, Particle filtering, Sequential Monte Carlo

1 Introduction

In this article, we focus in algorithms to perform inference in state space models. In such models, a key feature to perform parameter inference is to be able of computing expectations with respect to the law of hidden states conditionnally to the observations. In this paper, a new algorithm is introduced to approximate this computation when key quantities of the model are unknown, and can only be approximated by an unbiased estimator. We show the interest of our method in the context of partially observed diffusion processes.

2 Model and objectives

Let $\Theta \subset \mathbb{R}^q$ be a compact parameter space and $(X_t)_{t\geqslant 0}$ be defined as a weak solution to the following SDE in $\mathbb{R}^d$:

$$\mathrm{d}X_t = \alpha_\theta(X_t)\mathrm{d}t + \sigma_\theta(X_t)\mathrm{d}W_t \,, \tag{1}$$

where $\theta \in \Theta$, $(W_t)_{t\geqslant 0}$ is a standard Brownian motion in $\mathbb{R}^{\mathsf{d}}$, $\alpha_{\theta_*} : \mathbb{R}^d \to \mathbb{R}^d$ is the drift function and $\sigma_{\theta_*} : \mathbb{R}^d \to \mathbb{R}^{d\times \mathsf{d}}$ is the diffusion. It is assumed that the solution to (1) is partially observed at times $t_0 = 0, \ldots, t_n$, for a given $n \geqslant 1$, through an observation process $(Y_k)_{0\leqslant k\leqslant n}$ taking values in $\mathbb{R}^m$. For all $0 \leqslant k \leqslant n$, the distribution of Y_k given $(X_t)_{t\geqslant 0}$ depends on $X_k = X_{t_k}$ only and has density $g_{k;\theta_*}$ with respect to the Lebesgue measure. The distribution of X_0 has density χ with respect to the Lebesgue measure and for all $0 \leqslant k \leqslant n-1$, the conditional distribution of X_{k+1} given $(X_t)_{0\leqslant t\leqslant t_k}$ has density $q_{k+1;\theta_*}(X_k, \cdot)$.

In this setting, common learning objectives are the *state estimation problem*, which aims at recovering the underlying signal X_k at time t_k given the observations $Y_{0:n}$, where $a_{u:v}$ is a short-hand notation for $(a_u, \ldots, a_v)$, and the *parameter inference problem* which aims at approximating

$$\widehat{\theta}_n = \mathrm{argmax}_{\theta\in\Theta}\, \mathsf{L}_n(\theta) \;,$$

where $\mathsf{L}_n(\theta)$ is the likelihood of the observations. When θ is known, the state estimation problem is usually solved by approximating the posterior mean of X_k given the observations $Y_{0:n}$ when the model is driven by the parameter θ. In the context of parameter estimation, note that

$$\mathsf{L}_n(\theta) = \int \chi(x_0) g_{0;\theta}(x_0, Y_0) \prod_{k=0}^{n-1} \mathsf{r}_{k;\theta}(x_k, x_{k+1}) \mathrm{d}x_{0:n},$$

where, for all $0 \leqslant k \leqslant n$ and all $\theta \in \Theta$,

$$\mathsf{r}_{k;\theta}(x_k, x_{k+1}) = q_{k+1;\theta}(x_k, x_{k+1}) g_{k+1;\theta}(x_{k+1}, Y_{k+1}) \;.$$

Expectation Maximization based algorithms are appealing solutions to obtain an estimator of $\hat{\theta}_n$. The pivotal concept of the EM algorithm is that the intermediate quantity defined by

$$\theta \mapsto Q(\theta, \theta') = \mathbb{E}_{\theta'} \left[\sum_{k=0}^{n-1} \log \mathsf{r}_{k;\theta}(X_k, X_{k+1}) \middle| Y_{0:n} \right] \tag{2}$$

may be used as a surrogate for $\mathsf{L}_n(\theta)$ in the maximization procedure, where $\mathbb{E}_{\theta'}$ is the expectation under the joint distribution of the latent states and the observations when the model is parameterized by θ. In the context of HMMs, the gradient of the log-likelihood can also be expressed as an expectation of an additive functional of the hidden states given $Y_{0:n}$ (Cappé et al., 2005, Chapter 10, or Gloaguen et al., 2019 in the context of SDEs).

A key feature here is that all the relevant estimators rely on computing, for some parameters θ and θ':

$$\mathbb{E}_{\theta'} \left[h_{0:n,\theta}(X_{0:n}) | Y_{0:n} \right],$$

where $h_{0:n,\theta}$ is an *additive functional*, *i.e.* satisfying:

$$h_{0:n,\theta} : x_{0:n} \mapsto \sum_{k=0}^{n-1} \tilde{h}_{k;\theta}(x_k, x_{k+1}),$$

where $\tilde{h}_{k;\theta}$ is a functional depending on the estimator.

For any $\theta \in \Theta$, $0 \leqslant k_1 \leqslant k_2 \leqslant n$ and any bounded and measurable function h on $(\mathbb{R}^d)^{k_2-k_1+1}$, define the *joint smoothing distributions* as:

$$\phi_{k_1:k_2|n;\theta}[h] := \mathbb{E}_\theta \left[h(X_{k_1:k_2}) | Y_{0:n} \right] \;. \tag{3}$$

For all $0 \leqslant k \leqslant n$, $\phi_{k;\theta} = \phi_{k:k|k;\theta}$ are the *filtering distributions*.

In the following, θ is dropped from the notations for better clarity when there is no possible confusion. As noted for instance in Cappé et al. (2005), although the objective is to obtain approximation of smoothing distributions, the filtering distribution is crucial as, for additive functionals,

$$\phi_{0:n|n}[h_{0:n}] = \phi_n \left[\mathbf{T}_n[h_{0:n}]\right],$$

where

$$\mathbf{T}_n[h_{0:n}](X_n) = \mathbb{E}\left[h_n(X_{0:n})|X_n, Y_{0:n}\right]. \tag{4}$$

As a key consequence of the additive property, for all $1 \leqslant k \leqslant n$

$$\mathbf{T}_k[h_{0:k}](X_k) = \mathbb{E}\left[\mathbf{T}_{k-1}[h_{0:(k-1)}](X_{k-1}) + \tilde{h}_{k-1}(X_{k-1}, X_k)\middle| X_k, Y_{0:k-1}\right]. \tag{5}$$

However, the exact computation of all these key expectations is not possible in general state spaces.

Using a Sequential Monte Carlo methods, an alternative is to approximate ϕ_n by weighted samples $\{(\omega_n^\ell, \xi_n^\ell)\}_{\ell=1}^N$ to compute recursively, for each $1 \leqslant \ell \leqslant N$ an approximation τ_n^ℓ of $\mathbf{T}_n[h_{0:n}](\xi_n^\ell)$ so that the estimator of $\phi_{0:n|n}[h_{0:n}]$ is defined as

$$\phi_{0:n|n}^N[h_{0:n}] := \sum_{\ell=1}^N \frac{\omega_n^\ell}{\sum_{j=1}^N \omega_n^j} \tau_n^\ell \,. \tag{6}$$

3 Online sequential Monte Carlo smoothing

In the case of POD processes, SMC methods cannot be used straightforwardly as the transition densities q_k, $0 \leqslant k \leqslant n-1$, are unknown. To overcome these issues, following Fearnhead et al. (2008), consider the following assumption. Let $(\mathsf{U}, \mathcal{B}(\mathsf{U}))$ be a general state space.

H1 For all $\theta \in \Theta$ and $k \geqslant 0$, there exists a Markov kernel on $(\mathbb{R}^d \times \mathbb{R}^d, \mathcal{B}(\mathsf{U}))$ with density $\mathbf{K}_{k;\theta}$ with respect to a reference measure μ on $(\mathsf{U}, \mathcal{B}(\mathsf{U}))$ and a positive mapping $\bar{\mathsf{r}}_{k;\theta}$ on $\mathbb{R}^d \times \mathbb{R}^d \times \mathsf{U}$ such that, for all $(x, x') \in \mathbb{R}^d \times \mathbb{R}^d$,

$$\int \mathbf{K}_{k;\theta}(x, x'; z)\bar{\mathsf{r}}_{k;\theta}(x, x'; z)\mu(\mathrm{d}z) = \mathsf{r}_{k;\theta}(x, x') \,.$$

3.1 Filtering

Let $(\xi_0^\ell)_{\ell=1}^N$ be independent and identically distributed according to an instrumental proposal density ρ_0 on $\mathbb{R}^d$ and define the importance weights $\omega_0^\ell := \chi(\xi_0^\ell)/\rho_0(\xi_0^\ell)$, where χ is the density of the distribution of X_0, see Section 2. For any bounded and measurable function f defined on $\mathbb{R}^d$,

$$\phi_0^N[f] := \Omega_0^{-1} \sum_{\ell=1}^N \omega_0^\ell f(\xi_0^\ell) \,, \quad \text{where} \quad \Omega_0 := \sum_{\ell=1}^N \omega_0^\ell \,.$$

is a consistent estimator of $\phi_0[f]$. Then, for all $k \geqslant 1$, once the observation Y_k is available, the weighted particle sample $\{(\omega_{k-1}^\ell, \xi_{k-1}^\ell)\}_{\ell=1}^N$ is transformed into a new weighted particle sample approximating ϕ_k. This update step is carried through in two steps, *selection* and *mutation*, using the auxiliary sampler introduced in Pitt and Shephard (1999). New indices and particles $\{(I_k^\ell, \xi_k^\ell, \zeta_k^\ell)\}_{\ell=1}^N$ are simulated independently from the instrumental distribution with density on $\{1, \ldots, N\} \times \mathbb{R}^d \times \mathsf{U}$:

$$\upsilon_k(\ell, x, z) \propto \omega_{k-1}^\ell \vartheta_{k-1}(\xi_{k-1}^\ell) p_{k-1}(\xi_{k-1}^\ell, x) \times \mathbf{K}_k(\xi_{k-1}^\ell, x; z) \,,$$

where ϑ_{k-1} is an adjustment multiplier weight function and p_{k-1} a Markovian transition density. In practice, this step is performed as follows.

1. Sample I_k^ℓ in $\{1,\dots,N\}$ with probabilities proportional to $\{\omega_{k-1}^j \vartheta_{k-1}(\xi_{k-1}^j)\}_{1\leqslant j\leqslant N}$.
2. Sample ξ_k^ℓ with distribution $p_{k-1}(\xi_{k-1}^{I_k^\ell}, \cdot)$.
3. Sample ζ_k^ℓ with distribution $\mathbf{K}_k(\xi_{k-1}^{I_k^\ell}, \xi_k^\ell; \cdot)$.

For any $\ell \in \{1,\dots,N\}$, ξ_k^ℓ is associated with the importance weight defined by:

$$\omega_k^\ell := \frac{\overline{\mathsf{r}}_{k-1}(\xi_{k-1}^{I_k^\ell}, \xi_k^\ell; \zeta_k^\ell)}{\vartheta_{k-1}(\xi_{k-1}^{I_k^\ell}) p_{k-1}(\xi_{k-1}^{I_k^\ell}, \xi_k^\ell)} \tag{7}$$

to produce the following approximation of $\phi_k[f]$:

$$\phi_k^N[f] := \Omega_k^{-1} \sum_{\ell=1}^N \omega_k^\ell f(\xi_k^\ell) \ , \quad \text{where} \quad \Omega_k := \sum_{\ell=1}^N \omega_k^\ell \ .$$

3.2 Smoothing

In the context of additive functionals, the forward-only smoothing algorithm introduced in Del Moral et al. (2010) proposes a particle approximation of (4) that can be computed *online* using the recursion (5). This algorithm has a computational complexity which grows *quadratically* with the number of particles N. This computational cost can be reduced when the transition density of the hidden states is upper bounded following Olsson et al. (2017) by applying the accept-reject sampling approach proposed in Douc et al. (2011) and illustrated in Dubarry and Le Corff (2011). Following Gloaguen et al. (2019), the backward statistics $\mathbf{T}_{k+1}[h_{0:k+1}](\xi_{k+1}^i)$, where $\mathbf{T}_{k+1}$ is defined in (4), are estimated, for all $1 \leqslant i \leqslant N$, as follows,

$$\tau_{k+1}^i = \frac{1}{\widetilde{N}} \sum_{j=1}^{\widetilde{N}} \left(\tau_k^{J_{k+1}^{(i,j)}} + \tilde{h}_k \left(\xi_k^{J_{k+1}^{(i,j)}}, \xi_{k+1}^i \right) \right) \ ,$$

where $\widetilde{N} \geqslant 1$ is a sample size which is typically small compared to N and where $(J_{k+1}^{(i,j)}, \zeta_{k+1}^{(i,j)})$, $1 \leqslant j \leqslant \widetilde{N}$, are i.i.d. in $\{1,\dots,N\} \times \mathsf{U}$ with distribution

$$\overline{v}_k^i(\ell, z) \propto \omega_k^\ell \overline{\mathsf{r}}_k(\xi_k^\ell, \xi_{k+1}^i; z) \mathbf{K}_k(\xi_k^\ell, \xi_{k+1}^i; z) \ .$$

In Gloaguen et al. (2018), it is assumed that, for all $0 \leqslant k \leqslant n$ and $0 \leqslant i \leqslant N$, there exists an upper bound $\bar{\varepsilon}_k^i$ such that

$$\sup_{\ell,\zeta} \overline{\mathsf{r}}_k(\xi_k^\ell, \xi_{k+1}^i; \zeta) \leqslant \bar{\varepsilon}_k^i \ . \tag{8}$$

Then, for all $(i, z) \in \{1,\dots,N\} \times \mathsf{U}$,

$$\omega_k^\ell \overline{\mathsf{r}}_k(\xi_k^\ell, \xi_{k+1}^i; z) \mathbf{K}_k(\xi_k^\ell, \xi_{k+1}^i; z) \leqslant \bar{\varepsilon}_k \omega_k^\ell \mathbf{K}_k(\xi_k^\ell, \xi_{k+1}^i; z) \ .$$

Therefore, the following accept-reject mechanism algorithm may be used to sample from $\overline{v}_k^i$.

1. A candidate (J^*, ζ^*) is sampled in $\{1,\dots,N\} \times \mathsf{U}$ as follows:
   (a) J^* is sampled with probabilities proportional to $(\omega_k^\ell)_{\ell=1}^N$;
   (b) ζ^* is sampled independently with distribution $\mathbf{K}_k(\xi_k^{J^*}, \xi_{k+1}^i; \zeta^*)$.
2. (J^*, ζ^*) is then accepted with probability $\overline{\mathsf{r}}_k(\xi_k^{J^*}, \xi_{k+1}^i; \zeta^*)/\bar{\varepsilon}_k$ and, upon acceptance,

$$J_{k+1}^{(i,j)} = J^* \ .$$

3.3 Unbiased estimators of the transition densities

The algorithm described above strongly relies on assumption H1. In the context of SDEs, when $g_{k+1;\theta}$ is available explicitly, this boils down to finding an unbiased estimate $\widehat{q}_{k+1;\theta}(x,y;\zeta)$ of $q_{k+1;\theta}(x,y)$ and defining

$$\overline{\mathsf{r}}_{k;\theta}(x,y;\zeta) = \widehat{q}_{k+1;\theta}(x,y;\zeta) g_{k+1;\theta}(x_{k+1}, Y_{k+1}) \; .$$

Andersson et al. (2017) and Fearnhead et al. (2017) proposed such an estimator which can be used under generic assumptions, much less restrictive than the unbiased estimator of Fearnhead et al. (2008). A key advantage of this method is that it can be computed using only first derivatives of α_θ and second derivatives of σ_θ, and using a traditional Euler scheme.

The stability of this estimator is studied in Fearnhead et al. (2017) which provides L_p controls for the weight $\mathsf{w}_{s_{N_k}}$. The resulting parametrix algorithm is a highly flexible procedure to obtain such an unbiased estimate for a much broader class of diffusions than Poisson based estimations which require strong assumptions. However, the parametrix estimator of the transition density may be negative, and has no reason to satisfy (8). Thus, the SMC algorithms described above cannot be implemented.

4 Backward importance sampling for PODs

4.1 Positive parametrix estimates

Following Fearnhead et al. (2010), Walds identity for martingales may be used to obtain a new estimator from the parametrix approach, which is guaranteed to be positive. This estimator is defined up to an unknown constant of proportionality, which is removed when the importance weights are normalized in equation (9). This approach, uses extra simulation to obtain positiveness. This is done while ensuring that the weights remain unbiased up to a common constant of proportionality. Assume that the distribution $\mathbf{K}_k$ of the additional random variables ζ_k and the estimator $\overline{\mathsf{r}}_k$ are obtained with the parametrix estimator.

Particle filtering. For all $k \geqslant 0$, the Wald-based random weight particle filtering proceeds as follows.

1. For all $1 \leqslant i \leqslant N$, sample a new particle as described in Section 3.1.
   (a) Sample I_k^i in $\{1,\dots,N\}$ with probabilities proportional to $\{\omega_{k-1}^j \vartheta_{k-1}(\xi_{k-1}^j)\}_{1\leqslant j\leqslant N}$.
   (b) Sample ξ_k^i with distribution $p_{k-1}(\xi_{k-1}^{I_k^i}, \cdot)$.
2. For all $1 \leqslant i \leqslant N$, set $\omega_k^i = 0$.
3. While there exists $i_* \in \{1,\dots,N\}$ such that $\omega_k^{i_*} \leqslant 0$, for all $1 \leqslant i \leqslant N$, sample ζ_k^i with distribution $\mathbf{K}_k(\xi_{k-1}^{I_k^i}, \xi_k^i; \cdot)$ (i.e. compute a parametrix estimator of the transition density) and set

$$\omega_k^i = \omega_k^i + \frac{\overline{\mathsf{r}}_{k-1}(\xi_{k-1}^{I_k^i}, \xi_k^i; \zeta_k^i)}{\vartheta_{k-1}(\xi_{k-1}^{I_k^i}) p_{k-1}(\xi_{k-1}^{I_k^i}, \xi_k^i)} \; .$$

Backward simulation. For all $1 \leqslant i \leqslant N$, the backward importance sampling step proceeds then as follows.

1. For all $1 \leqslant j \leqslant \widetilde{N}$, sample $J_{k+1}^{(i,j)}$ in $\{1,\dots,N\}$ with probabilities proportional to $(\omega_k^i)_{i=1}^N$.
2. For all $1 \leqslant j \leqslant \widetilde{N}$, set $\varpi_k^{(i,j)} = 0$.

3. While there exist $j_* \in \{1, \dots, \widetilde{N}\}$ such that $\varpi_k^{(i,j)} \leqslant 0$, for all $1 \leqslant j \leqslant \widetilde{N}$, sample $\zeta_k^{(i,j)}$ with distribution $\mathbf{K}_k(\xi_k^{J_{k+1}^{(i,j)}}, \xi_{k+1}^i; \cdot)$ and set

$$\varpi_k^{(i,j)} = \varpi_k^{(i,j)} + \bar{\mathsf{r}}_k(\xi_k^{J_{k+1}^{(i,j)}}, \xi_{k+1}^i; \zeta_k^{(i,j)}) \; .$$

4.2 AR-free online smoothing

As the positive parametrix-based estimate does not satisfy the upper bound condition of (8), the statistics are updated recursively with an importance sampling step: for all $1 \leqslant i \leqslant N$,

$$\tau_{k+1}^i = \sum_{j=1}^{\widetilde{N}} \frac{\varpi_k^{(i,j)}}{\mathcal{W}_k^i} \left(\tau_k^{J_{k+1}^{(i,j)}} + \tilde{h}_k \left(\xi_k^{J_{k+1}^{(i,j)}}, \xi_{k+1}^i \right) \right) \; , \tag{9}$$

where $\varpi_k^{(i,j)}$, $1 \leqslant j \leqslant \widetilde{N}$ are computed using the parametrix estimate combined with Wald's identity. Then, the estimator of the conditional expectation of the additive functional is set as

$$\phi_{0:n|n}^{N,\mathrm{IS}}[h_{0:n}] := \sum_{i=1}^{N} \frac{\omega_n^i}{\Omega_n} \tau_n^i \; .$$

This estimator does not rely on an accept reject mechanism and is therefore less computationally intensive and can be used under reasonable assumptions for many SDEs. In addition, as shown in Section **??**, this does not affect the statistical efficiency of the algorithm.

5 Discussion

This paper proposes a solution to overcome the two main challenges when it comes to perform online smoothing for generic SDEs i.e. obtaining a positive and almost surely bounded estimate of the transition density to run the backward acceptance rejection mechanism.

1. Note that the proposed backward importance sampling may be used to approximate expectations under the smoothing distributions for general state space hidden Markov models and is not restricted to POD processes. This approach leads to significant gains in computational time for similar performance as the acceptance rejection approach.

2. The proposed estimator, unlike the existing methods such as GPE-based algorithms, applies to a large range of multivariate diffusion processes.

Performances of the algorithm were shown on an extensive numerical simulation settings.

References

Andersson, P., Kohatsu-Higa, A., et al. (2017). Unbiased simulation of stochastic differential equations using parametrix expansions. *Bernoulli*, 23(3):2028–2057.

Cappé, O., Moulines, E., and Rydén, T. (2005). *Inference in Hidden Markov Models*. Springer.

Del Moral, P., Doucet, A., and Singh, S. S. (2010). A backward particle interpretation of feynman-kac formulae. *ESAIM: Mathematical Modelling and Numerical Analysis*, 44(5):947–975.

Douc, R., Garivier, A., Moulines, E., and Olsson, J. (2011). Sequential monte carlo smoothing for general state space hidden markov models. *The Annals of Applied Probability*, 21(6):2109–2145.

Dubarry, C. and Le Corff, S. (2011). Fast computation of smoothed additive functionals in general state-space models. *2011 IEEE Statistical Signal Processing Workshop (SSP)*, pages 197–200.

Fearnhead, P., Latuszynski, K., Roberts, G. O., and Sermaidis, G. (2017). Continuous-time importance sampling: Monte carlo methods which avoid time-discretisation error. *arXiv preprint arXiv:1712.06201*.

Fearnhead, P., Papaspiliopoulos, O., Roberts, G., and Stuart, A. (2010). Random weight particle filtering of continuous time stochastic processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):497–512.

Fearnhead, P., Papaspiliopoulos, O., and Roberts, G. O. (2008). Particle filters for partially observed diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(4):755–777.

Gloaguen, P., Etienne, M.-P., and Le Corff, S. (2018). Online sequential monte carlo smoother for partially observed diffusion processes. *EURASIP Journal on Advances in Signal Processing*, 2018(1):9.

Gloaguen, P., Le Corff, S., and Olsson, J. (2019). Pseudo marginal sequential monte carlo methods for general state spaces. applications to recursive maximum likelihood. *arXiv*, (-):–.

Olsson, J., Westerborn, J., et al. (2017). Efficient particle-based online smoothing in general hidden markov models: the paris algorithm. *Bernoulli*, 23(3):1951–1996.

Pitt, M. K. and Shephard, N. (1999). Filtering via simulation: Auxiliary particle filters. *Journal of the American statistical association*, 94(446):590–599.

Algorithme de Newton stochastique pour l'estimation des paramètres de la régression logistique

Bernard Bercu [1] & Antoine Godichon-Baggioni [2] & Bruno Portier [3]

[1] *Institut de Mathématiques de Bordeaux, UMR 5251, 351 cours de la libération, 33405 Talence cedex, France, bernard.bercu@u-bordeaux.fr*

[2] *Laboratoire de Probabilités, Statistique et Modélisation, UMR 8001, 4 Place Jussieu, 75005 Paris, France, antoine.godichon_baggioni@upmc.fr*

[3] *Laboratoire de Mathématiques de l'INSA, place Emile Blondel, 76131 Mont-Saint-Aignan cedex, France, bruno.portier@insa-rouen.fr*

Résumé. La régression logistique est souvent utilisée comme modèle lorsque l'on doit traiter des variables binaires. Elle a de nombreuses applications en machine learning, sciences sociales, économétrie... Afin d'estimer les paramètres inconnus de la régression logistique lorsque les données arrivent de manière séquentielle, on se concentre sur un algorithme stochastique. Plus précisément, on s'intéresse au comportement asymptotique d'un nouvel algorithme de Newton stochastique. Celui-ci permet de mettre facilement à jour les estimateurs et d'avoir des pas adaptés à toutes les directions. On établit la convergence presque sûre ainsi que la normalité asymptotique des estimateurs ainsi obtenus.

Mots-clés. Algorithme de Newton stochastique, régression logistique, optimisation en ligne

Abstract. Logistic regression is a common model used when the output is a binary random variable. It has a wide range of applications including machine learning, social sciences, econometry... In order to estimate the unknown parameters of logistic regression with data streams arriving sequentially, we focus our attention on a recursive stochastic algorithm. More precisely, we investigate the asymptotic behavior of a new stochastic Newton algorithm. It enables to easily update the estimates when the data arrive sequentially and to have research steps in all directions. We establish the almost sure convergence of our stochastic Newton algorithm as well as its asymptotic normality.

Keywords. Stochastic Newton algorithm, logistic regression, online optimization

1 Introduction

La régression logistique est souvent utilisée comme modèle lorsque l'on doit traiter des variables binaires qui a de nombreuses application en machine learning, sciences sociales,

économétrie... Dans ce qui suit, on considère une suite (X_n, Y_n) de couples de variables aléatoires à valeurs dans $\mathbf{R}^d \times \{0,1\}$, et on suppose que les X_n sont indépendants et identiquement distribués. On suppose également que la loi de Y_n sachant X_n est une loi de Bernoulli. Plus précisément, on considère $\theta = (\theta_0, \ldots, \theta_p)^T \in \mathbf{R}^{p+1}$, on pose $\Phi_n = \left(1, X_n^T\right)^T$ et on suppose

$$\mathcal{L}\left(Y_n|\Phi_n\right) = \mathcal{B}\left(\pi\left(\theta^T\Phi_n\right)\right) \qquad \text{avec} \qquad \pi(x) = \frac{\exp(x)}{1+\exp(x)}.$$

L'objectif est donc d'estimer θ. Pour cela, on considère la fonction convexe G définie pour tout $h \in \mathbf{R}^{p+1}$ par

$$G(h) = \mathbf{E}\left[\log\left(1+\exp\left(h^T\Phi\right)\right) - h^T\Phi Y\right]$$

où $\mathcal{L}\left(Y|\Phi\right) = \mathcal{B}\left(\pi\left(\theta^T\Phi\right)\right)$ et Φ à la même loi que Φ_1. On peut alors vérifier que θ est un zéro du gradient de G, i.e

$$\nabla G(\theta) = \mathbf{E}\left[\left(\pi\left(\theta^T\Phi\right) - Y\right)\Phi\right] = 0. \tag{1}$$

Sous des hypothèses usuelles de convexité sur la fonction G, θ est alors l'unique minimiseur de G. Comme il n'existe pas de solution explicite de l'équation (1), il est nécessaire d'utiliser des algorithmes d'optimisation pour estimer θ. Générallement, lorsque la taille d'échantillon est fixée, on cherche à approcher le minimiseur de la fonction empirique générée par l'échantillon via des algorithmes d'otpimisation usuels (gradient, Newton,...). Cependant, lorsque les données arrivent de manière séquentielle, il peut être plus approprié et efficace de considérer des méthodes en ligne tels que l'algorithme du gradient stochastique et sa version moyennée (Robbins et Monro (1951), Polyak et Juditsky (1991), Gadat et Panloup (2017), Godichon-Baggioni (2019)). Cependant, ce type d'algorithmes implique de prendre le même pas pour toutes les directions, et peut conduire à de mauvaises estimations lorsque la matrice Hessienne de la fonction que l'on cherche à minimiser a des valeurs propres à des échelles significativement différentes. On va donc sintéresser à un algorithme de type Newton stochastique de la forme

$$\theta_{n+1} = \theta - \frac{1}{n+1}\overline{S}_n^{-1}\left(\pi\left(\theta_n^T\Phi_{n+1}\right) - Y_{n+1}\right)\Phi_{n+1}$$

où $\overline{S}_n^{-1}$ est un estimateur recursif de $H^{-1} = \left(\nabla^2 G(\theta)\right)^{-1}$. De plus, comme l'inversion de matrice peut s'avérer coûteuse en terme de temps de calculs, on construit S_n^{-1} de manière récursive à l'aide d'une formule d'inversion de Riccati (Duflo (1996)), aussi appelée formule d'inversion de Sherman-Morrison. Sous certaines hypothèses, on peut montrer que les estimateurs obtenus sont asyptotiquement efficaces.

2 L'algorithme

Dans ce qui suit, on suppose que les hypothèses suivantes sont vérifiées.

(A1) Le vecteur Φ admet un moment d'ordre 2 et la matrice $\mathbf{E}\left[\Phi\Phi^T\right]$ est définie positive.

(A2) La matrice Hessienne $H = \nabla^2 G(\theta)$ est positive.

Ces hypothèses assurent la stricte convexité de la fonction G et donc que θ est l'unique minimiseur de celle-ci. De plus, ces hypothèses assurent que la fonction G est deux fois continuement différentiable et pour tout $h \in \mathbf{R}^{p+1}$,

$$\nabla G(h) = \mathbf{E}\left[\pi\left(h^T\Phi\right)\Phi\right] - \mathbf{E}\left[Y\Phi\right],$$
$$\nabla^2 G(h) = \mathbf{E}\left[\pi\left(h^T\Phi\right)\left(1-\pi\left(h^T\Phi\right)\right)\Phi\Phi^T\right].$$

Lorsque les données arrivent de manière séquentielle, on considère alors l'algorithme de Newton stochastique défini de manière récursive par

$$\hat{a}_{n+1} = \pi\left(\theta_n^T\Phi_{n+1}\right)\left(1-\pi\left(\theta_n^T\Phi_{n+1}\right)\right)$$
$$\theta_{n+1} = \theta_n - \frac{1}{n+1}\overline{S}_n^{-1}\Phi_{n+1}\left(\pi\left(\theta_n^T\Phi_{n+1}\right) - Y_{n+1}\right)$$
$$S_{n+1}^{-1} = S_n^{-1} - \alpha_{n+1}\left(1+\alpha_{n+1}\Phi_{n+1}^T S_n^{-1}\Phi_{n+1}\right)^{-1} S_n\Phi_{n+1}\Phi_{n+1}^T S_n^{-1}$$

où θ_0 est borné, $\overline{S}_n^{-1} = (n+1)S_n^{-1}$, S_0 est symétrique et positive (on peut prendre I_{d+1} par exemple), et la suite (α_n) est définie pour tout $n \geq 1$ par

$$\alpha_n = \max\left\{\hat{a}_n, \frac{c_\beta}{n^\beta}\right\}$$

avec $c_\beta > 0$ et $\beta \in (0, 1/2)$. Cet argument de troncature permet de contrôler la plus grande valeurs propre de S_n^{-1} et ainsi obtenir des résultats de convergence pour nos estimateurs. En effet, grâce à la formule de Riccati, on a

$$S_n = \sum_{k=1}^{n} \alpha_k \Phi_k \Phi_k^T + S_0$$

et $\overline{S}_n$ (resp. $\overline{S}_n^{-1}$) est donc un estimateur naturel de H (resp. H^{-1}).

3 Résultats de convergence

Le théorème suivant donne la convergence presque sûre des estimateurs.

Théorème 1 *Supposons que les hypothèses **(A1)** et **(A2)** sont vérifiées. Alors*

$$\theta_n \xrightarrow[n\to+\infty]{p.s} \theta \qquad et \qquad \overline{S}_n \xrightarrow[n\to+\infty]{p.s} H.$$

De plus, pour tout $\delta > 0$,

$$\|\theta_n - \theta\|^2 = o\left(\frac{(\ln n)^{1+\delta}}{n}\right) \quad p.s$$

Enfin, si Φ *admet un moment d'ordre strictement plus grand que* 2,

$$\|\theta_n - \theta\|^2 = O\left(\frac{\ln n}{n}\right) \quad p.s.$$

A noter que le théorème précédent implique notamment la convergence presque sûre de $\overline{S}_n^{-1}$ vers H^{-1}. Le théorème suivant donne les vitesses de convergence presque sûre de l'estimateur de la Hessienne et de son inverse.

Théorème 2 *Supposons que les hypothèses* ***(A1)*** *et* ***(A2)*** *sont vérifiées. Si* Φ *admet un moment d'ordre* 4*, alors*

$$\left\|\overline{S}_n - H\right\|^2 = O\left(\frac{1}{n^{2\beta}}\right) \quad p.s \qquad et \qquad \left\|\overline{S}_n^{-1} - H^{-1}\right\|^2 = O\left(\frac{1}{n^{2\beta}}\right) \quad p.s$$

A noter que l'on n'a pas une vitesse optimale pour les estimateurs de la Hessienne, ce qui est dûe à la troncature. Finalement, le théorème suivant donne l'efficacité asymptotique des estimateurs de Newton stochastiques.

Théorème 3 *Supposons que les hypothèses* ***(A1)*** *et* ***(A2)*** *sont vérifiées. Si* Φ *admet un moment d'ordre* 4*, alors*

$$\sqrt{n}\left(\theta_n - \theta\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, H^{-1}\right).$$

4 Simulations

4.1 Le modèle

On considère un vecteur aléatoire X de $\mathbf{R}^d$ avec $d = 10$ avec des coordonnées indépendantes et suivant une loi uniforme sur $[0, 1]$. On considère le paramètre $\theta = (-9, 0, 3, -9, 4, -9, 15, 0, -7, 1, 0)^T$. Ce modèle est intérréssant car les valeurs propres de H sont à des échelles très différentes (cf Table 1).

0.1239	$2.832\ 10^{-3}$	$2.822\ 10^{-3}$	$2.816\ 10^{-3}$	$2.778\ 10^{-3}$	$2.806\ 10^{-3}$
$2.651\ 10^{-3}$	$2.517\ 10^{-3}$	$2.1567\ 10^{-3}$	$9.012\ 10^{-4}$	$4.422\ 10^{-4}$	

Table 1: Estimation des valeurs propres de H par ordre décroissant.

4.2 Comparaison des différents algorithmes

On compare ici les performances de quatre algorithmes: l'algorithme de Newton stochastique tronqué (TSN), non tronqué (SN), l'algorithme de gradient stochastique (SG) et sa version moyennée (ASG). A noter que les paramètres pour la descente de gradient ont été choisis par validation croisée.

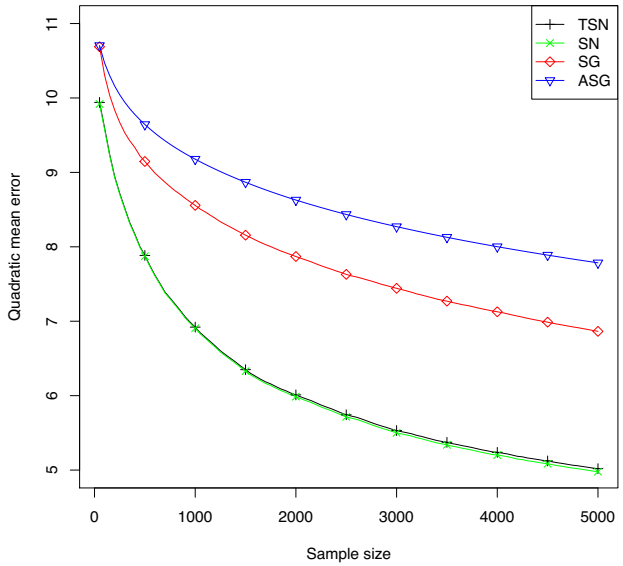


Figure 1: Evolutions de l'erreur quadratique moyenne des différents estimateurs.

La figure 1 montre que les algorithmes de Newton stochastiques se comportent mieux que les algorithmes de type gradient. Les mauvais comportements de ceux-ci sont dûs au fait que les valeurs propres de la Hessienne sont à différentes échelles, et il n'est alors pas judicieux d'avoir le même pas pour chaque direction.

Finalement, Figure 2, on voit que les performances des algorithmes de Newton stochastiques sont proches de celles de la version déterministe (NR).

Bibliographie

Duflo, M. (1996), Algorithmes stochastiques, Springer Berlin
Gadat, S. and Panloup, F. (2017), Optimal non-asymptotic bound of the Ruppert-Polyak averaging without strong convexity, arXiv preprint arXiv:1709.03342
Godichon-Baggioni, A. (2019), Lp and almost sure rates of convergence of averaged

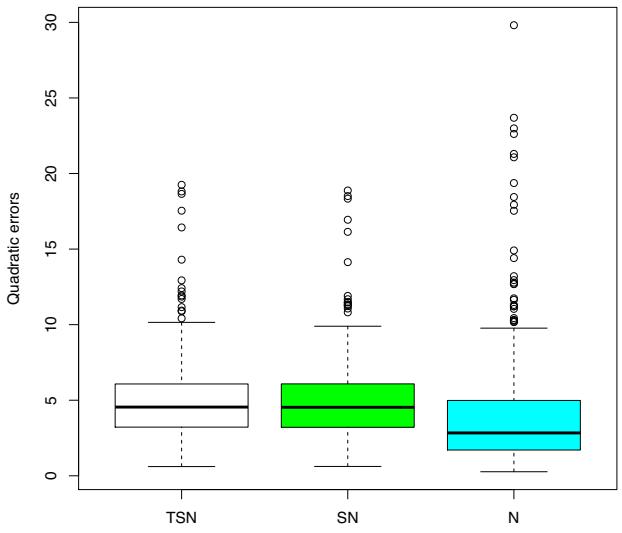


Figure 2: Boxplots des erreurs quadratiques pour les algorithmes TSN, SN et NR.

stochastic gradient algorithms: locally strongly convex objective, ESAIM: PS, 23 841–873,
Polyak, B. and Juditsky, A. (1992), Acceleration of stochastic approximation by averaging, SIAM journal on control and optimization, 30.4 838–855,
Robbins, H. and Monro, S. (1951), A stochastic approximation method, The annals of mathematical statistics, 400–407

Lissage de données fonctionnelles par estimation de leur régularité locale

Steven Golovkine[1] & Nicolas Klutchnikoff[2] & Valentin Patilea[3]

[1]*Renault, CREST, steven.s.golovkine@renault.com*
[2]*Université Rennes 2, IRMAR, nicolas.klutchnikoff@univ-rennes2.fr*
[3]*ENSAI, CREST, valentin.patilea@ensai.fr*

Résumé. Avec les récentes avancées technologiques, de plus en plus d'objets sont équipés de capteurs leur permettant, par exemple, de connaître la position d'autres objets de leur environnement. Ces capteurs fournissent un grand nombre de signaux pouvant être modélisés comme des données fonctionnelles entachées d'un bruit. Dans ce travail, nous supposons que ces données sont enregistrées avec un bruit d'échelle inconnue. Nous nous intéressons donc à l'estimation adaptative du signal grâce à une estimation ponctuelle de la régularité des fonctions sous-jacentes.

Mots-clés. Données fonctionnelles, Estimation non-paramétrique, Régularité

Abstract. With recent technological advances, more and more objects are equipped with sensors that allow them, for example, to know the position of other objects in their environment. These sensors provide a large amount of data that can be modelled as functional data. We assume that these data are recorded with a noise of unknown scale. In this work, we are therefore interested in estimating the signal of interest using an estimation of the smoothness of the underlying functions.

Keywords. Functional data, Non-parametric estimation, Smoothness

1 Introduction

Les capteurs sont de plus en plus présents dans notre vie quotidienne. Ceux-ci fournissent un grand nombre de données pouvant être modélisées comme données fonctionnelles. Comme ces capteurs ne sont pas parfait, il est raisonable de supposer que les données enregistrées le soient avec un certain bruit.

Supposons un échantillon de N courbes provenant d'un même processus aléatoire et éventuellement mesurées à des instants différents. De plus, ces courbes sont détériorées par un bruit aléatoire. Notre but est de définir une procédure, basée sur un sous-échantillon de N_0 courbes bruitées, permettant d'estimer la régularité ponctuelle de l'ensemble de courbes et de permettre d'estimer chacune des $N_1 = N - N_0$ courbes restantes. Pour les différentes applications, N_1 peut être beaucoup plus grand que N_0.

2 Modèle

Soit I un interval compact de $\mathbb{R}$. Considérons $X^{(1)}, \ldots, X^{(N)}$, des réalisations indépendantes du processus stochastique $X = (X_t : t \in I)$ ayant des trajectoires continues. Pour chaque $1 \leq n \leq N$, soit M_n un entier positif et soit $T_m^{(n)}, 1 \leq m \leq M_n$, les temps d'échantillonnage aléatoires de la courbe $X^{(n)}$. Ces temps sont obtenus comme des réalisations indépendantes d'une variable aléatoire T prenant ses valeurs dans I. Les entiers $M_1, \ldots, M_N$ sont des réalisations indépendantes d'une variable aléatoire discrète M de moyenne μ. Nous supposons que les réalisations de X, M et T sont mutuellement indépendantes. Ainsi, nous observons les paires $\left(Y_m^{(n)}, T_m^{(n)}\right) \in \mathbb{R} \times I$ avec $Y_m^{(n)}$ défini comme

$$Y_m^{(n)} = X^{(n)}(T_m^{(n)}) + \varepsilon_m^{(n)}, \qquad 1 \leq m \leq M_n, \quad 1 \leq n \leq N,$$

et les $\varepsilon_m^{(n)}$ sont des réalisations indépendantes de ε, une variable aléatoire de moyenne nulle et de variance σ^2.

Soit $t_0 \in I$ fixé. Nous cherchons à estimer $X^{(N_0+1)}(t_0), \ldots, X^{(N)}(t_0)$. Le cadre est un problème de régression non-paramétrique dépendant de la régularité de la trajectoire du processus X. Nous proposons une procedure adaptive pour l'estimation de la régularité des trajectoires générées par X à un point donné utilisant une partie N_0 des courbes. À partir de cette estimation de la régularité, un lissage *optimal* est réalisé sur le reste des $N - N_0$ courbes générées par le même processus stochastique.

Soit $H(\cdot) : I \mapsto (0, 1]$ et $L(\cdot) : I \mapsto (0, \infty)$, deux fonctions hölderiennes et notons $L_{t_0} = L(t_0)$ et $H_{t_0} = H(t_0)$. Définissons le voisinage de t_0 de la façon suivante:

$$J_\mu(t_0) = (t_0 - |I|/\log\mu, t_0 + |I|/\log\mu) \cap I,$$

avec $|I|$ la longueur de l'intervalle I.

Nous faisons l'hypothèse que le processus stochastique X vérifie la condition:

$$\mathbb{E}\left[(X_u - X_v)^2\right] \asymp L_{t_0}^2 |u - v|^{2H_{t_0}}, \quad u, v \in J_\mu(t_0). \tag{1}$$

La quantité H_{t_0} correspond à la *régularité locale du processus X au point t_0*. Dans le cas, où les trajectoires de X admettent une dérivée d'ordre $s \geq 1$ presque surement, cette définition de régularité locale est utilisée sur la dérivée d'ordre s de la réalisation du processus. Notre intéret premier est d'estimer H_{t_0}. Une problématique similaire a été considéré par Blanke et Vial (2014) dans le cas d'une seule trajectoire observée sans bruit, sur un pas régulier et sous hypothèse d'un processus gaussien.

L'intéret final de ce travail est le débruitage des N_1 courbes restantes. Ainsi, pour tout estimateur $\widehat{X}$ de X, définissons le risque associé à cette estimateur par :

$$\mathcal{R}(\widehat{X}) = \mathcal{R}(\widehat{X}; \mu, N_0, N_1) = \mathbb{E}\left[\max_{1 \leq n_1 \leq N_1} \left|\widehat{X}^{(N_0+n_1)} - X^{(N_0+n_1)}\right|^2\right]. \tag{2}$$

L'optimalité à laquelle nous faisons références dans ce travail est celle définie par rapport à ce risque. Pour des problèmes spécifiques, sous des conditions plus particulières, il est possible d'en déduire d'autres choix de paramètres de lissage. Voir, par exemple, [Carroll *et al.* (2013)].

3 Estimation locale de la régularité

Dans cette partie, nous présentons l'estimateur local de la régularité. Soit $K_0 \in \mathbb{N}$ et considérons un échantillon $(T_1, \ldots, T_M)$ de taille M et de loi T. Nous extrayons de cet échantillon le sous vecteur des K_0 plus proches valeurs de t_0, $\left(T_{(1)}, \ldots, T_{(K_0)}\right)$. En général, t_0 sera un point intérieur à I, et ainsi, $T_{(1)} \leq t_0 \leq T_{(K_0)}$.

Pour la construction de l'estimateur, nous avons besoin de définir les deux événements suivant :

$$\mathcal{A}_n = \mathcal{A}_n(\mu, N_0) = \left\{ M_n \geq K_0, T_{(1)}^{(n)} \in J_\mu(t_0), \ldots, T_{(K_0)}^{(n)} \in J_\mu(t_0) \right\}, \quad 1 \leq n \leq N_0,$$

$$\mathcal{B} = \{ M \geq K_0, T_{(1)} \in J_\mu(t_0), \ldots, T_{(K_0)} \in J_\mu(t_0) \}$$

et notons $\mathbf{1}_{\mathcal{A}_n}$ et $\mathbf{1}_{\mathcal{B}}$, les indicatrices associées. Soit $\mathbb{E}_{\mathcal{B}}(\cdot) = \mathbb{E}(\cdot \mathbf{1}_{\mathcal{B}})$. En utilisant (1), pour $1 \leq k \leq l \leq K_0$, nous avons

$$\mathbb{E}_{\mathcal{B}} \left[(X_{T_{(l)}} - X_{T_{(k)}})^2 \right] \asymp L_{t_0}^2 \mathbb{E}_{\mathcal{B}} |T_{(l)} - T_{(k)}|^{2H_{t_0}}.$$

Définissons, pour k tel que $8k - 7 \leq K_0$, la quantité suivante:

$$\widehat{\theta}_k = \frac{1}{N_0} \sum_{n=1}^{N_0} \left[Y_{(2k-1)}^{(n)} - Y_{(k)}^{(n)} \right]^2 \mathbf{1}_{\mathcal{A}_n}.$$

Un estimateur de la régularité locale H_{t_0} est, dans le cas où σ^2 est connu,

$$\widehat{H}_{t_0}(k, \sigma^2) = \begin{cases} \frac{\log(\widehat{\theta}_{2k-1} - 2\sigma^2) - \log(\widehat{\theta}_k - 2\sigma^2)}{2 \log 2} & \text{si } \widehat{\theta}_{2k-1} > 2\sigma^2 \text{ et } \widehat{\theta}_k > 2\sigma^2 \\ 0 & \text{sinon} \end{cases},$$

et, dans le cas où σ^2 est inconnu,

$$\widehat{H}_{t_0}(k) = \begin{cases} \frac{\log(\widehat{\theta}_{4k-3} - \widehat{\theta}_{2k-1}) - \log(\widehat{\theta}_{2k-1} - \widehat{\theta}_k)}{2 \log 2} & \text{si } \widehat{\theta}_{4k-3} > \widehat{\theta}_{2k-1} > \widehat{\theta}_k \\ 0 & \text{sinon} \end{cases}.$$

En particulier, sous certaines conditions sur le processus X, sur la densité de T, sur les moments du bruit ε, sur la concentration de M autour de sa moyenne et sur K_0 bien choisi comme fonction de μ, nous montrons que, pour μ suffisamment grand,

$$\mathbb{P}\left(|\widehat{H}_{t_0}(k) - H_{t_0}| > \log^{-2}(\mu) \right) \leq \frac{1}{\mu}.$$

Le résultat non-asymptotique sur la régularité locale de X permet de construire une fenêtre (quasiment) optimale pour le lissage des $N - N_0$ courbes.

4 Estimation de la fenêtre de lissage

Le lissage des N_1 se fait par régression à noyaux. Il nous manque donc une estimation de la fenêtre de lissage $b_n,\ n = N_0 + 1, \ldots, N$. Ainsi, cette fenêtre sera différente pour chaque courbe à débruiter.

Un estimateur possible pour b_n est le suivant :

$$b_n = \left(\frac{\sigma^2 \|K\|^2}{\frac{H_{t_0}}{[H_{t_0}]!} L_{t_0} \int |K(v)||v|^{H_{t_0}} dv} \times \frac{1}{M_n} \right)^{\frac{1}{2H_{t_0}+1}} \tag{3}$$

où $K : \mathbb{R} \to \mathbb{R}$ est un noyau, que l'on peut choisir comme étant celui d'Epanechnikov par exemple et $[H_{t_0}]$ est la partie entière de H_{t_0}.

Dans cette formule, les quantités suivantes sont à estimer : σ^2, H_{t_0} and L_{t_0}. Le choix du noyau est laissé à l'utilisateur.

- **Estimateur de σ^2 :**

  Nous proposons l'estimateur suivant pour σ^2 :

$$\widehat{\sigma^2} = \frac{1}{N_0} \sum_{n=1}^{N_0} \frac{1}{2(M_n - 1)} \sum_{l=2}^{M_n} \left(Y_{n,(l)} - Y_{n,(l-1)} \right)^2$$

- **Estimateur de H_{t_0} :**

  L'estimateur de H_{t_0} est développé à la section précédente.

- **Estimateur de L_{t_0} :**

  Différents estimateur de L_{t_0} sont possible suivant notre connaissance de la densité des points d'échantillonnage et de σ^2. Considérons uniquement le cas où l'on ne connait ni l'un ni l'autre, ce qui est usuellement le cas.

  Définissons, pour k tel que $8k - 7 \leq K_0$, la quantité suivante:

$$\widehat{\eta}_k = \frac{1}{N_0} \sum_{n=1}^{N_0} \left| T_{n,(2k-1)} - T_{n,(k)} \right|^{2H_{t_0}}. \tag{4}$$

  Un estimateur de la constante L_{t_0} est

$$\widehat{L}_{t_0}(k) = \left(\frac{\widehat{\theta}_{2k-1} - \widehat{\theta}_k}{\widehat{\eta}_{2k-1} - \widehat{\eta}_k} \right)^{1/2}.$$

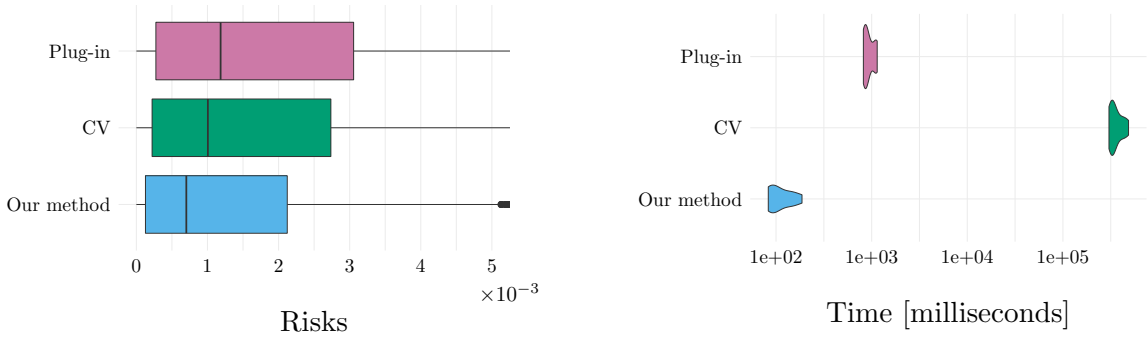
Ainsi, un estimateur de la fenêtre b_n est donné par :

$$\widehat{b}_{n,t_0}(k) = \left(\frac{\widehat{\sigma^2}\|K\|^2}{\frac{\widehat{H}_{t_0}(k)}{[\widehat{H}_{t_0}(k)]!}\widehat{L}_{t_0}(k)\int |K(v)||v|^{\widehat{H}_{t_0}(k)}dv} \times \frac{1}{M_n} \right)^{\frac{1}{2\widehat{H}_{t_0}(k)+1}}.$$

Finalement, le lissage des courbes est réalisé en utilisant l'estimateur de Nadaraya-Watson:

$$\widehat{X}^{(n)}(t) = \frac{\sum_{m=1}^{M_n} K((t - T_m^{(n)})/\widehat{b}_{n,t}(k))Y_m^{(n)}}{\sum_{m=1}^{M_n} K((t - T_m^{(n)})/\widehat{b}_{n,t}(k))}, \quad \forall t \in [0,1].$$

Nous étudions la qualité de cet estimateur de la fenêtre de lissage par rapport à deux estimateurs classiques de celle-ci que sont la validation croisée et l'estimateur *plug-in* [Ruppert *et al.* (1995)] en considérant le risque définit en (2). De plus, une expérimentation empirique du temps d'exécution d'un échantillon de $N = 1000$ courbes est développée pour chaque méthode. La figure 1 présente ces résultats dans le cas où X est un mouvement brownien, et donc dans le cas où la régularité locale du processus est de $H_{t_0} = 0.5$ pour tout point t_0.



(a) Risque ponctuel en $t_0 = 0.5$ (b) Temps de calcul (log scale)

Figure 1: Résultats obtenus dans le cas où X est un mouvement brownien

La méthode peut s'étendre aux cas où la régularité est supérieur à 1 en estimant la régularité des dérivées successives. Enfin, le cadre peut aussi s'élargir aux cas où la variance du bruit n'est plus supposée constante mais d'esperance contionnnelle constante, ce qui permet de considérer un bruit hétéroscédastique.

Bibliographie

Blanke, D., Vial, C. (2014). *Global Smoothness Estimation of a Gaussian Process from General Sequence Designs*, *Electronic Journal of Statistics*, Vol. 8, pp. 1152–1187.

Carroll, R.J., Delaigle, A., Hall, P. (2013). *Unexpected properties of bandwidth choice when smoothing discrete data for constructing a functional data classifier*, *Ann. Statist.*, Vol. 41, no. 6, 2739–2767.
Ruppert, D., Sheather, S. J. and Wand, M. P. (1995). *An Effective Bandwidth Selector for Local Least Squares Regression*, *Journal of the American Statistical Association*, Vol. 90, no. 432, pp. 1257–1270.
Tsybakov, A. (2009). *Introduction to Nonparametric Estimation*, Springer, New York.

Forecasting high resolution electricity demand data with additive models including smooth and jagged components

Umberto Amato[1], Anestis.Antoniadis[2], Italia De Feis[3], Yannig Goude[4] & Audrey Pichavant[5]

[1]*Institute of Applied Sciences and Intelligent Systems "E. Caianiello",* umberto.amato@cnr.it

[2]*LJK, Department of Statistics, University Grenoble Alpes, France,* Anestis.Antoniadis@univ-grenoble-alpes.fr

[3] *Institute for Applications of Computing "M. Picone", Italian National Research Council, Napoli, Italy, Email: i.defeis@iac.cnr.it*

[4]*DF, OSIRIS, 7 bd Gaspard Monge, 91120 Palaiseau, France, Université Paris Sud, France,* yannig.goude@edf.fr

[4]*EDF, OSIRIS, 7 bd Gaspard Monge, 91120 Palaiseau, France,* audrey.pichavant@edf.fr

Résumé. La prévision à court terme de consommation électrique est une entrée essentielle pour l'optimisation des systèmes électriques. Le développement des smarts grids induit de nouveaux challenges et opportunités, notamment la question de la prévision à des niveaux faibles d'agrégation. Les modèles GAM sont très utilisés dans ce domaine. Ils permettent d'obtenir de bons niveaux de performances tout en étant facilement interprétables par les prévisionnistes. Néanmoins, ces modèles supposent que les données de consommation soient relativement régulières. A de faibles niveaux d'agrégation, pour modéliser le caractère très volatile de ces données nous proposons d'introduire des composantes irrégulières dans ces modèles additifs. Nous montrons sur des données réelles l'intérêt de ces modèles hybrides en terme de performance prédictive.

Mots-clés. GAM, ondelettes, prévision de consommation électrique.

Abstract. Short-Term Load Forecasting (STLF) is a fundamental instrument in the efficient operational management and planning of electric utilities. Emerging smart grid technologies pose new challenges and opportunities. Although load forecasting at the aggregate level has been extensively studied, electrical load forecasting at fine-grained geographical scales of households is more challenging. Among existing approaches, semi-parametric generalized additive models (GAM) have been increasingly popular due to their accuracy, flexibility, and interpretability. Their applicability is justified when forecasting is addressed at higher levels of aggregation, since the aggregated load pattern contains relatively smooth additive components. High resolution data are highly volatile, forecasting the average load using GAM models with smooth components does not provide meaningful information about the future demand. Instead, we need to incorporate irregular and volatile effects to enhance the forecast accuracy. We focus on the analysis of

such hybrid additive models applied on smart meters data and show that it leads to improvement of the forecasting performances of classical additive models at low aggregation levels.

Keywords. GAM, wavelets, electricity consumption forecasting.

1 Introduction

Load forecasting is a crucial part of electric power system operations. Forecasting at a local scale is a complex problem. In [8] the authors argue from an extensive study of a dataset from PG&E Northern California region that hour ahead forecasting errors can go from a 29% Mean Absolute Percentage Error (MAPE) for a single individual residential consumer to less that 10% for groups of 100 customers. The study relies on different forecasting models and the authors mention that for more complicated tasks like day ahead forecasting, one needs to think carefully about the model to use. Our work is motivated by these points. We propose a new approach adaptive with time and above all with the level of aggregation.

We focus on partially linear additive models (PLAMs) based on function basis decompositions of their nonlinear additive components (e.g. [6], [9], [1]). These semi-parametric models combine the flexibility of fully nonparametric models and the simplicity of multiple regression models. In many applications the nonlinear components are assumed to be smooth and are therefore approximated by their decompositions into splines bases. In case the components are less regular in terms of smoothness, as it can be for example with low level aggregated data, the splines approximation can be inadequate and one should choose less regular basis functions, an example of which are wavelets.

2 Methodology: estimation and model selection in PLAMs

PLAMs retain the parsimony and interpretability of linear models and the flexibility of nonparametric additive regression, by allowing a linear component for some predictors presumed to have a strictly linear effect, and an additive structure for other predictors, reducing the "curse of dimensionality".

Given observations $\{(Y_i, \mathbf{X}_i, \mathbf{T}_i)\}_{i=1}^n$, where Y_i is the response, $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^T$ and $\mathbf{T}_i = (T_{i1}, \ldots, T_{iq})^T$ are vectors of covariates, the PLAM assumes that

$$Y_i = b + \mathbf{X}_i^T \boldsymbol{\beta} + \sum_{j=1}^{q} f_j(T_{ij}) + \epsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where b is the intercept, $\boldsymbol{\beta}$ is the $p \times 1$ vector of unknown coefficients for linear terms, f_j are unknown nonlinear real valued components and the ϵ_i's are i.i.d random variables with mean 0 and variance σ^2 independent of the covariates. In order to ensure that the model is identifiable, one requires that the linear covariates are centred and that identifiability conditions $\int f_j(t)dt = 0$, $j = 1, \ldots, q$ hold.

2.1 Smooth additive components

We will assume that the unknown nonparametric (smooth) additive components f_j belong to the subspace of centred functions within the Sobolev space of order m, $\mathcal{W}_2^m$ equipped with the Sobolev norm $\|f\|_m = \sqrt{\int \left(f^{(m)}(x)\right)^2 dx}$, where $f^{(m)}$ denotes the mth derivative of f.

Under these smoothness assumptions, the $f_j(t)$ can be well approximated by their expansion on O'Sullivan splines basis functions $\{B_\ell^{(j)}\}_{\ell \in \mathbb{N}}$ introduced by [7]:

$$f_j(t) \approx \sum_{\ell=1}^{m_j} \alpha_\ell^{(j)} B_\ell^{(j)}(t), \ j = 1, \ldots, q, \tag{2}$$

where m_j is an appropriate truncation index allowed to increase to infinity with n.

2.2 Irregular additive components

Let us consider now PLAMs with nonlinear additive components that are less smooth. To capture key characteristics of variations and of inhomogeneity in each f_j, $j = 1, \ldots, q$, and to exploit their sparse wavelet coefficients representations, we will assume that f_j belong to the (inhomogeneous) Besov space on the unit interval $\mathcal{B}^s_{\pi,r}([0,1])$ with $s+1/\pi-1/2 > 0$. The parameter π is a degree measuring the function's inhomogeneity, s is a measure of its smoothness. Roughly speaking, the (not necessarily integer) parameter s indicates the number of function's (fractional) derivatives, where their existence is required in an L^π-sense; the additional parameter r is secondary in its role, allowing for additional fine tuning of the definition of the space (e.g. see [5]).

As with splines basis expansions of smooth functions we approximate the nonparametric additive components using wavelet bases (see [3]). For each f_j, we use its expansion on wavelet basis functions $\{W_\ell^{(j)}\}_\ell$:

$$f_j(t) \approx \sum_{\ell=1}^{K_j} \gamma_\ell^{(j)} W_\ell^{(j)}(t), \tag{3}$$

where K_j is an appropriate truncation index allowed to increase to infinity with n. A function f_j within some Besov ball can be well approximated by the above expansion and its estimation is equivalent to estimate the wavelet coefficient vector $\boldsymbol{\gamma}^{(j)} = (\gamma_1^{(j)}, \ldots, \gamma_{K_j}^{(j)})^T$.

Similarly to the spline case, as alluded to in [3] we can also define the regression design matrices containing wavelet basis functions evaluated at the observations of the corresponding predictors (see [2]).

2.3 Hybrid PLAM

In the previous subsection the additive components were assumed to posses a similar degree of regularity: they were either belonging to Sobolev spaces or to Besov spaces. However, a limitation was the use of a single class of basis functions, either splines or wavelets. A loss of efficiency occurs when the additive part is composed of both smooth functions and functions with much less regularity, a class of models that we are calling hybrid PLAM.

Our main contribution is to combine the methods used in the previous subsections and obtain an estimator that can deal with and overcome the difficulties given from the non-linear part of such hybrid partially additive models.

The basic idea is exploring the advantage of using a hybrid fitting for the additive part combining regression splines and wavelets together, and use again a grouped model selection methods for solving the estimation and variable selection problems.

Hereafter we consider a random sample $\{Y_i, \mathbf{X}_i, \mathbf{T}_i^s, \mathbf{T}_i^w\}_{i=1,\dots,n}$, related through the hybrid partially linear additive model (HPLAM)

$$Y_i = \mathbf{X}_i^T \boldsymbol{\beta} + \sum_{j=1}^{q_s} f_j^s(T_{ij}^s) + \sum_{j=1}^{q_w} f_j^w(T_{ij}^w) + \epsilon_i, \quad i = 1, \dots, n, \tag{4}$$

where, $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})^T$ is the p-dimensional covariate vector representing the linear regression components, $\boldsymbol{\beta}$ is the $p \times 1$ vector of corresponding regression coefficients, f_j^s are unknown smooth functions of T_{ij}^s, where $\mathbf{T}_i^s = (T_{i1}^s, \dots, T_{iq_s}^s)^T$ is a q_s-dimensional nonlinear covariate vector with values in $[0,1]^{q_s}$, f_j^ws are unknown non-smooth functions of T_{ij}^w, where $\mathbf{T}_i^w = (T_{i1}^w, \dots, T_{iq_w}^w)^T$ is a q_w-dimensional nonlinear covariate vector with values in $[0,1]^{q_w}$ and where the errors ϵ_i form a sequence of i.i.d. Gaussian random variables with mean 0 and variance σ^2 independent of the predictor variables $\mathbf{X}_i$, $\mathbf{T}_i^s$ and $\mathbf{T}_i^w$.

3 Forecasting smart meter data at various levels of aggregation

Electricity consumption data at the level of individual households have several distinctive features and a wide variety of shapes. The purpose of this section is to analyse how forecast errors scale with aggregation levels using the semi-parametric forecasting models introduced in the previous sections.

We use the data collected during a smart metering trial conducted by the Commission for Energy Regulation (CER) in Ireland (see [4]). The data set contains half-hourly measurements of electricity consumption gathered from 4623 consumers (residential customers and small-to-medium enterprises) between 14 July 2009 and 31 December 2010.

We model the ith individual electric load Y_i at time t as

$$Y_i(t) = m(i) + \mu_i(t) + g_i(T(t)) + \epsilon_i(t),\ i = 1, .., n,\ t = 1, .., J, \tag{5}$$

where $m(i) = (X\boldsymbol{\beta})_i$ is the average effect of p time-independent variables on the load curve; X is the $n \times p$ design matrix associated to these variables; $\boldsymbol{\beta}$ is the corresponding p-dimensional vector of parameters; $\mu_i(t)$ is a time dependent longitudinal effect modelled by its decomposition in a nonparametric functional basis, including the day-of-week (Dow_t) as exogenous variable, for identifiability reasons $\sum_t \mu_i(t) = 0$. $g_i(T(t))$ is the effect of temperature $T(t)$ on the load profile of the ith client, and again, for identifiability reasons, we assume that $\sum_t g_i(T(t)) = 0$; $\epsilon_i(t)$ is a noise process. We have tested three different models for the combined longitudinal effect $\mu_i(t)+g_i(T(t))$, using two different expressions for $\mu_i(t)$, denoted hereafter $\mu_i^s(t)$ (smooth) and $\mu_i^w(t)$ (wavelet), combined with either a wavelet approximation $g_i^w(T(t))$ or a spline approximation $g_i^s(T(t))$ for the effect of the temperature.

We simulate an online forecasting procedure at different aggregation level and conduct an intensive simulation with random sub-sampling of the population at each level of aggregation. Additionally to the proposed models which can be used for mid term forecasts (up to a week ahead here) we implement a correction term on the errors assuming an autoregressive structure of the noise $\epsilon_i(t)$ in the model of equation (5). We compute the RMSEs (Root Mean Square Error) at each aggregation levels on the last 5 months. On the first 23 days of each of these months we compute the fitting error of the mid term models. On the last 7 days we calculate the forecasting error of the mid term models and the associated short term correction models.

Figure 1 reports the forecasting errors on the testing data, which measure the predictive performance of the estimation methods. We plot with solid line the mean ratio of the mid term models with respects to the RMSE of the smooth model (with splines) and in dashed line the ones of the short term correction models.

Indeed, the spline based model seems to work best across low levels of aggregation (for $L \leq 20$). At higher levels of aggregation (for $L \geq 30$) wavelets and hybrid work better than splines with hybrid model always slightly outperforming wavelets. The improvement goes from around 2 percent to 10 percent in RMSE. Statistical tests confirm these findings.

References

[1] U. Amato, A. Antoniadis, and I. De Feis. Additive model selection. *Statistical Methods & Applications*, 25(4):519–564, Nov 2016.

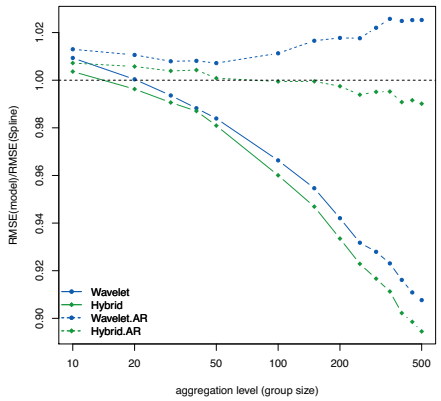


Figure 1: Mean ratio (over the range of 50 groups and 5 months) of RMSE (forecasting set) of models (wavelet in blue, hybrid in green) to RMSE of the spline model for various aggregation levels (logarithmic scale).

[2] U. Amato, A. Antoniadis, I. De Feis, and Y. Goude. Estimation and group variable selection for additive partial linear models with wavelets and splines. *South African Statistical Journal*, 51:235–272, 2017.

[3] A. Antoniadis and J. Fan. Regularization of wavelet approximations. *Journal of the American Statistical Association*, 96(455):939–967, 2001.

[4] Commission for Energy Regulation. Cer smart metering project - electricity customer behaviour trial, 2009-2010 [dataset]. 1st edition. *Irish Social Science Data Archive*, SN: 0012-00, 2012.

[5] D. L. Donoho and I. M. Johnstone. Minimax estimation via wavelet shrinkage. *Ann. Statist.*, 26(3):879–921, 06 1998.

[6] T. Hastie and R. Tibshirani. Generalized additive models. *Statist. Sci.*, 1(3):297–310, 08 1986.

[7] F. O'Sullivan. A statistical perspective on ill-posed inverse problems. *Statistical Science*, 1(4):502–518, 1986.

[8] R. Sevlian and R. Rajagopal. A scaling law for short term load forecasting on varying levels of aggregation. *International Journal of Electrical Power & Energy Systems*, 98:350–361, 06 2018.

[9] S. Wood. *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, 2 edition, 2017.

Apprentissage de modèles CHARME avec des réseaux de neurones profonds

José G. Gómez-García [1] · Jalal Fadili [2] · Christophe Chesneau [3]

[1] *Université Paris-Est Créteil, LAMA, UMR CNRS 8050.*
ESIPE, 71 rue Saint-Simon, 94000 Créteil.
jose-gregorio.gomez-garcia@u-pec.fr

[2] *Normandie Université, ENSICAEN, UNICAEN, GREYC, UMR CNRS 6072.*
ENSICAEN, 6 Bd du Maréchal Juin, 14050 Caen.
Jalal.Fadili@ensicaen.fr

[3] *Normandie Université, UNICAEN, LMNO, UMR CNRS 6139.*
LMNO, Sciences 3, Campus 2, Bd du Maréchal Juin, 14000 Caen.
christophe.chesneau@unicaen.fr

Résumé. Dans cette note, nous considérons un modèle appelé CHARME (Conditional Heteroscedastic Autoregressive Mixture of Experts). En quelques mots, c'est un modèle de mélange généralisé de séries chronologiques non linéaire et non paramétrique AR-ARCH. Nous garantissons la stabilité (ergodicité et stationnarité) du modèle sous certaines conditions de type Lipschitz pour les fonctions d'autorégression et de volatilité, lesquelles sont beaucoup plus faibles que celles présentées dans la littérature existante. Ce résultat et la propriété d'approximation universelle de réseaux de neurones (RN), possiblement avec des architectures profondes (RNP), nous fournit les bases pour développer une théorie d'apprentissage pour les fonctions d'autorégression-basées-sur-RN du modèle. En outre, la consistance forte et la normalité asymptotique de l'estimateur des poids et des biais des RN considéré sont garanties sous de faibles conditions.

Mots-clés. modèle AR-ARCH non-paramétrique ; réseaux de neurones profonds ; modèles de mélange ; séquence à changement de régime markoviens ; dépendance τ-faible ; ergodicité ; stationnarité ; identifiabilité ; consistance ; signaux d'EEG.

Abstract. In this note, we consider a model called CHARME (Conditional Heteroscedastic Autoregressive Mixture of Experts). Roughly speaking, this is a class of generalized mixture of nonlinear nonparametric AR-ARCH time series. We guarantee the stability (ergodicity and stationarity) of the model under certain Lipschitz-type conditions on the autoregression and volatility functions, which are much weaker than those presented in the current literature. This result and the universal approximation property of neural networks (NN), possibly with deep architectures (DNN), provides us with the bases for developing a learning theory for the NN-based autoregressive functions of the model.By the way, the strong consistency and asymptotic normality of the considered estimator of the NN weights and biases are guaranteed under weak conditions.

Keywords. Nonparametric AR-ARCH ; deep neural network ; mixture models ; Markov switching ; τ-weak dependence ; ergodicity, stationarity ; consistency ; EEG signals.

1 Introduction

Dans l'analyse de séries chronologiques, il est commun d'étudier les modèles tels que : AR, ARMA, ARCH, GARCH, etc. ; ou plus généralement, le modèle CHARN

$$X_t = f(X_{t-1}, \ldots, X_{t-p}, \theta^0) + g(X_{t-1}, \ldots, X_{t-p}, \lambda^0)\epsilon_t, \qquad t \in \mathbb{Z}, \tag{1}$$

où f, g sont des fonctions inconnues et $(\epsilon_t)_{t\in\mathbb{Z}}$ est un bruit blanc indépendant. Cependant, dans la pratique, il n'est pas toujours réaliste de supposer que le processus observé ait la même tendance f et la même volatilité g à chaque instant t. Entre autre, c'est le cas des signaux d'EEG, voir Lo *et al.* (2009), où l'on peut observer des changements de comportement, même brusques, lesquelles on ne peut pas les modéliser même en utilisant les modèles localement stationnaires. C'est pour cela que nous nous concentrons sur un modèle plus général, appelé CHARME, qui prend en compte ces changement brusques de comportement.

Pour définir ce modèle, considérons l'espace de Banach $(E, \|\cdot\|)$, doté de sa tribu borélienne $\mathcal{E}$. L'espace produit E^p est alors naturellement doté de sa tribu produit $\mathcal{E}^{\otimes p}$. Le modèle **CHARME(p)**, à valeurs dans E, est la série chronologique définie par

$$X_t = \sum_{k=1}^{K} \xi_t^{(k)} \left(f_k(X_{t-1}, \ldots, X_{t-p}, \theta_k^0) + g_k(X_{t-1}, \ldots, X_{t-p}, \lambda_k^0)\epsilon_t \right) \quad t \in \mathbb{Z}, \tag{2}$$

où

- pour chaque $k \in [K] := \{1, 2, \ldots, K\}$, $f_k : E^p \times \Theta_k \longrightarrow E$ et $g_k : E^p \times \Lambda_k \longrightarrow \mathbb{R}$ sont respectivement les fonctions d'autorégression et de volatilité, avec des espaces de paramètres respectifs Θ_k et Λ_k, $\mathcal{E}^{\otimes p} \times \mathcal{B}(\Theta_k)$- et $\mathcal{E}^{\otimes p} \times \mathcal{B}(\Lambda_k)$- mesurables, où $\mathcal{B}(\Theta_k)$ est la tribu borélienne sur Θ_k et pareillement pour Λ_k ;
- $(\epsilon_t)_t$ est un bruit blanc indépendant à valeurs dans E ;
- $\xi_t^{(k)} = \mathbb{I}_{\{R_t=k\}}$, avec $\mathbb{I}_{\mathcal{C}}$ désignant la fonction caractéristique de $\mathcal{C}$ (*i.e.*, elle vaut 1 sur $\mathcal{C}$ et 0 sinon), où $(R_t)_{t\in\mathbb{Z}}$ est une séquence de variables aléatoires indépendantes à valeurs dans l'espace fini $[K]$, qui est en plus indépendante du bruit blanc $(\epsilon_t)_{t\in\mathbb{Z}}$. Par la suite, on pose $\pi_k = \mathbb{P}(R_0 = k)$.

Le modèle (2) peut être étendu au cas $p = \infty$. Nous l'appellerons alors modèle CHARME à mémoire infinie et que nous désignerons par CHARME(∞). Dans ce cadre, l'espace d'états du modèle est le sous-ensemble de $E^{\mathbb{N}}$:

$$E^{\infty} := \left\{ (x_k)_{k>0} \in E^{\mathbb{N}} : \ x_k = 0 \text{ for } k > N, \text{ for some } N \in \mathbb{N}^* \right\},$$

doté de sa tribu produit $\mathcal{E}^{\otimes\mathbb{N}}$.

Il est clair que le modèle (2) contient le modèle (1) (cela correspond au cas $K = 1$ en (2)). D'ailleurs, des applications de ce modèle ont été traitées d'une manière directe ou indirecte dans plusieurs domaines de recherche. Voir par exemple : Tadjuidje-Kamgaing, J. (2005), Weigend, A.S. and Shi, S. (2000), Kirch, C. and Kamgaing, T. (2012) et Liehr *et al.* (1999).

2 Ergodicité et stationnarité des modèles CHARME

Le résultat suivant nous fournit des conditions pour avoir la stabilité du modèle dont la preuve est donnée dans la Section 8.1 de Gómez-García *et al.* (2020).

Théorème 1. *Considérons le modèle CHARME(∞), i.e.,* (2) *avec* $p = \infty$. *Supposons qu'il existe des séquences non-négatives* $(a_i^{(k)})_{i\geq 1, k\in[K]}$ *et* $(b_i^{(k)})_{i\geq 1, k\in[K]}$ *telles que, pour tout* $x, y \in E^\infty$ *et tout* $k \in [K]$,

$$\|f_k(x,\theta_k^0) - f_k(y,\theta_k^0)\| \leq \sum_{i=1}^{\infty} a_i^{(k)} \|x_i - y_i\|, \quad |g_k(x,\theta_k^0) - g_k(y,\theta_k^0)| \leq \sum_{i=1}^{\infty} b_i^{(k)} \|x_i - y_i\| \quad (3)$$

Notons $A_k = \sum_{i=1}^{\infty} a_i^{(k)}$, $B_k = \sum_{i=1}^{\infty} b_i^{(k)}$ *et* $C(m) = 2^{m-1} \sum_{k=1}^{K} \pi_k \left(A_k^m + B_k^m \|\epsilon_0\|_m^m\right)$. *Alors, nous obtenons les affirmations suivantes :*

(i) si $c := C(1) < 1$, *alors il existe une solution strictement stationnaire* $(X_t)_{t\in\mathbb{Z}}$ *du modèle CHARME(∞) appartenant à* $\mathbb{L}^1$.

(ii) si en plus $C(m) < 1$ *pour certain* $m > 1$, *alors cette solution appartient à* $\mathbb{L}^m$.

Remarque 1.

(1.1) Le résultat précédent est également valable dans le cas $p < \infty$. En effet, il suffit de prendre $a_i^{(k)} = b_i^{(k)} = 0$ pour tout $i > p$ et tout $k \in [K]$ dans les inégalités (3).

(1.2) Remarquons que le modèle CHARME(∞) (2) avec $p = \infty$ peut être réécrit comme une séquence de Markov $X_t = F(X_{t-1}, X_{t-2}, \ldots; \tilde{\xi}_t)$, $t \in \mathbb{Z}$, via la fonction

$$F(x; (\xi^{(0)}, \ldots, \xi^{(K)})) = \sum_{k=1}^{K} \xi^{(k)} \left(f_k(x,\theta_k^0) + g_k(x,\lambda_k^0)\xi^{(0)}\right), \quad (4)$$

avec des innovations $\tilde{\xi}_t := (\epsilon_t, \xi_t^{(1)}, \ldots, \xi_t^{(K)}) = (\epsilon_t, \xi_t) \in E \times B_e$, où $B_e := \{e_1, \ldots, e_K\}$ est la base canonique de $\mathbb{R}^K$. Sous les hypothèses du Théorème 1, la fonction F est continue car les fonctions $f_k(\cdot, \theta_k^0)$ et $g_k(\cdot, \lambda_k^0)$ sont continues par la condition (3). Il découle alors de (Doukhan, P. and Wintenberger, O, 2008, Lemma 5.5) et de la complétude de $\mathbb{L}^m$, qu'il existe une fonction mesurable H telle que le processus CHARME(∞) peut être écrit comme $X_t = H(\tilde{\xi}_t, \tilde{\xi}_{t-1}, \ldots)$. C'est-à-dire : le processus CHARME(∞) peut être représenté par un décalage de Bernoulli causal. En outre, sous ces hypothèses, $(X_t)_{t\in\mathbb{Z}}$ est le seul décalage de Bernouli causal, solution à (2) avec $p = \infty$. Donc, la solution $(X_t)_{t\in\mathbb{Z}}$ est automatiquement un processus ergodique. Enfin, le théorème ergodique implique la LFGN pour ce processus. Cette conséquence du Théorème 1 sera un résultat clé pour établir la consistance forte lorsqu'il s'agit d'estimer les fonctions d'autorégression et de volatilité du modèle CHARME(p).

(1.3) Stockis *et al.* (2010) montre l'ergodicité du modèle CHARME(p) avec $p < \infty$, sous réserve de multiple conditions. En particulier, les auteurs demandent la régularité du bruit blanc $(\epsilon_t)_{t\in\mathbb{Z}}$. En revanche, nous n'avons pas besoin de cette restriction ici.

3 Estimation des paramètres du modèle : consistance

Soit $(X_t)_{1-p\leq t\leq n}$ $n+p$ observations de la solution strictement stationnaire $(X_t)_{t\in\mathbb{Z}}$ du modèle (2) (cette solution existe grâce au Théorème 1). Supposons que le nombre d'états K est connu et que nous avons accès aux observations des variables cachées iid $(R_t)_{1-p\leq t\leq n}$, ou bien, des variables $(\xi_t^{(k)})_{1-p\leq t\leq n,k\in[K]}$. Une hypothèse similaire peut être trouvée dans la littérature pour des cas spécieux du modèle CHARME. Voir, *e.g.*, Tadjuidje-Kamgaing, J. (2005) et Stockis *et al.* (2010).

Notre objectif est d'étudier un estimateur non linéaire des paramètres

$$(\theta^0,\lambda^0) := (\theta_1^0,\ldots,\theta_K^0,\lambda_1^0,\ldots,\lambda_K^0)$$

du modèle CHARME(p) (2) à partir des observations $(X_t)_{1-p\leq t\leq n}$ et $(\xi_t^{(k)})_{1-p\leq t\leq n,k\in[K]}$. Cet objectif est atteint en résolvant le problème de minimisation

$$\begin{gathered}(\widehat{\theta}_n,\widehat{\lambda}_n)\in \operatorname{Argmin}_{(\theta,\lambda)\in\Theta\times\Lambda} Q_n(\theta,\lambda), \text{ où}\\ Q_n(\theta,\lambda) := \frac{1}{n}\sum_{t=1}^{n}\sum_{k=1}^{K}\xi_t^{(k)}\ell\big(X_t, f_k(X_{t-1},\ldots,X_{t-p},\theta_k), g_k(X_{t-1},\ldots,X_{t-p},\lambda_k)\big).\end{gathered} \tag{5}$$

Ici, $\ell : E\times E\times\mathbb{R}\longrightarrow\mathbb{R}\cup\{+\infty\}$ est une certaine fonction de coût. En général, ℓ devrait satisfaire $\ell(u,u,\tau)=0$, $\forall\tau$.

Afin de présenter notre résultat de consistance, il est plus commode de définir les processus

$$Y_t=(X_{t-p},X_{t-p+1},\ldots,X_t) \quad \text{et} \quad \xi_t=(\xi_t^1,\ldots,\xi_t^K), \quad t\in\mathbb{Z}.$$

Soit $(E^{p+1}\times B_e,\mathcal{E}^{\otimes(p+1)}\otimes\Xi,P)$ l'espace de probabilité commun, dans lequel sont définis les vecteurs aléatoires Y_t et ξ_t. Adoptons la notation suivante :

$$h(Y_t,\xi_t,\theta,\lambda) := \sum_{k=1}^{K}\xi_t^{(k)}\ell\big(X_t, f_k(X_{t-1},\ldots,X_{t-p},\theta_k), g_k(X_{t-1},\ldots,X_{t-p},\lambda_k)\big). \tag{6}$$

En utilisant des arguments complexes du calcul des variations (en particulier sur les intégrands normaux et l'épi-convergence (voir Rockafellar, R.T. (1976) et Rockafellar, R.T. and Wets, R.J.B. (1998)), nous pouvons établir la consistance de l'estimateur (5) sous de faibles conditions. En particulier, sans la nécessité d'avoir un échantillon iid ni la différentiabilité de la fonction Q_n. Ceci est résumé dans le théorème suivant :

Théorème 2. *Soit $(X_t)_{t\in\mathbb{Z}}$ une solution strictement stationnaire et ergodique du modèle* (2) *(elle existe grâce au Théorème 1 avec $C(m)<1$ pour certain $m\geq 1$). Considérons les conditions raisonnables (***A.1***)-(***A.7***) de Gómez-García et al. (2020). Alors,*

(i) chaque point d'accumulation de $(\widehat{\theta}_n,\widehat{\lambda}_n)_{n\in\mathbb{N}}$ appartient à $\operatorname{Argmin}_{(\theta,\lambda)\in\Theta\times\Lambda}\mathbb{E}h(Y,\xi,\theta,\lambda)$ p.s.

(ii) si de plus la suite $(Q_n)_{n\in\mathbb{N}}$ est équi-coercitive, et $\operatorname{Argmin}_{(\theta,\lambda)\in\Theta\times\Lambda}\mathbb{E}h(Y,\xi,\theta,\lambda)=\{\theta^0,\lambda^0\}$, alors $(\widehat{\theta}_n,\widehat{\lambda}_n)\to(\theta^0,\lambda^0)$ et $Q_n(\widehat{\theta}_n,\widehat{\lambda}_n)\to\mathbb{E}h(Y,\xi,\theta^0,\lambda^0)$ p.s.

4 Apprentissage du modèle avec des RNP

Il est connu que, étant donné toute fonction cible continue f et une précision cible $\epsilon > 0$, les réseaux de neurones (RN) avec suffisamment paramètres (poids et biaises) judicieusement choisis donnent une approximation de la fonction pour une erreur de taille ϵ. Cette propriété d'approximation universelle des RN, nous permet de considérer le modèle CHARME(p) (2) avec les fonctions f_k et g_k exactement modélisées par des RN, avec $E = \mathbb{R}^d$. Pour une introduction de réseaux de neurones (profonds), voir Section 2.2 de Gómez-García *et al.* (2020). Avec les mêmes notations de cette dernière section, pour chaque $k \in [K]$, soit $\theta_k = \Big((W_k^{(1)}, b_k^{(1)}), \ldots, (W_k^{(L_k)}, b_k^{(L_k)})\Big)$, où $W_k^{(l)}$ et $b_k^{(l)}$ sont respectivement la matrice des poids et le vecteur de biaises de la l-ème couche du RN f_k. Similairement $\lambda_k = \Big((\bar{W}_k^{(1)}, \bar{b}_k^{(1)}), \ldots, (\bar{W}_k^{(\bar{L}_k)}, \bar{b}_k^{(\bar{L}_k)})\Big)$ pour le RN g_k. De plus, nous considérons la même fonction d'activation φ pour toutes les couches des RN f_k, g_k, avec $k \in [K]$.

Ergodicité et Stationnarité. En considérant les notations du Théorème 1, les précédentes notations et en notant $\|W_k^l\|$ la norme spectrale de la matrice correspondante, on peut démontrer que

$$A_k = \big(\mathrm{Lip}(\varphi)^{L_k - 1} \prod_{l=2}^{L_k} \|W_k^{(l)}\|\big) \sum_{i=1}^{p} \|W_{k,i}^{(1)}\| \quad \text{et} \quad B_k = \big(\mathrm{Lip}(\varphi)^{\bar{L}_k - 1} \prod_{l=2}^{\bar{L}_k} \|\bar{W}_k^{(l)}\|\big) \sum_{i=1}^{p} \|\bar{W}_{k,i}^{(1)}\|.$$

Par conséquence, si $C(m) = 2^{m-1} \sum_{k=1}^{K} \pi_k \left(A_k^m + B_k^m \|\epsilon_0\|_m^m\right) < 1$ pour un certain $m \geq 1$, il existe une solution strictement stationnaire du modèle CHARME(p)-basé-sur-RN.

Consistance. Les conditions (**A.1**)-(**A.7**) de Gómez-García *et al.* (2020) sont satisfaites pour les RN f_k et g_k, pour tout $k \in [K]$ (cela a été montré dans le cité article). Donc, l'existence de la solution stationnaire et ergodique du modèle CHARME(p)-basé-sur-RN, implique le Théorème 2(i).

Pour pouvoir appliquer le Théorème 2(ii), nous avons besoin d'une certaine équi-coercitivité et unicité des vrais paramètres (θ^0, λ^0). Ceux-ci sont discutées et assurées dans la Section 6.2.1 de Gómez-García *et al.* (2020).

5 Commentaires

Établir la normalité asymptotique de l'estimateur (5) est très complexe dans un cadre variationnel. C'est pour cela que nous nous restreignons aux arguments habituels de la théorie d'inférence statistique qui demandent, en particulier, la dérivabilité d'ordre trois de la fonction Q_n. En plus, pour simplifier les résultats, nous prenons $\ell(u, v, \tau) = \|u - v\|^2/\tau^2$ et $g_k \equiv 1$, pour tout $k \in [K]$. Sous ces conditions et restrictions, nous établissons

la normalité asymptotique de l'estimateur (5). Les détails peuvent être trouvés dans la Section 5 de Gómez-García *et al.* (2020) et seront aussi discutés lors de cette présentation.

Références

Doukhan, P. and Wintenberger, O. (2008) *Weakly dependent chains with infinite memory.* Stochastic Processes and their Applications, 118 :1997–2013.

Gómez-García, JG., Fadili, J. and Chesneau, C. (2020) *Learning CHARME models with (deep) neural networks.* arxiv preprint arxiv :2002.03237.

Kirch, C. and Kamgaing, T. (2012) *Testing for parameter stability in nonlinear autoregressive models.* Journal of Time Series Analysis, 33(3) :365–385.

Liehr, S., Pawelzik, K., Kohlmorgen, J. and Moler, K.R. (1999) *Hidden markov mixtures of experts with an application to eeg recordings from sleep.* Th. of Biosc, 118 :246–260.

Lo, M.T., Tsai, P.H., Lin, P.F., Lin, C. and Hsin, Y.L. (2009) *The nonlinear and nonstationary properties in eeg signals : probing the complex fluctuations by hilbert-huang transform.* Advances in Adaptive Data Analysis, 1(3) :461–482.

Rockafellar, R.T. (1976) *Integral functionals, normal integrands and measurable selections.* In J. Gossez and L. Waelbroeck, editors, *Nonlinear Operators nd the Calculus of Variations*, number 543 in Lecture Notes in Mathematics, pages 157–207. Springer.

Rockafellar, R.T. and Wets, R.J.B. (1998) *Variational Analysis.* Springer.

Stockis, J-P., Franke, J. and Tadjuidje Kamgaing, J. (2010) *On geometric ergodicity of charme models.* Journal of Time Series Analysis, 31 :141–152.

Tadjuidje-Kamgaing, J. (2005) *Competing neural networks as model for nonstationary financial time series.* PhD thesis, University of Kaiserslautern.

Weigend, A.S. and Shi, S. (2000) *Predicting daily probability distributions of s&p500 returns.* Journal of Forecasting, 19(4) :375–392.

Yarotsky, D. (2017) *Error bounds for approximations with deep relu networks.* Neural Networks, 94 :103–114, 2017.

Scale matrix estimation under data-based loss in high and low dimensions

Mohamed Anis Haddouche [1], Dominique Fourdrinier [2] & Fatiha Mezoued [3]

[1] *Université de Normandie, INSA Rouen, UNIROUEN, UNIHAVRE, LITIS, avenue de l'Université, BP 8, 76801 Saint-Étienne-du-Rouvray, France.*
mohamed.haddouche@insa-rouen.fr et École Nationale Supérieure de Statistique et d'Économie Appliquée (ENSSEA), LAMOPS, Tipaza, Algeria.
[2] *Université de Normandie, UNIROUEN, UNIHAVRE, INSA Rouen, LITIS, avenue de l'Université, BP 12, 76801 Saint-Étienne-du-Rouvray, France.*
dominique.fourdrinier@univ-rouen.fr
[3] *École Nationale Supérieure de Statistique et d'Économie Appliquée (ENSSEA), LAMOPS, Tipaza, Algeria.*
famezoued@yahoo.fr

Résumé. Nous considérons le problème d'estimation de la matrice d'échelle Σ du modèle additif $Y_{p\times n} = M+\mathcal{E}$, du point de vue de la théorie de la décision. Ici, p représente le nombre de variables, n le nombre d'observations, M une matrice de paramètres inconnus de rang $q < p$ et $\mathcal{E}$ un bruit aléatoire de distribution à symétrie elliptique, de matrice de covariance proportionnelle à $I_n \otimes \Sigma$. Ce problème d'estimation est abordé sous une représentation canonique où la matrice d'observation Y est décomposée en deux matrices, à savoir, $Z_{q\times p}$ qui résume l'information contenue dans M et une matrice $U_{m\times p}$, où $m = n-q$, qui résume l'information suffisante pour l'estimation de Σ. Comme les estimateurs naturels de la forme $\hat{\Sigma}_a = a\,S$ (où $S = U^{\mathrm{T}}\,U$ et a est une constante positive) se comportent mal lorsque $p > m$ (S n'est pas inversible), nous proposons des estimateurs alternatifs de la forme $\hat{\Sigma}_{a,G} = a\big(S + S\,S^{+}G(Z,S)\big)$ où S^{+} est l'inverse de Moore-Penrose de S (qui coïncide avec l'inverse S^{-1} lorsque S est inversible). Nous fournissons des conditions sur la matrice de correction $SS^{+}G(Z,S)$ telles que $\hat{\Sigma}_{a,G}$ améliore $\hat{\Sigma}_a$ sous le coût basé sur les données $L_S(\Sigma,\hat{\Sigma}) = \mathrm{tr}\big(S^{+}\Sigma\,(\hat{\Sigma}\,\Sigma^{-1} - I_p)^2\big)$. Nous adoptons une approche unifiée des deux cas où S est inversible ($p \le m$) et S est non inversible ($p > m$).

Mots-clés. Distribution à symétrie elliptique, coût basé sur les données, identité de type Stein-Haff , matrice de covariance, matrice d'échelle.

Abstract. We consider the problem of estimating the scale matrix Σ of the additif model $Y_{p\times n} = M + \mathcal{E}$, under a theoretical decision point of view. Here, p is the number of variables, n is the number of observations, M is a matrix of unknown parameters with rank $q < p$ and $\mathcal{E}$ is a random noise, whose distribution is elliptically symmetric with covariance matrix proportional to $I_n \otimes \Sigma$. We deal with a canonical form of this model where Y is decomposed in two matrices, namely, $Z_{q\times p}$ which summarizes the information

contained in M, and $U_{m\times p}$, where $m = n-q$, which summarizes the sufficient information to estimate Σ. As the natural estimators of the form $\hat{\Sigma}_a = a\,S$ (where $S = U^{\mathrm{T}}\,U$ and a is a positive constant) perform poorly when $p > m$ (S non-invertible), we propose estimators of the form $\hat{\Sigma}_{a,G} = a\big(S + S\,S^{+}\,G(Z,S)\big)$ where S^{+} is the Moore-Penrose inverse of S (which coincides with S^{-1} when S is invertible). We provide conditions on the correction matrix $SS^{+}G(Z,S)$ such that $\hat{\Sigma}_{a,G}$ improves over $\hat{\Sigma}_a$ under the data-based loss $L_S(\Sigma,\hat{\Sigma}) = \mathrm{tr}\big(S^{+}\Sigma\,(\hat{\Sigma}\,\Sigma^{-1} - I_p)^2\big)$. We adopt a unified approach of the two cases where S is invertible ($p \le m$) and S is non-invertible ($p > m$).

Keywords. Eliptically symmetric distributions, data-based loss, Stein-Haff type identity, covariance matrix, scale matrix.

1 Introduction

Consider the following additive model

$$Y = M + \mathcal{E}, \qquad \mathcal{E} \sim ES(0_{np}, I_n \otimes \Sigma), \tag{1}$$

where Y is an observed $n \times p$ matrix, M denotes an $n \times p$ matrix of unknown parameters and $\mathcal{E}$ is an $n \times p$ elliptically symmetric distributed noise with unknown covariance matrix proportional to $I_n \otimes \Sigma$, where Σ is an unknown $p \times p$ invertible scale matrix and I_n is the n-dimensional identity matrix. Note that, the class of elliptically symmetric distributions encompasses a large number of important distributions such as Gaussian, Cauchy, exponential, Student and Weibull distributions. Our main assumption is that M is of low-rank, that is,

$$\mathrm{rank}(M) = q < p \tag{2}$$

Note that Model (1) is a common alternative representation of the multivariate low-rank regression model $Y = X\,\beta + \mathcal{E}$, where X is an $n \times q$ matrix of known constants of rank $q < p$ and β is an $q \times p$ matrix of unknown parameters. In the Gaussian setting, Model (1) arises in many fields that require to estimate M as in signal processing, image processing, collaborative filtering. Thus, it has been considered by various authors such as Candès and Recht (2009), Ji et al. (2010) and Candès et al. (2013). Recently, Canu and Fourdrinier (2017) introduced the extended elliptical setting in Model (1). It is worth noting that many estimation procedures of M rely on an accurate estimation of the scale matrix Σ, which is the aim of this paper.

Thanks to the low-rank assumption in (2), there exists a $n \times n$ orthogonal matrix $Q = (Q_1 Q_2)$, with $Q_2^{\mathrm{T}}\,M = 0$, so that the canonical form of Model (1) is given by

$$Q^{\mathrm{T}}\,Y = \begin{pmatrix} Q_1^{\mathrm{T}} \\ Q_2^{\mathrm{T}} \end{pmatrix} Y = \begin{pmatrix} Z \\ U \end{pmatrix} = \begin{pmatrix} \theta \\ 0 \end{pmatrix} + Q^{T}\,\mathcal{E}, \tag{3}$$

where Z and U are respectively $q \times p$ and $m \times p$ matrices with $m = n - q$ (*cf.* Canu and Fourdrinier (2017) for more details). Note that, the canonical form (3) separates information about the mean structure Z and the information concerning the scale U, since $S = U^T U$ summarizes the information to estimate Σ. Now, we restrict our attention to the setting where the joint density of Z and U is of the form

$$(z, u) \mapsto |\Sigma|^{-n/2} f\big[\, \mathrm{tr}\{(z-\theta)\Sigma^{-1}(z-\theta)^{\mathrm{T}}\} + \mathrm{tr}\{\Sigma^{-1}\, u^{\mathrm{T}} u\}\,\big]\ , \tag{4}$$

for some function f.

In the following, $E_{\theta,\Sigma}$ will denote the expectation with respect to the density (4) and $E^*_{\theta,\Sigma}$ the expectation with respect to the density

$$(z, u) \mapsto \frac{1}{K^*}|\Sigma|^{-n/2}\, F^*\big[\, \mathrm{tr}\{\,(z-\theta)\Sigma^{-1}(z-\theta)^{\top}\} + \mathrm{tr}\{\Sigma^{-1}\, u^{\mathrm{T}}\, u\}\,\big]\ ,$$

where $F^*(t) = \frac{1}{2}\int_t^\infty f(\nu)\, d\nu$ and the normalizing constant K^* is assumed to be finite. Note that, in the setting of a multivariate normal distribution, since $F^* = f$, these two expectations coincide.

As mentioned by James and Stein (1961), the natural estimators of the form $\hat{\Sigma}_a = a\, S$ (where a is a positive constant) perform poorly. Therefore, we consider alternative estimators of the form $\hat{\Sigma}_{a,G} = a\,(S + SS^+G(Z,S))$ and we derive dominance results under the data-based loss function

$$L_S(\hat{\Sigma}, \Sigma) = \mathrm{tr}\big(S^+\Sigma\,\big(\hat{\Sigma}\Sigma^{-1} - I_p\big)^2\big), \tag{5}$$

and its associated risk

$$R(\hat{\Sigma}, \Sigma) = E_{\theta,\Sigma}[L_S(\hat{\Sigma}, \Sigma)]\,, \tag{6}$$

where $\hat{\Sigma}$ is an estimator of Σ and S^+ is the Moore-Penrose inverse of S. It is worth noticing that this type of loss function, called data-based since it involves S^+, was introduced by Efron and Morris (1976). Since then, it was considered by various authors, in a Gaussian setting by Kubokawa and Srivastava (2008) and Tsukuma and Kubokawa (2015), and in a spherical setting by Fourdrinier and Strawderman (2015).

The two main features of our approach is that we consider the general elliptically symmetric distribution context and we unify the two cases where S is non-invertible ($p > m$) and S is invertible ($p \leq m$). The primary decision-theoretic results are presented in Section 2. More precisely, we derive a sufficient condition on the correction matrix function $SS^+G(Z,S)$ for which $\hat{\Sigma}_{a,G}$ improves on $\hat{\Sigma}_a$ under the data-based loss in (5). In Section 3, we provide numerical results through simulations.

2 Main result

Among the usual estimators $\hat{\Sigma}_a = a\,S$, there exists $a_o > 0$ such that $\hat{\Sigma}_{a_o}$ is optimal (that is, the risk of $\hat{\Sigma}_{a_o}$ is less than or equal to the risk of $\hat{\Sigma}_a$, for any $a > 0$); this is

$$a_o = \frac{1}{K^*(p \vee m)}\,,$$

where $p \vee m = \max\{p, m\}$ (*cf.* Haddouche (2019) for a proof). The improvement over the class of $a\,S$'s will be shown through the improvement of

$$\hat{\Sigma}_{a_o,G} = a_o\,(S + SS^+G(Z,S))\,, \tag{7}$$

over $\hat{\Sigma}_{a_o} = a_o\,S$, where

$$G(Z,S) = \frac{t}{\mathrm{tr}(S^+)}\,SS^+$$

and t is a positive constant. Note that the choice of this specific form of $G(Z,S)$ is motivated by the estimator considered by Konno (2009) in the normal case. We give sufficient conditions on the corrected factor $SS^+G(Z,S)$, that is on the constant t, such that the risk difference

$$\Delta(G) = R(\hat{\Sigma}_{a_o,G}, \Sigma) - R(\hat{\Sigma}_{a_o}, \Sigma)$$

between $\hat{\Sigma}_{a_o,G}$ and $\hat{\Sigma}_{a_o}$ is non-positive. Of course, $\Delta(G) \leq 0$ makes only sense if and only if $R(\hat{\Sigma}_{a_o,G}, \Sigma) < \infty$. It is shown in Haddouche (2019) that this occurs as soon as the expectations $E_{\theta,\Sigma}\left[\|S^+\,G\|_F^2\right]$, $E_{\theta,\Sigma}\left[\|\Sigma^{-1}SS^+G\|_F^2\right]$, $E_{\theta,\Sigma}\left[\mathrm{tr}(\Sigma\,S^+)\right]$ and $E_{\theta,\Sigma}\left[\mathrm{tr}(\Sigma^{-1}\,S)\right]$ are finite. In that case,

$$\Delta(G) = a_o^2\,K^*\,E_{\theta,\Sigma}\big[\mathrm{tr}\left(\Sigma^{-1}SS^+G\left\{I_p + S^+G + SS^+\right\}\right)\big] - 2\,a_o\,E_{\theta,\Sigma}\big[\mathrm{tr}\left(S^+\,G\right)\big]\,. \tag{8}$$

The dependence of the risk difference in (8) on the unknown parameter Σ^{-1} is problematic. As a remedy, we apply the Stein-Haff type identity in the framework of elliptically symmetric distribution given in Fourdrinier Haddouche and Mezoued (2019).

Lemma 1 *Let $G(z,s)$ be a $p \times p$ matrix function such that, for any fixed z, $G(z,s)$ is weakly differentiable with respect to s. Assume that $E_{\theta,\Sigma}\left[\left|\mathrm{tr}\left(\Sigma^{-1}S\,S^+\,G\right)\right|\right] < \infty$. Then we have*

$$E_{\theta,\Sigma}\left[\mathrm{tr}\left(\Sigma^{-1}\,SS^+\,G\right)\right] = K^*\,E^*_{\theta,\Sigma}\left[\mathrm{tr}\left(2\,SS^+\,\mathcal{D}_s\{SS^+G\}^{\mathrm{T}} + (m - (p \wedge m) - 1)\,S^+\,G\right)\right].$$

Thanks to this identity, sufficient conditions for improvement of $\hat{\Sigma}_{a_o,G}$ over $\hat{\Sigma}_{a_o}$, are given in the following theorem (*cf.* Haddouche (2019) for a proof) through an upper bound of the risk difference in (8).

Theorem 1 *Consider a density of the form (4). Let*

$$\hat{\Sigma}_{a_o,G} = a_o \left(S + \frac{t}{\mathrm{tr}(S^+)} SS^+ \right) \tag{9}$$

where t is a positive constant. Then $\hat{\Sigma}_{a_o,G}$ improves over $\hat{\Sigma}_{a_o}$ as soon as

$$0 \leq t \leq \frac{2\left((p \wedge m) - 1\right)}{(p \vee m) - (p \wedge m) + 1}\,.$$

where $p \wedge m = \min\{p, m\}$.

3 Numerical study

We deal here with the non-invertible case ($p > m$) for a Gaussian distribution ($K^* = 1$) where the scale matrix have an autoregressive structure of the form $(\Sigma)_{ij} = 0.9^{|i-j|}$. Note that simulation on the Student distributions are under study. We evaluate numerically the performance of the alternative estimator $\hat{\Sigma}_{a_o,G}$ in (9) where $a_o = 1/p$ and $t = 2\,(m-1)/(p-m+1)$, through the percentage relative improvement in average loss PRIAL of $\hat{\Sigma}_{a_o,G}$ over $\hat{\Sigma}_{a_o}$ defined as

$$\mathrm{PRIAL}(\hat{\Sigma}_{a_o,G}) = \frac{\text{average loss of } \hat{\Sigma}_{a_o} - \text{average loss of } \hat{\Sigma}_{a_o,G}}{\text{average loss of } \hat{\Sigma}_{a_o}} \times 100\,,$$

which is reported in the following table.

p	m	PRIAL (%)
20	4	15.00
20	8	18.56
20	12	25.56
20	16	47.034
100	20	3.39
100	40	4.19
100	60	5.76
100	80	10.42

Results of 1000 Monte Carlo simulation for $(\Sigma)_{ij} = 0.9^{|i-j|}$.

For $p = 20$ and $p = 100$, the PRIAL increases with the values of m. Note that, when $p = 20$ and $m = 16$ the PRIAL is close de 50 %. Note that the data-based Loss is much more discriminant then the usual quadratic loss for which the PRIAL is lower.

Bibliographie

Candès, E. J. and Sing-Long, C. A. and Trzasko, J. D. (2013). Unbiased Risk Estimates for Singular Value Thresholding and Spectral Estimators.*IEEE Transactions on signal processing* 61:4643-4657

Candès, E. J. and Recht, B. (2009). Exact Matrix Completion via Convex Optimization, *Foundations of Computational Mathematics*, 9(6): 717.

Efron, B. and Morris, C. (1976). Multivariate empirical Bayes and estimation of covariance matrices. *Annals of Statistics.* 4(1):22-32

Canu, S. and Fourdrinier, D. (2017). Unbiased risk estimates for matrix estimation in the elliptical case, *Journal of Multivariate Analysis*, 158, pp. 60-72

Foudrinier, D. Haddouche, M. A. and Mezoued, F. (2019). Scale matrix estimation of an elliptically symmetric distribution in high and low dimensions, *Université de Rouen Normandie et ENSSEA Tipaza,Technical report*

Fourdrinier, D. and Strawderman, W.E. (2015). Robust minimax Stein estimation under invariant data-based loss for spherically and elliptically symmetric distributions. *Metrika.* (4)78:461-484

Haddouche, M.A. (2019). Scale matrix estimation. *Ph.D disertation, Normandie Université*. Chapter 4

James, W. and Stein, C. (1961). Estimation with Quadratic Loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics.* 361–379

Ji, H. and Liu, C. and Shen, Z. and Xu, Y. (2010). Robust video denoising using low rank matrix completion *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition.* 100:2237-2253

Kubokawa, T. and Srivastava, M.S. (2008). Estimation of the precision matrix of a singular Wishart distribution and its application in high-dimensional data. *Journal of Multivariate Analysis.* (9)99:1906-1928

Konno, Y. (2009). Shrinkage estimators for large covariance matrices in multivariate real and complex normal distributions under an invariant quadratic los. *Journal of Multivariate Analysis* 100:2237-2253

Tsukuma, H. and Kubokawa, T. (2015). A unified approach to estimating a normal mean matrix in high and low dimensions. *Journal of Multivariate Analysis* 139:312-328

Finding "twin" electrical load curves for new customers using deep learning

Honorine Royer [1] & Anne Philippe [2] & Philippe Charpentier [3] & Laurent Bozzi [4]

[1] *EDF R&D - 7 boulevard Gaspard Monge, 91120 Palaiseau, honorine.royer@edf.fr*
[2] *Laboratoire de Mathématiques Jean Leray - Université de Nantes, 2 rue de la Houssinière, 44322 Nantes, anne.philippe@univ-nantes.fr*
[3] *EDF R&D - 7 boulevard Gaspard Monge, 91120 Palaiseau, philippe.charpentier@edf.fr*
[4] *EDF R&D - 7 boulevard Gaspard Monge, 91120 Palaiseau, laurent.bozzi@edf.fr*

Résumé. Nous étudions différentes approches de machine learning pour prévoir des courbes de charge électrique individuelles à partir de variables clients. Plus précisément, nous utilisons un auto-encodeur pour faire de la réduction de dimension sur les courbes. Nous appliquons ensuite des réseaux de neurones profonds à propagation avant, des processus gaussiens et des "deep Gaussian processes" pour estimer la fonction de régression expliquant la couche latente de l'autoencodeur à l'aide des prédicteurs (variables clients). Nous mettons en œuvre les méthodes citées pour un cas d'usage EDF sur la consommation d'électricité normalisée des clients non résidentiels.

Mots-clés. apprentissage profond, courbes de charge électrique, méthodes bayésiennes, processus gaussiens, inférence variationnelle, autoencodeurs

Abstract. We explore different machine learning methods to predict individual electrical load curves using customers informations. More precisely, we use an autoencoder to perform dimensionality reduction on the curves. We then apply deep feedforward neural networks, Gaussian processes and deep Gaussian processes to estimate the regression function between the latent space of the autoencoder and the predictors (customers' variables). We implement the methods mentioned above on a use-case from EDF concerning the scaled electricity consumption of non-residential customers.

Keywords. deep learning, electrical load curves, Bayesian methods, Gaussian processes, variational inference, autoencoders

1 Introduction

The transformations occurring on the French electricity market lead companies like Electricité de France (EDF) to innovate and provide new customer services. Those services are developed using statistical methods applied to customer data. Our study focuses on one use case that aims at finding for a new customer, a load curve among a library of known load curves, that suits the customer. Learning from customers with a relatively

long data history, we want to be able to find similar curve shapes for new or potential customers. This study focuses on non-residential customers, labelled by their contract power, specifically $C2$ and $C4$ customers. Those two groups differ on their consumption, $C2$ being larger consumers than $C4$. For a new client or a prospect, we want to find, among a library of scaled consumption curves, one that, in terms of shape, would be similar to the one of the potential client. The dataset used for the study is provided by EDF and contains about 7% of $C4$ customers and 93% of $C2$ customers. For both types of clients, we have their consumption time series (load curves), over one year, by half hourly period and their billing informations (such as the business sector, or the peak power per month). We aim at predicting $C4$ consumptions, but we lack historical data on them. Hence we use informations from the bigger database of $C2$ customers to find a "twin" load curve that corresponds to $C4$ clients. Due to the imbalanced database, the weight of the $C4$ group is too small for the $C4$ customers to impact strongly the global estimation, hence they are not used during training in this study.

The load curves of non-residential customers are highly heterogeneous though not very thermosensitive, their period of activity can vary substantially as well (for instance some are active only during certain months, while others are active throughout the whole year). This heterogeneity makes prediction and forecasting tasks challenging, but those tasks are essential to companies like EDF, e.g. to evaluate the sourcing on the wholesale market.

In this study, we focus on deep learning approaches. Deep learning is a field of machine learning with many recent successful applications in various domains. It mostly refers to artificial neural networks such as feedforward neural networks, convolutional neural networks, deep belief networks etc. . For an extended review of deep learning models, refer to Goodfellow et al. (2016). We first apply an autoencoder to reduce the dimension of the load curves and then we model the latent space using various types of neural networks, some including skip connections that allow to capture information that might have been lost between different layers. We also use Gaussian processes and deep Gaussian processes to that end. Deep Gaussian processes (DGP) are deep models made of composition of functions, where each function is drawn from a Gaussian process prior. See Damianou & Lawrence (2013) for the complete specification of this type of model, and Bui et al. (2016) for a detailed survey on DGP applied to regression.

The complete methodology is described in Section 2. Section 3 contains the application of the method to EDF data.

2 Our Methodology

We want to estimate the individual load curves written $\mathbf{X}$ as a function of the clients' features $\mathbf{V}$ constituted of billing informations. Then, the estimated function is used to predict the load curves of new customers. Due to the high dimension of the curves $\mathbf{X}$, we work on a condensed version of them. To predict the curves of $C4$ customers, we build our

models on the $C2$ customers' database $(\mathbf{X}^{C2}, \mathbf{V}^{C2})$ while assuming that the two groups have similar behaviors that allow us to apply the same models. This hypothesis is quite strong and could be relaxed in future work. The methodology consists in the following steps :

1. *Reduction of dimension using an autoencoder*

   We use a particular architecture of neural networks called autoencoder to reduce the dimension of the load curves. Recall that a one-hidden-layer neural network is defined by :

   $$\mathbf{h} = \sigma_1(\mathbf{W}_1^T \cdot \mathbf{x} + \mathbf{b}_1), \tag{1}$$
   $$\mathbf{y} = \sigma_2(\mathbf{W}_2^T \cdot \mathbf{h} + \mathbf{b}_2), \tag{2}$$

   The autoencoder is a particular case of neural network, as described in equations (1) and (2), with $\mathbf{y} = \mathbf{x}$, and is defined by :

   $$\mathbf{I} = e(\mathbf{x}), \tag{3}$$
   $$\mathbf{x} = d(\mathbf{I}), \tag{4}$$

   where e and d are the encoding and the decoding functions respectively, $\mathbf{I}$ is the latent or encoded space. The functions e and d are made of one or several layers illustrated in equations (1) and (2).
   Let n be the number of $C2$ customers. On $(\mathbf{X}_k^{C2})_{1\leq k\leq n}$, the set of $C2$ load curves, we train an autoencoder. We then extract the latent space written $\mathbf{I}^{C2} = (\mathbf{I}_k^{C2})_{1\leq k\leq n}$.

2. *Modelling of the latent space* $\mathbf{I}$ *with the features* $\mathbf{V}$

   Considering the following non linear regression model :

   $$\mathbf{I} = f(\mathbf{V}) + \varepsilon,$$

   where f is the regression function and ε some error term, we want to estimate f in high dimension hence we focus on different approaches. We estimate the function f with the following models :

   - **NN1** : a one-hidden-layer feedforward neural network.
   - **GP2NN1** : a 2 layers feedforward neural network composed of a Dense layer and a Gaussian process layer.
   - **RN1** : a 9 layers neural network with skip connections.
   - **GP1RN1**: a 10 layers neural network with skip connections and one Gaussian process layer.
   - **GP2RN1** : an 11 layers neural network with skip connections and two Gaussian process layers.

Hereafter, the estimator of each model is denoted $\hat{f}_{C2}$.

3. *Prediction the latent space of the $C4$ curves*

   For a $C4$ customer, we predict the latent variables using their billing informations $\mathbf{V}^{C4}$:

$$\hat{\mathbf{I}}^{C4} = \hat{f}_{C2}(\mathbf{V}^{C4}).$$

4. *Identifying the "twin" load curves*

   For a $C4$ customer, we have the estimated latent variables $\hat{\mathbf{I}}^{C4}$ and we seek the nearest neighbor in $(\mathbf{I}_k^{C2})_{1\leq k\leq n}$. More precisely, we want to find the value of k for which the L_1 distance between $\hat{\mathbf{I}}^{C4}$ and $\mathbf{I}_k^{C2}$ is minimal :

$$\hat{\mathbf{k}} = \text{argmin}_{1\leq k\leq n}|\mathbf{I}_k^{C2} - \hat{\mathbf{I}}^{C4}|.$$

   We then predict, for a $C4$ customer the associated load curve $(\mathbf{X}_{\hat{k}}^{C2})$.

Remark 1 *When building the latent space, the ReLU activation function in the autoencoder creates several nil columns. They are removed in the second part of the method. We also focus on columns with an "acceptable" rate of zero values, which means that any latent variable that exceeds said rate is not considered for our study.*

Remark 2 *DGP and Gaussian process layers are Bayesian models, their posterior distributions is intractable. They are approximated using variational methods (see Blei et al. (2017) for a detailed survey).*

3 Results and Discussion

For our implementation and experiments, we use `R` software (R Core Team (2019)), TensorFlow (Abadi et al. (2015)) and TensorFlow Probability (Dillon et al. (2017)). To evaluate the performances at each stages, we consider the relative $L1$ error. The error between $\mathbf{Z}$ and $\hat{\mathbf{Z}}$, its estimation, is :

$$\mathcal{E}(\mathbf{Z}, \hat{\mathbf{Z}}) = \frac{\sum_{i=1}^{n} |Z_i - \hat{Z}_i|}{\sum_{i=1}^{n} |Z_i|}. \tag{5}$$

We first estimate the functions e and d (recall the equations (3) and (4)) on the $C2$ load curves. We evaluate the autoencoder's properties in terms of the reconstruction error $\mathcal{E}(\mathbf{X}^{C2}, \hat{d}(\hat{e}(\mathbf{X}^{C2})))$. This error is calculated on the $C4$ load curves as well to see if the autoencoder is robust. In Figure 1,we show the boxplots of the reconstruction error on the $C2$ load curves (on the left) and on the $C4$ load curves (on the right). As expected, the reconstruction error on the $C4$ curves is slightly higher than on the $C2$ curves, and they both seem to have the same order of magnitude. This validates the assumption that the same latent variables summarize well both databases.

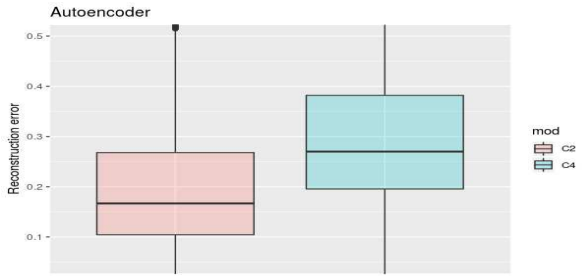


Figure 1: Boxplots of the reconstruction errors $\mathcal{E}(\mathbf{X}^{C2}, \hat{d}(\hat{e}(\mathbf{X}^{C2})))$ and $\mathcal{E}(\mathbf{X}^{C4}, \hat{d}(\hat{e}(\mathbf{X}^{C4})))$.

Remark 3 *To evaluate the stability of the autoencoder, we have run it six times on the $C2$ database. The six replications showed very similar errors so only one of them is displayed.*

We extract the latent space predicted by the autoencoder on the $C4$ curves $\mathbf{I}^{C4}$ and compare it with its estimations $\hat{\mathbf{I}}^{C4}$. We compute the error terms $\mathcal{E}(\mathbf{I}^{C4}, \hat{\mathbf{I}}^{C4})$ and plot boxplots for each model in Figure 2. Comparing the median of errors, the best model is **GP2RN1**, however, the interquartile range is slightly larger than the one of the **RN1**. The median of errors of **RN1** and **NN1** are the second and third lowest of the five models respectively.

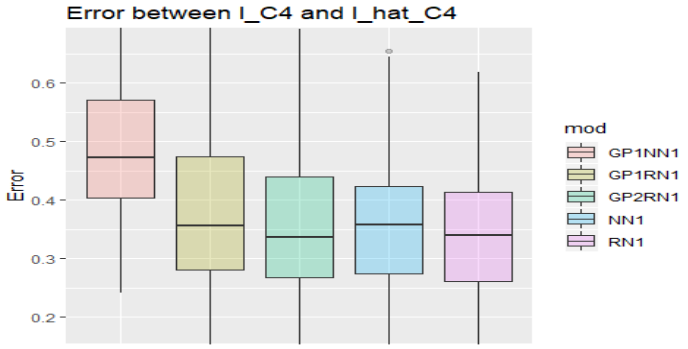


Figure 2: Boxplots of the errors $\mathcal{E}$ between $\mathbf{I}^{C4}$ and $\hat{\mathbf{I}}^{C4}$ for each model.

We represent the boxplots of the error $\mathcal{E}(\mathbf{X}^{C4}, \mathbf{X}^{C2}_{\hat{k}})$ for each model (see Figure 3), after minimizing the distances between $\hat{\mathbf{I}}^{C4}$ and $\mathbf{I}^{C2}$. Here, the same models as in Figure 2 seem to have the lowest median error. The **GP2RN1** performances are slightly deteriorated, but are very similar to the **RN1** and **NN1** ones. Due to the complexity of this problem (consumption time series by half-hourly period over one year), the errors obtained are quite satisfactory. The poor performances of the more complex models can be explained by the difficulty to estimate them on this size of data, thus **NN1**, the simpler model,

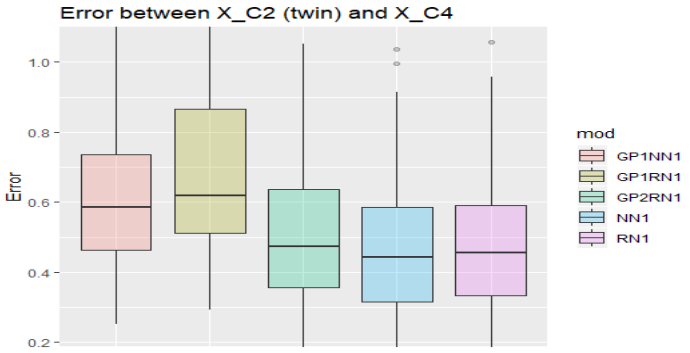


Figure 3: Boxplots of the errors $\mathcal{E}$ between the twin load curve $\mathbf{X}^{C2}$ and real load curve $\mathbf{X}^{C4}$ for each model.

gives out the best performances. Those results are in line with Ockam's Razor principle. The benefit of using **GP2RN1** is to get prediction intervals, because GP layers output distributions. Hence, it could be used to build prediction intervals on the load curves of each consumer.

An alternative to finding the "twin" load curve is to decode the predicted latent spaces by taking $\hat{d}(\hat{\mathbf{I}}^{C4})$. However, to apply this, we need to model every latent variable. Here, we choose to model only the variables in the latent space that contain a reasonable rate of zero values. Future research could include the modelling of said variables in addition to the initial methods considered. We also may consider using transfer learning to add some $C4$ contributions to our models as a way to counter the imbalanced database.

Bibliography

Abadi, M. et al. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112.518 : 859-877.

Bui, T., Hernández-Lobato, D., Hernandez-Lobato, J., Li, Y., & Turner, R. (2016). Deep Gaussian processes for regression using approximate expectation propagation. In *International conference on machine learning* , pp. 1472-1481.

Damianou, A., & Lawrence, N. (2013). Deep Gaussian processes. In *Artificial Intelligence and Statistics*, pp. 207-215.

Dillon, J. V. et al. (2017). Tensorflow distributions. *arXiv preprint arXiv:1711.10604*.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). Deep learning.*MIT press*, 29.

R Core Team (2019). R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing, Vienna, Austria* https://www.R-project.org/

Bayesian inference for transfer learning

Loïc Iapteff [1], Julien Jacques [2], Matthieu Rolland [3] & Benoit Celse [4]

[1] *loic.iapteff @ifpen.fr, Rond-point de l'échangeur de Solaize, 69360 Solaize*
[2] *julien.jacques @univ-lyon2.fr, 5 Avenue Pierre Mendès France, 69500 Bron*
[3] *matthieu.rolland @ifpen.fr, Rond-point de l'échangeur de Solaize, 69360 Solaize*
[4] *benoit.celse @ifpen.fr, Rond-point de l'échangeur de Solaize, 69360 Solaize*

Résumé. Le groupe IFP commercialise des catalyseurs et doit s'engager sur leur performance. Il est donc nécessaire de disposer de modèles prédictifs fiables pour chaque nouvelle génération de catalyseurs. Ces modèles sont construits à partir de données expérimentales très couteuses. Afin d'optimiser les coûts, notre ambition est de réduire le nombre d'expérimentations nécessaires pour estimer un modèle associé à un nouveau type de catalyseur, en transférant l'information contenue dans les modèles d'anciennes générations. Cet article décrit nos travaux sur le transfert de modèle linéaire par inférence bayésienne.

Mots-clés. Transfer learning, inférence bayésienne, modèle linéaire

Abstract. IFP group develops catalysts and has to guarantee their performances. It is therefore crucial to have good predictive models for all new catalysts. These models are built upon very expensive experimental data. In order to minimize costs, we aim at reducing the number of new data points to measure to fit a model on the new catalyst, that is by using the knowledge available in the previous model. This paper describes our work on linear model transfer using Bayesian inference.

Keywords. Transfer learning, Bayesian inference, linear model

1 The problem

IFP group develops and sells catalysts to the chemical, bio chemical producers. Catalysts are solids that render the reaction feasible, faster, and / or at lower temperature and pressure. The performance must be guaranteed and it is therefore crucial to have good predictive models for all new catalysts. These models are built upon very expensive experimental data and generally without accounting for the former catalysts' datasets. In order to minimize costs, we aim at reducing the number of new data points to measure on the new catalyst by transferring the models. By transferring, we mean use the knowledge available in the previous model and mix it with the new data points to build a good prediction model with a minimum of experimentation.

The problem is a transfer learning problem. Let's define a domain as $D = (X, P(X))$ with X a feature space and $P(X)$ its probability distribution, and an associated task

$T = (Y, f)$ with f the function used to predict $y \in Y$ given $\mathbf{x} \in X$. Pan & Yang (2010) define transfer learning as follow: "Given a source domain D_s and learning task T_s, a target domain D_t and learning task T_t, transfer learning aims to help improve the learning of the target predictive function f_t in D_t using the knowledge in D_s and T_s, where $D_s \neq D_t$, or $T_S \neq T_T$". In our case, $D_s = D_t$ and $T_s \neq T_t$ as the catalyst and its performance is different, which put us in the inductive transfer learning case (Pan & Yang (2010)). The catalyst changes, so reaction will change and features have no reason to follow a particularly different distribution. There are different methods to solve these problems, such as transferring knowledge of instances, features or parameters.

We want a model to predict the output property Y_i with some information on feed and operating conditions, described using 12 features identical for the source and the target catalyst. Different models are tested to predict the output property on source data, specifically linear model, support vector, multi-layer perceptron, random forest, gradient boosting and kriging (Matheron (1969)). Best predictions are achieved with kriging but the linear model also offers satisfying results. Therefore, in this work the linear model is considered for its simplicity.

The model for the source catalyst is

$$Y_i = \beta_{s0} + \sum_{j=1}^{p} \beta_{sj} X_{ij} + \epsilon_i \tag{1}$$

with $\epsilon_i \sim \mathcal{N}(0, {\sigma_s}^2)$ and $p = 12$. Let $\hat{\theta}_s$ be the maximum likelihood estimate of $\theta_s = (\beta_{s0}, \ldots, \beta_{sp}, \sigma_s^2)$. The training data set available for the source catalyst is assumed to be sufficiently large such that $\hat{\theta}_s$ can be considered as satisfying estimate of θ_s.

Our goal is to estimate the same model but for the target catalyst

$$Y_i = \beta_{t0} + \sum_{j=1}^{p} \beta_{tj} X_{ij} + \epsilon_i \tag{2}$$

for which the available training data set is of a smaller size n_t.

We choose to focus on transfer knowledge of parameters because it's well adapted for the transformation of the linear model. Two approaches are considered. The first one is inspired from Bouveyron & Jacques (2010) and consists in identifying a link between θ_s and θ_t. The second one is a Bayesian approach and consider a Bayesian linear model for (2) with prior distribution on θ_t depending on θ_s.

2 Transfer models

The objective is to have a good model with as few points as possible. In the experiments presented in this paper, the performance of the transfer techniques are evaluated for

different numbers n_t of new points. The evaluation of the transferred model may depend on the sampling of training set. For this reason we present average results for 10 random samplings. For each sampling, RMSE score are evaluated on a test set independent of the training set. In the present industrial context, a model is considered satisfying if the RMSE score is lower than 0.005. With the source model the RMSE score is 0.0033 on the source data.

2.1 Transfer learning using parametric link

This method uses the idea of Bouveyron & Jacques (2010): keep some parameters unchanged for the target model ($\beta_{sj} = \beta_{tj}$ for some j), considering the influence doesn't change between both models, and then learn only others parameters.

If $\mathcal{M}$ is the set of index of parameters to be modified, then $\beta_{tj} = \beta_{sj}$ for $j \in \{1, ..., p\}\backslash\mathcal{M}$ and $\beta_{tj} = \lambda_j \beta_{sj}$ for $j \in M$. Only a reduced number of parameters have to be estimated for the target model. The challenge with this approach is the choice of $\mathcal{M}$. In this work $\mathcal{M}$ is selected by leave-one-out cross validation on the n_t target points, for all possible size of $\mathcal{M}$. For a given size n_t, we start by choosing a parameter to modify by taking the one minimizing the RMSE by leave one out on the n_t target points. Then a second parameter is chosen in the same way, this time learning a model by modifying these two parameters. We add one parameter at a time until we modify them all. With this approach, a performing model can be fitted with less points than if a totally new model is learned (Figure 1 left). For example, with 10 observations and modifying 1 parameter, RMSE is smaller than the objective of 0.005. In contrast, 30 observations are needed to a learned from scratch model to achieve such good results. With 100 observations, learned from scratch model is better. Changing a small number of parameters is more efficient for small training set but worse when a lot of data is available. An emerging challenge is the choice of the size of $\mathcal{M}$. Choosing a small size for $\mathcal{M}$ offers quick results. But by also trying to determine its size by cross validation, the results deteriorate.

Another remark, for a given size of $\mathcal{M}$, the modified parameters are not the same for small and large values of n_t where they do not change. In other words, we are not able to find the best parameters to transform with a few points. Assuming to know the best parameters to change, results are better in terms of number of target points, n_t (Figure 1 right). Parameters chosen to be modified are those chosen with 100 observations. So far, we have not been able to identify a technique to decide which are the best parameters to change on the available n_t points. For this reason, we explored Bayesian inference.

2.2 Transfer learning using Bayesian approach

In this section a Bayesian approach is used to learn parameters for the new linear model. The idea is to choose prior distributions depending on the source data. The model is then:

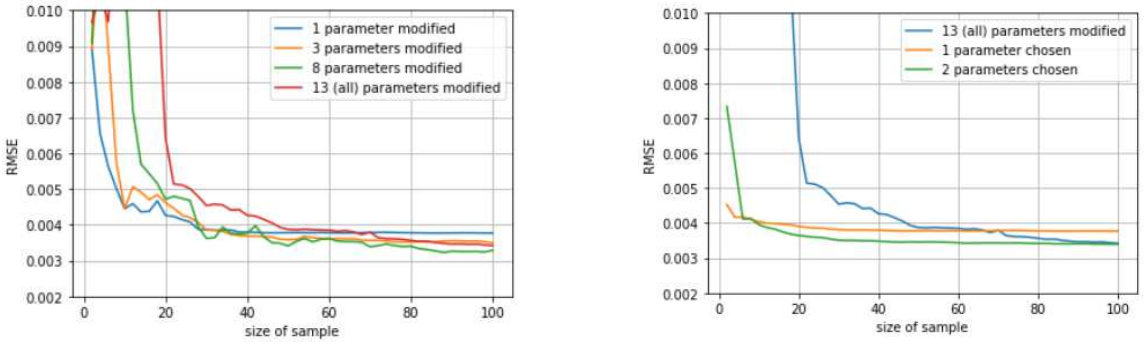


Figure 1: Graphs shows the evolution of RMSE according to n_t. On the left, parameters to modify are chosen by cross validation, on the right they are chosen knowing they are the best to modify.

$$\begin{aligned} Y_i &= \beta_{t0} + \sum_{j=1}^{p} \beta_{tj} X_{ij} + \epsilon_i, \\ \boldsymbol{\beta}_t &\sim \pi(\boldsymbol{\beta}_t), \\ \epsilon_i &\sim \mathcal{N}(0, {\sigma_t}^2), \\ \sigma_t &= \sigma_s, \end{aligned}$$

where $\boldsymbol{\beta}_t = (\beta_{t0}, \beta_{t1}, \ldots, \beta_{tp})^T$.

The Bayes Theorem gives that the posterior of $\boldsymbol{\beta}_t$ is

$$\pi(\boldsymbol{\beta}_t | \boldsymbol{Y}_t) = \frac{\pi(\boldsymbol{\beta}_t) f(\boldsymbol{Y}_t | \boldsymbol{\beta}_t)}{f(\boldsymbol{Y}_t)},$$

with $\boldsymbol{Y}_t = (Y_1, ..., Y_{n^t})^T$.

We consider different prior distribution $\pi(\boldsymbol{\beta}_t)$. The first one is the well known Zellner's prior (Zellner, 1986), also known as g-prior, for parameters $\boldsymbol{\beta}_t$:

$$\pi(\boldsymbol{\beta}_t) \sim \mathcal{N}(\widehat{\boldsymbol{\beta}}_s, g\sigma_t^2 (\boldsymbol{X}_t^T \boldsymbol{X}_t)^{-1}),$$

with $\widehat{\boldsymbol{\beta}}_s$ the maximum likelihood estimator (MLE) learned on the source data and $\boldsymbol{X}_t$ the $n^t \times (p+1)$ matrix of target observations, $(\boldsymbol{X}_t)_{i1} = 1$ for $i = 1, ..., n$.

Using such a prior, only the mean of the prior distribution depends on the source data. The variance of the prior depends only on target data and on a fixed parameter g. Notice that the posterior's mean using such a prior is a weighted average between the MLE and the mean of the prior. In our case, a weighted average between the MLE for source data and the MLE for target data is calculated: $\widehat{\boldsymbol{\beta}}_t = \frac{1}{g+1}(g\widehat{\boldsymbol{\beta}}_{MLE,t} + \widehat{\boldsymbol{\beta}}_s)$.

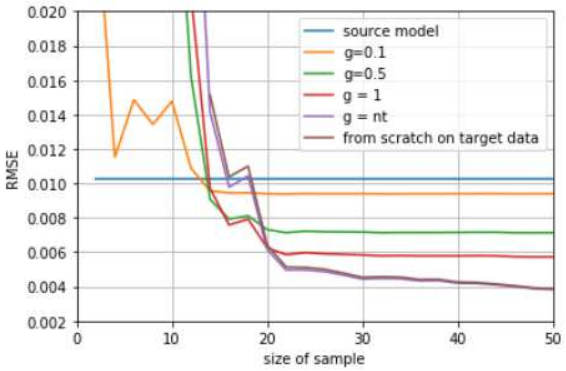


Figure 2: Comparison between an estimation of $\boldsymbol{\beta}_t$ with a Bayesian approach and a g-prior for different values of g, and a model learned without any prior.

With this prior, results are not satisfying whatever the value of g. $\widehat{\boldsymbol{\beta}}_s$ is not a good estimator for $\boldsymbol{\beta}_t$, thus averaging with the MLE for target doesn't improve results.

An idea to improve the results is to increase the information transferred. Source data is employed to learn both the mean and the variance of prior. Once again, a Gaussian prior is considered for parameters $\boldsymbol{\beta}_t$. The prior mean is $\widehat{\boldsymbol{\beta}}_s$. The variance is chosen as the variance of $\widehat{\boldsymbol{\beta}}_s$ scaled with a scalar λ. Let remark that when $\lambda = 1$, this prior for $\boldsymbol{\beta}_t$ corresponds to the posterior distribution of $\boldsymbol{\beta}_s$ estimated in a Bayesian model with an uniform prior for $\boldsymbol{\beta}_s$. The introduction of the factor λ allows more flat prior:

$$\pi(\boldsymbol{\beta}_t) = \mathcal{N}(\widehat{\boldsymbol{\beta}}_s, \lambda\sigma_s^2(\boldsymbol{X}_s^T\boldsymbol{X}_s)^{-1}).$$

The mean of the corresponding posterior distribution leads to:

$$\widehat{\boldsymbol{\beta}}_t = (\boldsymbol{X}_t^T\boldsymbol{X}_t + \sigma_t^2\lambda^{-1}\boldsymbol{\Sigma}^{-1})^{-1}(\boldsymbol{X}_t^T\boldsymbol{Y}_t + \sigma_t^2\lambda^{-1}\boldsymbol{\Sigma}^{-1}\widehat{\boldsymbol{\beta}}_s),$$

with $\boldsymbol{\Sigma} = \sigma_s^2(\boldsymbol{X}_s^T\boldsymbol{X}_s)^{-1}$.

The parameter value λ must yield the best performance with the fewest observations possible. When $\lambda \to \infty$ the posterior mean tends to the MLE. When $\lambda \to 0$ the posterior mean tends to the prior. To see the impact of lambda, different values are tried and RMSE evolution is evaluated (Figure 3 left). There is an optimum in the λ value in the range $100 - 1000$ for our data.

Data are normalized, thus parameters $\boldsymbol{\beta_t}$ take values in $[-1, 1]$. A compromise must be found between prior information and variability of parameters. A good choice of standard deviation for each parameter seems to be near from 1 to cover the $[-1, 1]$ interval and not be too wide. Taking a $\lambda = (mean((\Sigma_{jj})_{j=0,..,p}))^{-1} \simeq 800$ on our data yields an average variance of 1 for each parameter. This value of λ offers good results with few points (Figure 3 right). With this approach, performing models can be fitted with a number of target points smaller than without knowledge on old catalyst. RMSE score of 0.004 can be reached with only 5 target points instead of 50 for a learned from scratch model.

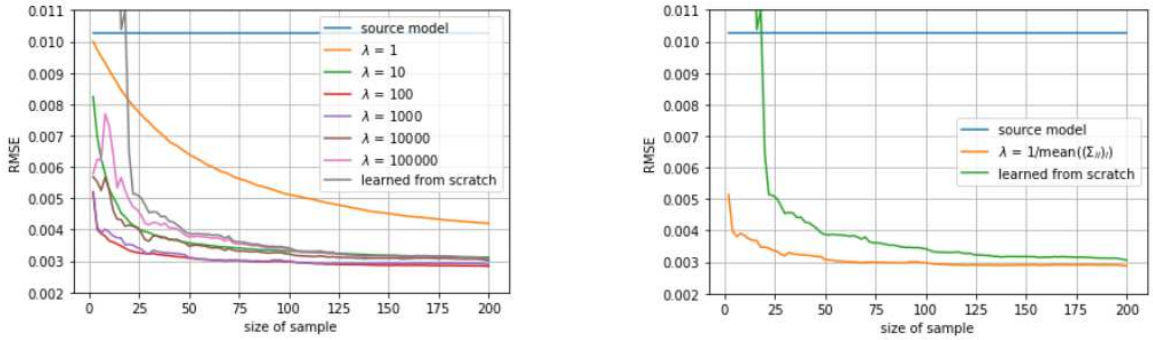


Figure 3: Impact of λ on $\pi(\boldsymbol{\beta}_t|\boldsymbol{Y}_t)$.

3 Perspectives

Bayesian inference with an optimized learning parameter (lambda) gives good results on transfering linear models on our data set. The knowledge from an old catalyst, namely the parameter covariance matrix and values, is shown to improve the quality of the linear model for the new catalyst. Future works will focus on 2 mains areas: the transfer of kriging models, still with a bayesian transfer knowledge, and the design of experiments by choosing the best target data to use for transfer.

References

Bouveyron, C., & Jacques, J. (2010). Adaptive linear models for regression: improving prediction when population has changed. *Pattern Recognition Letters*, 2237-2247.
Matheron, G. (1969) Le krigeage universel (Universal kriging) Vol. 1. Cahiers du Centre de Morphologie Mathematique, Ecole des Mines de Paris, Fontainebleau, 83 p.
Pan, S., & Yang, Q. (2010). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 1345-1359.
Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In: Goel, P. and Zellner, A., Eds., Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti, Elsevier Science Publishers, Inc., New York, 233-243

Estimation itérative en propagation d'incertitudes : réglage robuste de l'algorithme de Robbins-Monro

Bertrand Iooss [1]

[1] *EDF R&D, 6 Quai Watier, 78401 Chatou - bertrand.iooss@edf.fr*

Résumé.
En quantification d'incertitudes de modèles numériques, l'estimation de quantiles des sorties du modèle est réalisée usuellement par l'analyse statistique de l'échantillon complet de la variable étudiée. Cette approche n'est pas applicable lorsque des quantités prohibitives de données sont générées à chaque simulation. Ce problème peut être résolu grâce à une technique d'estimation à la volée (itérative) basée sur l'algorithme de Robbins-Monro. Nous étudions numériquement cet algorithme afin d'estimer une fonction quantile discrétisée à partir d'échantillons de taille limitée (quelques centaines d'observations). En pratique, la distribution de la variable sous-jacente étant inconnue, il est essentiel de définir des valeurs "robustes" des paramètres de l'algorithme, afin que les estimations des quantiles soient raisonnablement bonnes dans la plupart des situations.

Mots-clés. Quantification d'incertitudes, Quantile, Estimation itérative, Robbins-Monro, Moyennisation

Abstract.
In uncertainty quantification of numerical simulation models, the classical approach for quantile estimation requires availability of the full sample of the studied variable. This approach is not suitable at exascale as large ensembles of simulation runs would need to gather a prohibitively large amount of data. This problem can be solved thanks to an on-the-fly (iterative) approach based on the Robbins-Monro algorithm. We numerically study this algorithm for estimating a discretized quantile function from samples of limited size (a few hundreds observations). As in practice, the distribution of the underlying variable is unknown, the goal is to define "robust" values of the algorithm parameters, which means that quantile estimates have to be reasonably good in most situations.

Keywords. Uncertainty Quantification, Quantile, Iterative estimation, Robbins-Monro, Averaging

1 Introduction

Lors du développement et de l'utilisation des modèles de simulation numérique, les analyses d'incertitudes et de sensibilité sont des outils précieux (Smith, 2014). Elles nécessitent d'exécuter plusieurs (voire de nombreuses) fois le modèle de simulation avec différentes

valeurs des entrées du modèle (suivant des lois de probabilité prédéfinies) afin de calculer des quantités statistiques d'intérêt (notées QoI) sur les sorties du modèle (i.e. leur moyenne, variance, quantiles, indices de sensibilité, ...). La pratique usuelle consiste à stocker tous les résultats de simulation avant de calculer les QoI. Dans certains cas où des variables d'état dépendant du temps et de l'espace sont simulées, la masse de données produites rend prohibitifs leur stockage et leurs temps de lecture (nécessaires à l'estimation des QoI). Une solution proposée récemment dans Terraz et al. (2017) consiste à ne pas stocker les sorties des simulations en calculant les QoI à la volée. Cela amène à considérer des problèmes d'estimation statistique itérative, sujet relativement classique dans le traitement des gros volumes de données mais peu explorés dans les études d'incertitudes de modèles numériques.

Dans ce travail, nous nous intéressons à l'élaboration d'un algorithme d'estimation itératif en propagation d'incertitudes (dans la suite de Ribés et al., 2019), alors que l'analyse de sensibilité itérative a été étudiée dans Terraz et al. (2017). Nous nous focalisons sur l'estimation de quantiles, éléments essentiels pour le calcul d'intervalles de prédiction ou de tolérance, et pour la détection d'outliers, en particulier dans les études de sûreté (voir un exemple dans le domaine de l'ingénierie nucléaire dans Iooss and Marrel, 2019). En se restreignant (par souci de concision) à une sortie scalaire, nous cherchons un estimateur $\hat{q}_\alpha$ des α-quantiles q_α (de la variable aléatoire $Y \in \mathbb{R}$) définis par:

$$q_\alpha = \inf\{y \in \mathbb{R} \,|\, \mathbb{P}(Y \leq y) \geq \alpha\} \,, \tag{1}$$

avec $\alpha \in [\alpha_{\min}, \alpha_{\max}]$ où $\alpha_{\min}$ $(\in]0,1[)$ et $\alpha_{\max}$ $(\in]0,1[)$ sont les valeurs minimale et maximale des ordres des quantiles estimés. Dans notre étude, $\alpha_{\min}$ (resp. $\alpha_{\max}$) sera égal à 5% (resp. 95%). L'estimateur empirique de q_α, associant à l'échantillon i.i.d. $(Y_1, \ldots, Y_N)$ l'échantillon ordonné $(Y_{(1)}, \ldots, Y_{(N)})$, s'écrit $\hat{q}_\alpha^N = Y_{(\lfloor \alpha N \rfloor + 1)}$.

A la place de cet estimateur, nous étudions l'algorithme de Robbins-Monro (RM) (Robbins and Monro, 1951) bien connu pour l'estimation itérative de quantile. L'une des spécificités de notre étude, comme dans Tierney (1983), réside dans la faiblesse de la taille de l'échantillon disponible (quelques centaines d'observations). Ainsi, les propriétés asymptotiques de l'estimateur considéré, quoique apportant des garanties de convergence essentielles, seront peu exploitables pour le réglage des algorithmes.

L'algorithme RM consiste à mettre à jour l'estimateur courant du quantile (noté $q_\alpha(n)$) à chaque nouvelle observation Y_{n+1} avec $n \geq 1$ par la formule de récurrence

$$q_\alpha(n+1) = q_\alpha(n) - \frac{C}{n^\gamma}\left(\mathbb{1}_{Y_{n+1} \leq q_\alpha(n)} - \alpha\right) \,, \tag{2}$$

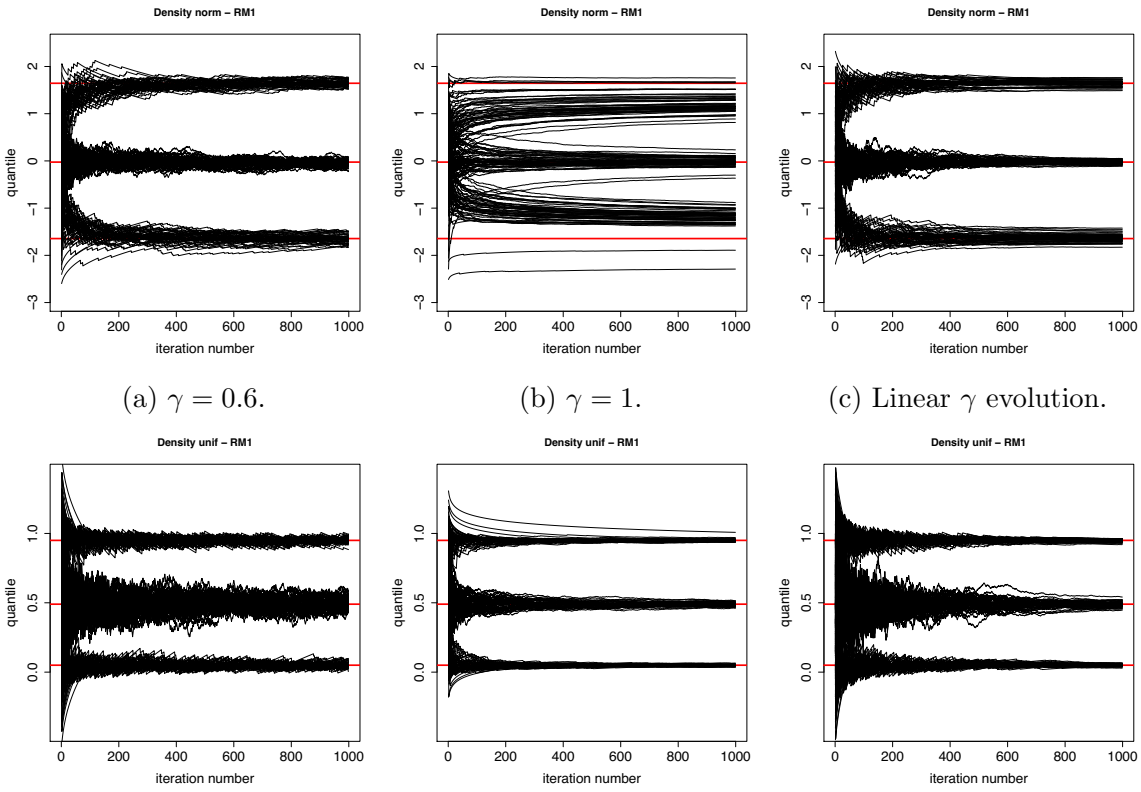
avec $q_\alpha(1) = Y_1$ (étape d'initialisation issue de la première donnée), $C > 0$ une constante et $\gamma \in]0,1]$ régissant la vitesse de descente de l'algorithme stochastique. A taille d'échantillon N finie, l'estimateur RM du α-quantile de Y est donc $\hat{q}_\alpha = q_\alpha(N)$. Cet estimateur est consistant et asymptotiquement gaussien pour $\gamma \in]0.5,1]$ (Duflo, 1997). La valeur de γ ne semble donc pas d'une importance cruciale mais, pour des N peu élevés, nous allons

voir dans la section 2 que son réglage est important. La section 3 discute du réglage de la constante C qui est primordial pour que le pas de descente soit d'amplitude convenable. La section 4 présente finalement une version moyennée de l'algorithme RM.

2 Réglage robuste de γ

Nous cherchons une valeur de γ qui donne des résultats "acceptables" quelle que soit la distribution de Y (inconnue en pratique). Notre test numérique considère les cas $Y \sim \mathcal{N}(0,1)$ et $Y \sim \mathcal{U}[0,1]$, avec $N = 1000$, $C = 1$ et trois ordres de quantile α (0.05, 0.5 et 0.95). Pour chacun de ces cas, la Figure 1 montre 50 trajectoires indépendantes de l'estimateur RM $q_\alpha(n)$ pour $n = 1, \ldots, N$ en considérant trois étalonnages différents de γ : 0.6, 1 et une variation linéaire en fonction de n qui s'écrit

$$\gamma(n) = 0.5 + 0.5\frac{n-1}{N-1} \, . \tag{3}$$



(a) $\gamma = 0.6$. (b) $\gamma = 1$. (c) Linear γ evolution.

(d) $\gamma = 0.6$. (e) $\gamma = 1$. (f) Linear γ evolution.

Figure 1: Simulations de trajectoires de l'algorithme RM ($N = 1000$, haut : $Y \sim \mathcal{N}(0,1)$, bas : $Y \sim \mathcal{U}[0,1]$). Les lignes rouges donnent les quantiles exacts d'ordre 0.05, 0.5 et 0.95.

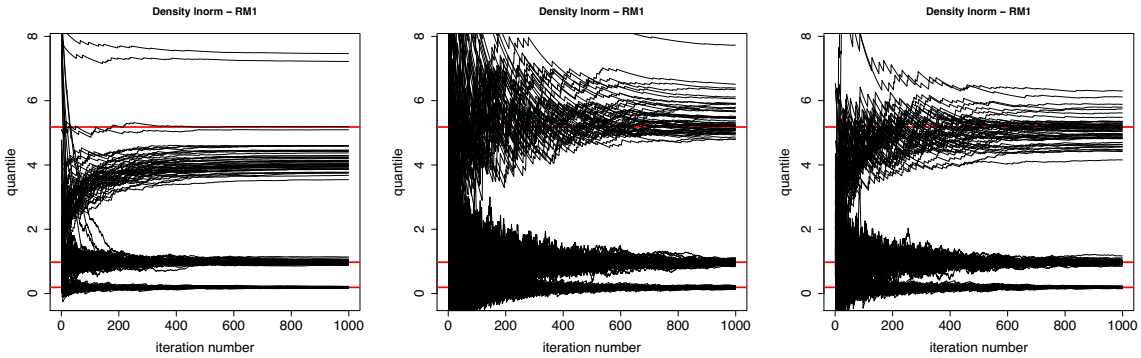
L'idée du profil de $\gamma(n)$ donné par l'Eq. (3) est d'avoir des fluctuations fortes de l'estimateur au début de l'algorithme (pour effacer sa dépendance aux valeurs de Y tirées en premier) puis des fluctuations faibles en fin d'algorithme (pour stabiliser l'estimateur aux dernières itérations). En effet, nous pouvons constater que les fluctuations avec $\gamma = 1$ sont trop faibles dans le cas gaussien ($\gamma = 0.6$ est satisfaisant dans ce cas-là) et les fluctuations avec $\gamma = 0.6$ sont trop fortes dans le cas uniforme ($\gamma = 1$ est satisfaisant dans ce cas-là). Le profil d'une variation linéaire de γ réalise un compromis entre ces deux cas extrêmes (et dans les nombreux autres tests réalisés). Par ailleurs, les propriétés théoriques asymptotiques de l'algorithme RM sont conservées avec le réglage de l'Eq. (3).

3 Réglage robuste de C

Dans la section précédente, la constante C a été fixée à 1. Ce choix s'avère catastrophique lorsque la variable considérée a une dispersion qui n'est pas de cet ordre de grandeur. Il faut rappeler qu'en pratique cette dispersion est inconnue. La Figure 2 montre 50 trajectoires indépendantes de l'estimateur RM $q_\alpha(n)$ pour $n = 1, \ldots, 1000$ et Y de loi lognormale ($\log(Y) \sim \mathcal{N}(0,1)$). γ est de profil linéaire et trois étalonnages différents de C sont testés : 1, 10 et un réglage adaptatif qui s'écrit

$$C(n) = |q_{\alpha_{\max}}(n-1) - q_{\alpha_{\min}}(n-1)| \,, \tag{4}$$

où $n \geq 2$ et $C(1) = |Y_2 - Y_1|$. Sur la Figure 2, il est clair que, pour le quantile d'ordre 0.95, C doit être suffisamment grand pour que les fluctuations soient importantes dès le début de l'algorithme RM. Le réglage adaptatif de C via l'Eq. (4) permet de réguler automatiquement ces fluctuations. De nombreux autres tests numériques sur des distributions de différents types ont permis de confirmer la justesse de ce choix.



(a) $C = 1$. (b) $C = 10$. (c) C adaptatif.

Figure 2: Simulations de trajectoires de l'algorithme RM ($N = 1000$, $Y \sim \mathcal{LN}(0,1)$). Les lignes rouges donnent les quantiles exacts d'ordre 0.05, 0.5 et 0.95.

4 Version moyennée de Robbins-Monro

Il est connu que la version moyennée de RM (notée ici RMM) converge plus rapidement que celle de l'Eq. (2). Nous avons cependant constaté que, si le quantile moyenné est introduit dans (2), les fluctuations de l'estimateur le long des itérations ne sont pas d'ampleur suffisante pour converger vers la valeur exacte. Il faut donc conserver la formulation (2) pour $q_\alpha(n)$ et stocker en plus, à chaque itération, l'estimateur moyenné (noté $\bar{q}_\alpha(n)$) :

$$\bar{q}_\alpha(n+1) = \bar{q}_\alpha(n) + \frac{q_\alpha(n+1) - \bar{q}_\alpha(n)}{n+1} \,, \tag{5}$$

avec $n \geq 1$ et $\bar{q}_\alpha(1) = Y_1$.

La Figure 3 compare les algorithmes RM et RMM pour $Y \sim \mathcal{N}(0,1)$, $N = 1000$ et le réglage adaptatif de C (cf. Eq. (4)) . Les quantiles sont estimés pour des ordres α discrétisés dans l'intervalle $[0.05, 0.95]$ par pas de 0.01. La métrique utilisée (en ordonnée) est l'erreur quadratique moyenne entre les quantiles exacts et les quantiles estimés. Les estimations sont répétées 100 fois de manière indépendante afin de capturer la variabilité des erreurs due à l'échantillonnage. L'estimateur de référence est l'estimateur empirique (qui n'est pas itératif). Sur cet exemple, les performances de RM et RMM avec un γ de profil linéaire sont similaires et proches de celles de l'estimateur empirique. Un γ constant et faible (égal à 0.6) donne des résultats encore meilleurs avec RMM (mais pas avec RM). En fait, la moyennisation dans RMM (qui fait converger plus rapidement l'estimateur du quantile) rend inutile l'augmentation du γ vers 1 que l'on a avec le profil linéaire. D'autres tests avec différentes distributions, non montrés ici, présentent des conclusions similaires.

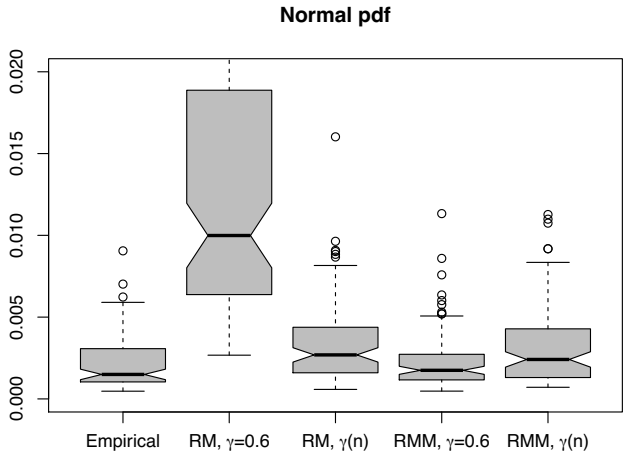


Figure 3: Erreurs quadratiques moyennes des fonctions quantiles discrétisées pour les estimateurs empirique, RM et RMM. $\gamma(n)$ est le réglage de γ en profil linéaire.

5 Conclusions

Ce travail a permis de dégager quelques heuristiques pour l'estimation itérative de quantile par l'algorithme RM avec un échantillon de faible taille. Le choix d'un C adaptatif est bénéfique dans tous les cas et le choix d'un γ de profil linéaire est robuste et doit être privilégié pour RM. Par contre, pour l'algorithme RMM, γ faible donne de meilleurs résultats. Enfin, l'utilisation de séquences de points bien réparties au lieu d'échantillons i.i.d (tests non montrés ici) permettent d'améliorer la précision des estimateurs de manière drastique, avec γ de profil linéaire et C adaptatif. Cette idée semble judicieuse et sera étudiée en profondeur dans le cas où la variable Y provient d'un modèle dont les entrées sont de grande dimension et où le choix d'un bon plan d'expériences (de type "space filling design") est important. Dans le même ordre d'idée, il sera fructueux de combiner l'algorithme RM et les techniques de simulation d'événements rares (cf. e.g. Kohler et al., 2014). Une autre perspective majeure de ce travail sera d'avoir accès à un intervalle de confiance sur le quantile estimé, quantité indispensable dans les applications pratiques.

6 Remerciements

Ce travail émane en partie du projet ANR international INDEX (ANR-18-CE91-0007) dédiés aux plans d'expériences incrémentaux. L'auteur remercie Luc Pronzato, Bernard Bercu, Alejandro Ribés et Clément Gauchy pour de fructueuses discussions.

Bibliographie

Duflo, M. (1997). *Random iterative models*. Springer, Berlin.
Iooss, B. and Marrel, A. (2019). Advanced methodology for uncertainty propagation in computer experiments with large number of inputs. *Nuclear Technology*, 205:1588–1606.
Kohler, M., Krzyżak, A., and Walk, H. (2014). Nonparametric recursive quantile estimation. *Statistics and Probability Letters*, 93:102–107.
Ribés, A., Terraz, T., Fournier, Y., Iooss, B., and Raffin, B. (2019). Large scale in transit computation of quantiles for ensemble runs. *Preprint, https://hal.inria.fr/hal-02016828*.
Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, 22:400–407.
Smith, R. (2014). *Uncertainty quantification*. SIAM.
Terraz, T., Ribes, A., Fournier, Y., Iooss, B., and Raffin, B. (2017). Large scale in transit global sensitivity analysis avoiding intermediate files. In *Proceedings the International Conference for High Performance Computing, Networking, Storage and Analysis (Supercomputing)*, Denver, USA.
Tierney, L. (1983). A space-efficient recursive procedure for estimating a quantile of an unknown distribution. *SIAM Journal on Scientific and Statistical Computing*, 4:706–711.

missKnockoff – controlled variable selection with missing values

Wei Jiang [1], Julie Josse [1] & Malgorzata Bogdan [2]

[1] *CMAP, École Polytechnique and Inria XPOP, France*
[2] *University of Wroclaw, Poland and Lund University, Sweden*

Abstract. Model selection with high-dimensional data becomes an important issue in the last two decades. With the presence of missing data, only a few methods are available to select a model. We propose a novel approach – missKnockoff, which stands for knockoff with missing values. The advantage of knockoff is that the model selection relies on the modeling of covariates, without requirement to know how the outcome depends on them. We propose multiple imputation before generating knockoff copies, and present several ways to aggregate the supports. Through extensive simulations, we demonstrate the performance in terms of power and FDR under a wide range of scenarios. Finally, we analyze a real dataset consisting of patients from Paris hospitals who underwent severe trauma.

Keywords. incomplete data, FDR control, multiple imputation

1 Introduction

1.1 Problem statement

In the contemporary statistical analysis, challenges involves a large quantity of covariates and nonlinear models: we consider a supervised learning setting with potential explanatory variables $X = (X_1, X_2, \cdots, X_p)$ and response variable Y. We then have access to n *i.i.d.* realizations $(X_{i1}, X_{i2}, \cdots, X_{ip})_{i\in[n]} \sim P_X$ and $(Y_i)_{i\in[n]} \sim P_{Y|X}$. Following classical statistical procedure, one may want to recover $P_{Y|X}$ by estimating the parameters in a given model, for example a linear regression. However, this estimation can be strongly biased considering a high-dimensional setting where p is large and even larger than n. But in most cases, we believe that Y only depends upon a small fraction of the entire variables. For example, in molecular genetics, a large number of genetic predictors is available but only a few are relevant for explaining a certain biological phenomena. The objective is to identify a subset $\mathcal{S} \subset [p]$ indexing important variables; in other words, variable X_j with $j \in \mathcal{H}_0 := [p]/\mathcal{S}$ is null, if $Y \perp\!\!\!\perp X_j \mid X_{-j}$, where X_{-j} denotes the remaining $p-1$ variables excluding X_j.

Candes et al. (2018) propose a framework knockoff in order to address the variable selection problem. They assume *no knowledge* of the conditional distribution $P_{Y|X}$ but the joint distribution of the covariates P_X is known. Therefore X is modeled and Y is

not, which is in contrast with the classical approach. Cases in which these assumptions hold can include, for instance,

- The distribution of covariates is known, such as genetic case-control studies (Consortium et al., 2007), and sensitivity analysis of numerical climate models (Saltelli et al., 2008).
- We have semi-supervised data: there is a large amount of unlabeled data in addition to the n labeled observations.

1.2 False discovery control

When selecting variables, we aim at controlling the False Discovery Rate (FDR) (Benjamini and Hochberg, 1995), which can be defined as follows: let $\hat{\mathcal{S}}$ be a model selection outcome through a certain procedure, then:

$$\text{FDR} \coloneqq \mathbb{E}\left[\frac{|\hat{\mathcal{S}} \cap \mathcal{H}_0|}{|\hat{\mathcal{S}}|}\right] = \mathbb{E}\left[\frac{\text{number of false positives}}{\text{number of selected variables}}\right],$$

where $|\cdot|$ denotes the length of a set. For the parametric model such as penalized linear regression, one available method is SLOPE (Bogdan et al., 2015) which penalizes larger coefficients more stringently. Differently, based on the assumption that we are capable to model X rather than Y, Barber et al. (2019); Candes et al. (2018) suggest a methodology named as Knockoff. Intuitively, Knockoff first generates a set of "fake" variables that depend on the original covariates and mimic its correlation structure. Then it returns true variables which are clearly more important than their knockoff copies according to some feature importance measures.

2 Knockoff with missing data

In the high-dimensional dataset, apart from the issue of model selection, the problematic of missing data can be ubiquitous. For example, genetic data obtained from microarray experiments often contain missing values for several reasons: insufficient resolution, image corruption, manufacturing errors, etc. The most common practice for dealing with missing data, listwise deletion, leads to estimation bias unless the missingness is randomly generated, and often results in significant information loss, especially for large data. The literature of missing values is abundant but only few methods are available to select a model when values are missing.

Throughout this paper, we assume the MAR (Missing at random) mechanism (Little and Rubin, 2002; Rubin, 1976) which implies that the missing data depends only on the observed variables, and the missing values mechanism can therefore be ignored when maximizing the likelihood (Little and Rubin, 2002).

In this paper, we propose a new methodology – missKnockoff, which stands for knockoff with missing values. we propose multiple imputation before generating knockoff copies, and the algorithm is composed of five stages:

1. Sample multiple plausible values for the missing elements;
2. Sample one set of knockoff copies based on each completed dataset;
3. For each pair of imputation and knockoff set, perform a supervised learning algorithm, such as cross-validation LASSO on response, then obtain fitted coefficient vectors and statistics;
4. Aggregate the statistics based on multiple knockoff (Holden and Hellton, 2018; Gimenez and Zou, 2018; Nguyen et al., 2019);
5. Calculate FDR threshold and decide whether to reject the variables.

Figure 2 illustrates the whole procedure of processing missing values and model selection via multiple knockoff. Through extensive simulations, we demonstrate the performance in terms of power and FDR control under a wide range of scenarios.

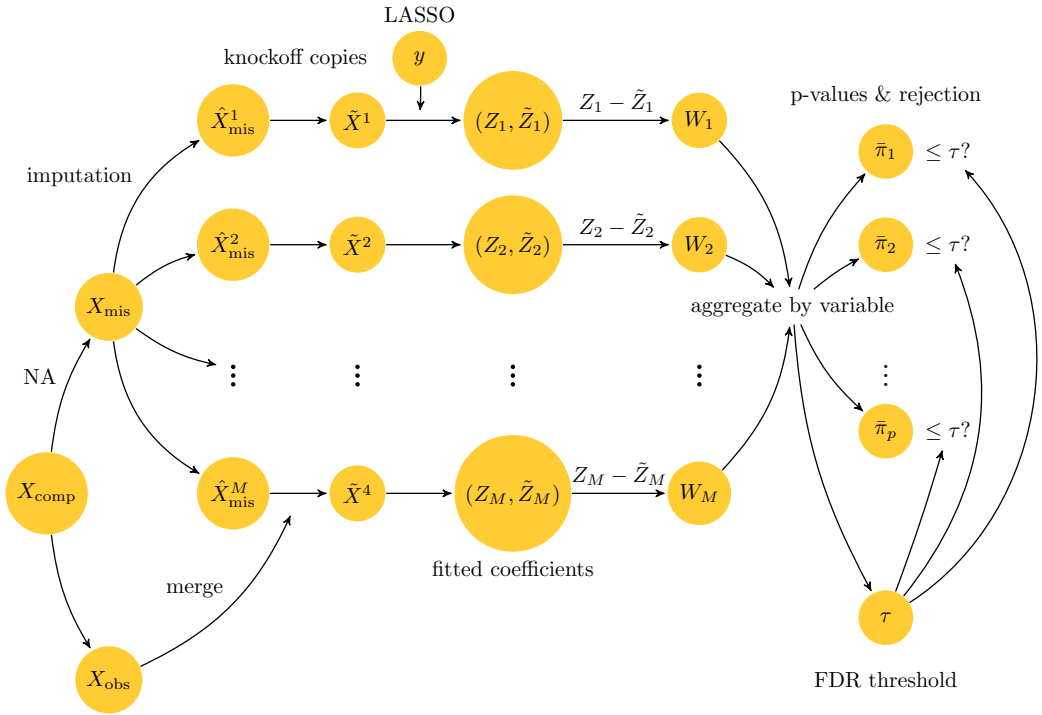


Figure 1: Diagram of stages for handling missing values in model selection via Knockoff.

References

Barber, R. F., Candès, E. J., et al. (2019). A knockoff filter for high-dimensional selective inference. *The Annals of Statistics*, 47(5):2504–2537.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

Bogdan, M., van den Berg, E., Sabatti, C., Su, W., and Cands, E. J. (2015). SLOPE—adaptive variable selection via convex optimization. *Ann. Appl. Stat.*, 9(3):1103–1140.

Candes, E., Fan, Y., Janson, L., and Lv, J. (2018). Panning for gold : Model-x knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577.

Consortium, W. T. C. C. et al. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661.

Gimenez, J. R. and Zou, J. (2018). Improving the stability of the knockoff procedure: Multiple simultaneous knockoffs and entropy maximization. *arXiv preprint arXiv:1810.11378*.

Holden, L. and Hellton, K. (2018). Multiple model-free knockoffs.

Little, R. J. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc.

Nguyen, B., Chevalier, J.-A., Thirion, B., and Arlot, S. (2019). Aggregation of multiple knockoffs. *preprint*.

Romano, Y., Sesia, M., and Candès, E. J. (2018). Deep knockoffs. *arXiv preprint arXiv:1811.06687*.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., and Tarantola, S. (2008). *Global sensitivity analysis: the primer*. John Wiley & Sons.

First order Sobol indices for physical models via inverse regression

Benoit KUGLER [1] & Florence FORBES [1] & Sylvain DOUTE [2]

[1] *Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LJK, 38000 Grenoble, France (firstname.lastname@inria.fr)*
[2] *Univ. Grenoble Alpes, CNRS, IPAG, 38000 Grenoble, France (sylvain.doute@univ-grenoble-alpes.fr)*

Résumé. Dans un contexte d'inversion Bayésienne de modèle physique, on souhaite effectuer une analyse de sensibilité pour comprendre et ajuster le modèle. Pour ce faire, on introduit des indicateurs inspirés des indices de Sobol, mais visant le modèle inverse. Comme le modèle inverse n'a en général pas d'expression analytique, on propose d'utiliser un modèle paramétrique pour l'approximer. Les paramètres de ce modèle peuvent être estimés par un algorithme EM. On peut ensuite exploiter l'expression analytique de la postérieur par une intégration numérique de type Monte-Carlo, ce qui permet une estimation efficace de ces pseudo indices de Sobol.

Mots-clés. Analyse de sensibilité, indices de Sobol, problème inverse, régression, apprentissage statistique

Abstract. In a bayesian inverse problem setting, we aim at performing sensitivity analysis to help understand and adjust the physical model. To do so, we introduce indicators inspired by Sobol indices but focused on the inverse model. Since this inverse model is not generally available in closed form, we propose to use a parametric surrogate model to approximate it. The parameters of this model may be estimated via standard EM inference. Then we can exploit its tractable form and perform Monte-Carlo integration to efficiently estimate these pseudo Sobol indices.

Keywords. Sensitivity analysis, Sobol indices, inverse problem, regression, statistical learning

1 Introduction

A wide class of problems from medical imaging [Mesejo et al., 2016, Lemasson et al., 2016, Nataraj et al., 2018] to astrophysics [Deleforge et al., 2015, Schmidt and Fernando, 2015] can be formulated as inverse problems [Tarantola, 2005, Giovannelli and Idier, 2015]. An inverse problem refers to a situation where one aims at determining the causes of a phenomenon from experimental observations of its effects. Such a resolution generally starts by the so-called *direct or forward* modelling of the phenomenon. It theoretically

describes how input parameters $\mathbf{x}$ are translated into effects $\mathbf{y}$. Then from experimental observations of these effects, the goal is to find the parameters values that best explain the observed measures.

Typical features or constraints that can occur in practice are that 1) both direct and inverse relationships are (highly) non-linear, *e.g.* the direct model is available but is a (complex) series of ordinary differential equations as in [Mesejo et al., 2016, Hovorka et al., 2004]; 2) the observations $\mathbf{y}$ are high-dimensional because they represent signals in time or spectra, as in [Schmidt and Fernando, 2015, Bernard-Michel et al., 2009, Ma et al., 2013]; 3) many such high-dimensional observations are available and the application requires a very large number of inversions, *e.g.* [Deleforge et al., 2015, Lemasson et al., 2016]; 4) the parameters $\mathbf{x}$ to be predicted is itself multi-dimensional with correlated dimensions so that predicting its components independently is sub-optimal, *e.g.* when there are known constraints such as their sum is one like for concentrations or probabilities, [Deleforge et al., 2015, Bernard-Michel et al., 2009].

A common issue when dealing with inverse problems is to be sure that the problem is well defined. In this regard some natural questions arise. Is the direct model one-to-one ? And if it's the case, are the output sensitive enough to the parameters ? If not, small noise in the observations may lead to high errors in predictions (high variance in probabilistic settings), making the model barely usable in practice. To answer these questions, one may use Sensitivity Analysis, which aims at providing qualitative or quantitative indicators on the variation of the output $\mathbf{y}$ with respect to the input $\mathbf{x}$. Sensitivity analysis may be useful on its own, to gain inner knowledge on the forward model. More over, in a setting where one has freedom to select the forward model, it can be used to enhance the quality of the inversion.

Among various methods, we focus in this paper on Sobol sensitivity analysis, applied to an inverse problem setup. We aim at computing what we call pseudo Sobol indices for the inverse model. We propose to use a surrogate model, estimated via a regression approach and exploited via numerical integration.

2 Sobol indices for an inverse problem

We start by specifying the notations for our inverse problem, before introducing the so-called Gaussian Locally-Linear Mapping model (GLLiM) ([Deleforge et al., 2015]) as a surrogate model, and proposing an efficient way to estimate inverse Sobol indices.

2.1 Context

The parameters and observations are assumed to be random variables $\mathbf{X} \in \mathbb{R}^L$ and $\mathbf{Y} \in \mathbb{R}^D$ of dimension L and D respectively where D is usually much greater than L. The forward model is then described by a likelihood function linking parameters values $\mathbf{x}$ to

the probability of observing some $\mathbf{y}$ and denoted by $\mathcal{L}_{\mathbf{x}}(\mathbf{y}) = p(\mathbf{y} \mid \mathbf{X} = \mathbf{x})$. We will further assume that the relationship between $\mathbf{x}$ and $\mathbf{y}$ is described by a known function F and that the uncertainties on the theoretical model are independent on the input parameter $\mathbf{x}$. In other words,

$$\mathbf{Y} = F(\mathbf{X}) + \epsilon \tag{1}$$

where ϵ is a random variable. For instance ϵ can be assumed to be a centered Gaussian variable with covariance matrix $\mathbf{\Sigma}$, so that $\mathcal{L}_{\mathbf{x}}(\mathbf{y}) = \mathcal{N}(\mathbf{y}; F(\mathbf{x}), \mathbf{\Sigma})$, where $\mathcal{N}(\ .\ ; F(\mathbf{x}), \mathbf{\Sigma})$ denotes the Gaussian pdf with mean $F(\mathbf{x})$ and covariance $\mathbf{\Sigma}$. We denote the prior density on $\mathbf{X}$ by $p(\mathbf{x})$.

Sensitivity analysis often deals with the forward model. Thus, first order Sobol indices are defined as

$$S_{l,d} = \frac{Var[\mathbb{E}[Y_d|X_l]]}{Var[Y_d]} \tag{2}$$

As we will show later on, our approach also yields estimation of Sobol indices for F. However, we mainly propose in this paper to study the sensitivity of the inverse model. It's challenging because it requires the expression of F^{-1} which is not available. Reversing the role of X and Y, we propose to define the inverse Sobol indices as

$$S^*_{d,l} = \frac{Var[\mathbb{E}[X_l|Y_d]]}{Var[X_l]} \tag{3}$$

We propose in the next section a model which enables the computation of S^* by exploiting an explicit density.

2.2 Surrogate model

We propose to use a learning approach. We approximate $(\mathbf{X}, \mathbf{Y})$ by a Gaussian Locally-Linear Mapping model (GLLiM) which builds upon Gaussian Mixture Models to approximate non linear functions ([Deleforge et al., 2015]). This is modeled by introducing a latent variable $Z \in \{1, \dots, K\}$ such that

$$\mathbf{Y} \ = \ \sum_{k=1}^{K} \mathbb{1}_{Z=k}(\boldsymbol{A}_k\mathbf{X} + \boldsymbol{b}_k + \boldsymbol{\epsilon}_k) \tag{4}$$

where $\mathbb{1}$ is the indicator function, $\boldsymbol{A}_k$ a $D \times L$ matrix and $\boldsymbol{b}_k$ a vector of $\mathbb{R}^D$ that define an affine transformation. Variable $\boldsymbol{\epsilon}_k$ corresponds to an error term which is assumed to be zero-mean and not correlated with $\mathbf{X}$ capturing both the observation noise and the reconstruction error due to the affine approximation.

In order to keep the posterior tractable, we assume that $\boldsymbol{\epsilon}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_k)$ and $\mathbf{X}$ is a mixture of K Gaussians : $p(\mathbf{x}|Z = k) = \mathcal{N}(\mathbf{x}; \boldsymbol{c}_k, \mathbf{\Gamma}_k)$ and $p(Z = k) = \pi_k$. The GLLiM model is thus characterized by the parameters $\boldsymbol{\theta} = \{\pi_k, \boldsymbol{c}_k, \mathbf{\Gamma}_k, \boldsymbol{A}_k, \boldsymbol{b}_k, \mathbf{\Sigma}_k\}_{k=1:K}$

This model can be learned against a training set, sampled along the prior on $\mathbf{X}$ and the likelihood defined in (1), via an EM algorithm. More specifically, we sample a dictionary $(\mathbf{x}_n, \mathbf{y}_n)_{n=1..N}$ where $\mathbf{x}_n$ are realizations of the prior $p(\mathbf{x})$ and $\mathbf{y}_n = F(\mathbf{x}_n) + \boldsymbol{\epsilon}_n$. We then run the EM algorithm on $(\mathbf{x}_n, \mathbf{y}_n)_{n=1..N}$ to estimate $\boldsymbol{\theta}$ and use the resulting GLLiM distribution denoted by p_G (and depending on $\boldsymbol{\theta}$) as a surrogate model for the pdf of $(\mathbf{X}, \mathbf{Y})$

2.3 Sobol indices for the GLLiM model

We have shifted the computation of Sobol indices from the real model to an approximated one, and in this section $(\mathbf{X}, \mathbf{Y})$ will thus denote random variables following the GLLiM distribution. The purpose is to exploit the tractable density p_G given by the GLLiM model. Indeed, from p_G, the conditional distribution is available in closed form :

$$p_G(\mathbf{x}|\mathbf{Y} = \mathbf{y}, \boldsymbol{\theta}) = \sum_{k=1}^{K} w_k^*(\mathbf{y}) \mathcal{N}(\mathbf{x}; \boldsymbol{A}_k^* \mathbf{y} + \boldsymbol{b}_k^*, \boldsymbol{\Sigma}_k^*) \tag{5}$$
$$\text{with } w_k^*(\mathbf{y}) = \frac{\pi_k \mathcal{N}(\mathbf{y}; \boldsymbol{c}_k^*, \boldsymbol{\Gamma}_k^*)}{\sum_{j=1}^{K} \pi_j^* \mathcal{N}(\mathbf{y}; \boldsymbol{c}_j^*, \boldsymbol{\Gamma}_j^*)}$$

A new parametrization $\boldsymbol{\theta}^* = \{\boldsymbol{c}_k^*, \boldsymbol{\Gamma}_k^*, \boldsymbol{A}_k^*, \boldsymbol{b}_k^*, \boldsymbol{\Sigma}_k^*\}_{k=1:K}$ is used to illustrate the similarity between the two conditional distributions. The parameters $\boldsymbol{\theta}^*$ are easily deduced from $\boldsymbol{\theta}$ as follows:

$$\begin{aligned}
\boldsymbol{c}_k^* &= \boldsymbol{A}_k \boldsymbol{c}_k + \boldsymbol{b}_k \\
\boldsymbol{\Gamma}_k^* &= \boldsymbol{\Sigma}_k + \boldsymbol{A}_k \boldsymbol{\Gamma}_k \boldsymbol{A}_k^\top \\
\boldsymbol{\Sigma}_k^* &= \left(\boldsymbol{\Gamma}_k^{-1} + \boldsymbol{A}_k^\top \boldsymbol{\Sigma}_k^{-1} \boldsymbol{A}_k\right)^{-1} \\
\boldsymbol{A}_k^* &= \boldsymbol{\Sigma}_k^* \boldsymbol{A}_k^\top \boldsymbol{\Sigma}_k^{-1} \\
\boldsymbol{b}_k^* &= \boldsymbol{\Sigma}_k^* \left(\boldsymbol{\Gamma}_k^{-1} \boldsymbol{c}_k - \boldsymbol{A}_k^\top \boldsymbol{\Sigma}_k^{-1} \boldsymbol{b}_k\right)
\end{aligned} \tag{6}$$

To compute $\mathbb{E}[X_l|Y_d]$ as required in (3), we observe that $(\mathbf{X}, Y_d)$ still follows a GLLiM distribution, with $D = 1$ and

$$\begin{aligned}
\boldsymbol{A}_k^{(d)} &:= \boldsymbol{A}_k[d, :] && \text{(row } d\text{)} \\
\boldsymbol{b}_k^{(d)} &:= \boldsymbol{b}_k[d] && \text{(coefficient } d\text{)} \\
\boldsymbol{\Sigma}_k^{(d)} &:= \boldsymbol{\Sigma}_k[d, d] && \text{(coefficient in row and column } d\text{)}
\end{aligned} \tag{7}$$

Since (5) is a Gaussian mixture, $\mathbb{E}[X_l|Y_d = y_d]$ is easy to compute from (5), with $\boldsymbol{\theta}$ adjusted as in (7) :

$$f_d(y_d) := \mathbb{E}[X_l|Y_d = y_d] = \sum_{k=1}^{K} w_k^*(y_d) \left(\boldsymbol{A}_k^* y_d + \boldsymbol{b}_k^*\right) \tag{8}$$

Unfortunately, the variance of $f_d(Y_d)$ is not straightforward. Thus, we propose to compute it via Monte-Carlo integration. It can be shown that $\mathbf{Y}$ follows a Gaussian mixture model, with parameters $(\pi_k, \boldsymbol{c}_k^*, \boldsymbol{\Gamma}_k^*)_{k=1..K}$: $p_G(\mathbf{y}|\boldsymbol{\theta}) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{y}; \boldsymbol{c}_k^*, \boldsymbol{\Gamma}_k^*)$, from which we can efficiently sample.

Finally, since $\mathbf{X}$ is also a mixture of Gaussian distributions, its variance has a closed form :

$$Cov(\mathbf{X}) = \sum_{k=1}^{K} \pi_k \left[\boldsymbol{\Gamma}_k + \boldsymbol{c}_k \boldsymbol{c}_k^\top\right] - \left(\sum_{k=1}^{K} \pi_k \boldsymbol{c}_k\right)\left(\sum_{k=1}^{K} \pi_k \boldsymbol{c}_k\right)^\top$$

To sum up, given a GLLiM model characterized by $\boldsymbol{\theta}$, we propose to compute the inverse Sobol indices $S_{d,l}^*$ by sampling $(y_d^{(n)})_{n=1..N}$ according to its mixture of Gaussians distribution and estimating $v_d := Var(f_d(Y_d))$ with Monte-Carlo integration using samples $(y_d^{(n)})_{n=1..N}$. The index $S_{d,l}^*$ can be then computed by

$$S_{d,l}^* = \frac{v_d}{Cov(\mathbf{X})_{l,l}}$$

3 Illustration on a photometric model

We apply the proposed approach to the Hapke's model, a highly-non linear model used in remote sensing. It is a semi-empirical photometric model that relates physically meaningful parameters to the reflectivity of a granular material for a given geometry of illumination and viewing. A geometry denoted by G is described by three angles (θ_0, θ, ϕ). Thus, our forward model takes the form $F(\mathbf{x}) = (f_{hapke,G_1}(\mathbf{x}), ..., f_{hapke,G_D}(\mathbf{x}))$ where D is the number of geometries, and $\mathbf{x} = (\omega, \bar{\theta}, b, c)$ are the physical parameters. The exact expression of f_{hapke} may be found for example in [Schmidt and Fernando, 2015]. The figure 1 plots the inverse Sobol indices for the parameter ω, with respect to the geometries.

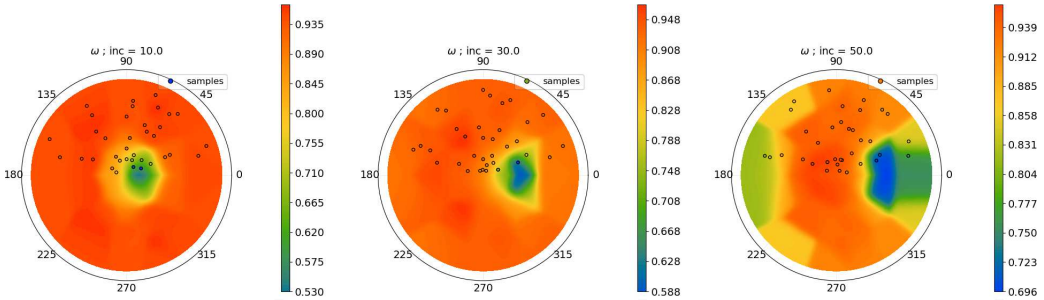


Figure 1: $D = 108$ geometries, regrouped according to the incidence θ_0. Radial coordinate is θ, polar angular coordinate is ϕ.

4 Conclusion

We proposed an efficient computation of sensitivity indicators inspired by Sobol indices using a surrogate model. We focused on the Sobol indices for inverse models. Regarding forward models, a surrogate model is not usually needed as the forward model is often available in closed form. Still, our approach also yields direct Sobol indices, by reversing the role of $\mathbf{X}$ and $\mathbf{Y}$. It could be used in a setup where only a dictionary of samples $(\mathbf{x}, \mathbf{y})$ is available, and not the functional model F.

So far, we have implicitly assumed that $\mathbf{Y}$ components where independent, which is a common assumption when using Sobol indices. If it is wrong, Sobol indices are still well defined, but the property of variance decomposition does not hold anymore. Studies on sensitivity analysis in the dependent case suggest to use generalized Sobol indices, which take into account the dependency between components. For instance, a future work could be to apply the methodology proposed in [Chastaing et al., 2014].

References

[Bernard-Michel et al., 2009] Bernard-Michel, C., Douté, S., Fauvel, M., Gardes, L., and Girard, S. (2009). Retrieval of Mars surface physical properties from OMEGA hyperspectral images using regularized sliced inverse regression. *Journal of Geophysical Research: Planets*, 114(E6).

[Chastaing et al., 2014] Chastaing, G., Prieur, C., and Gamboa, F. (2014). Generalized Sobol sensitivity indices for dependent variables: Numerical methods. *arXiv:1303.4372 [stat]*.

[Deleforge et al., 2015] Deleforge, A., Forbes, F., Ba, S., and Horaud, R. (2015). Hyper-Spectral Image Analysis with Partially-Latent Regression and Spatial Markov Dependencies. *IEEE Journal of Selected Topics in Signal Processing*, 9(6):1037–1048.

[Giovannelli and Idier, 2015] Giovannelli, J.-F. and Idier, J., editors (2015). *Regularization and Bayesian Methods for Inverse Problems in Signal and Image Processing: Giovannelli/Regularization and Bayesian Methods for Inverse Problems in Signal and Image Processing*. John Wiley & Sons, Inc., Hoboken, NJ, USA.

[Hovorka et al., 2004] Hovorka, R., Canonico, V., Chassin, L. J., Haueter, U., Massi-Benedetti, M., Federici, M. O., Pieber, T. R., Schaller, H. C., Schaupp, L., Vering, T., and Wilinska, M. E. (2004). Nonlinear model predictive control of glucose concentration in subjects with type 1 diabetes. *Physiological Measurement*, 25(4):905–920.

[Lemasson et al., 2016] Lemasson, B., Pannetier, N., Coquery, N., Boisserand, L. S. B., Collomb, N., Schuff, N., Moseley, M., Zaharchuk, G., Barbier, E. L., and Christen, T. (2016). MR Vascular Fingerprinting in Stroke and Brain Tumors Models. *Scientific Reports*, 6:37071.

[Ma et al., 2013] Ma, D., Gulani, V., Seiberlich, N., Liu, K., Sunshine, J. L., Duerk, J. L., and Griswold, M. A. (2013). Magnetic resonance fingerprinting. *Nature*, 495(7440):187–192.

[Mesejo et al., 2016] Mesejo, P., Saillet, S., David, O., Bénar, C., Warnking, J. M., and Forbes, F. (2016). A differential evolution-based approach for fitting a nonlinear biophysical model to fMRI BOLD data. *IEEE Journal of Selected Topics in Signal Processing*, 10(2):416–427.

[Nataraj et al., 2018] Nataraj, G., Nielsen, J.-F., Scott, C., and Fessler, J. A. (2018). Dictionary-Free MRI PERK: Parameter Estimation via Regression with Kernels. *IEEE Transactions on Medical Imaging*, 37(9):2103–2114.

[Schmidt and Fernando, 2015] Schmidt, F. and Fernando, J. (2015). Realistic uncertainties on Hapke model parameters from photometric measurements. *Icarus*, 260:73–93 (IF 2,84).

[Tarantola, 2005] Tarantola, A. (2005). *Inverse Problem Theory and Methods for Model Parameter Estimation*. Other Titles in Applied Mathematics. Society for Industrial and Applied Mathematics.

Partitionnement spectral et modèle à blocs stochastique dynamique : parcimonie et régularité

Nicolas Keriven[1] & Samuel Vaiter [2]

[1] *CNRS, GIPSA-lab.*
11 rue des Mathématiques, 38400 Saint-Martin-d'Hères, France.
nicolas.keriven@gipsa-lab.grenoble-inp.fr
[2] *CNRS, IMB, Université de Bourgogne.*
9 avenue Alain Savary, 21000 Dijon, France.
samuel.vaiter@u-bourgogne.fr

Résumé. Dans ce papier, nous analysons des variantes de l'algorithme classique de Partitionnement Spectral dans le cas d'un Modèle à Bloc Stochastique Dynamique (DSBM). Des résultats existants montrent que certaines analyses du cas statique s'étendent au cas dynamique dans le cas *relativement parcimonieux*, où le degré moyen croît logarithmiquement avec la taille du graphe. Les bornes d'erreur peuvent alors être améliorées lorsque le DSBM est suffisamment *régulier*, c'est-à-dire que les communautés ne changent pas trop au cours du temps. Nous améliorons cette analyse en établissant un lien entre la parcimonie et la régularité du modèle : plus le DSBM est régulier, plus il peut être parcimonieux. En particulier, une régularité raisonnable permet de traiter le cas *parcimonieux* (degré moyen borné). Nous étendons également notre analyse au Laplacien normalisé. Une conséquence de nos résultats est d'obtenir, à notre connaissance, la meilleure borne de concentration spectrale du Laplacien normalisé pour le SBM classique.

Mots-clés. Partitionnement Spectral, Modèle à Bloc Stochastique Dynamique, Graphes Parcimonieux

Abstract. In this paper, we analyse classical variants of the Spectral Clustering (SC) algorithm in the Dynamic Stochastic Block Model (DSBM). Existing results show that, in the relatively sparse case where the expected degree grows logarithmically with the number of nodes, guarantees in the static case can be extended to the dynamic case and yield improved error bounds when the DSBM is sufficiently *smooth* in time, that is, the communities do not change too much between two time steps. We improve over these results by drawing a link between the sparsity and the smoothness of the DSBM : the more regular the DSBM is, the more sparse it can be, while still guaranteeing consistent recovery. In particular, a mild condition on the smoothness allows to treat the *sparse* case with bounded degree. We also extend these guarantees to the normalized Laplacian, and as a by-product of our analysis, for the classical static SBM, we obtain to our knowledge the best spectral concentration bound available for the normalized Laplacian.

Keywords. Spectral Clustering, Dynamic Stochastic Block Model, Sparse Graph

1 Introduction

De nombreux phénomènes peuvent être modélisés par des réseaux évoluant dans le temps : les interactions au sein d'un réseau social, la diffusion de maladies infectieuses, ou encore la transmission de l'information dans un réseau de machines. Sur de tels *graphes dynamiques*, une tâche classique consiste à grouper les nœuds en communautés évoluant dans le temps, par exemple les utilisateurs d'un réseau social. Pour ce faire, les méthodes les plus utilisées sont des variantes du classique algorithme de *partitionnement spectral* (SC) adaptées pour le cas dynamique. Nous donnons dans cet article des garanties théoriques état-de-l'art pour deux variantes simples du SC, sous un modèle à blocs stochastiques (SBM) dynamique (DSBM). Des résultats plus généraux ainsi que les preuves peuvent être trouvés dans la version longue de l'article [4].

Il est connu qu'une quantité clé pour analyser les méthodes sur SBM est la probabilité moyenne de connection entre deux nœuds, que nous appellerons ici α_n pour un graphe à n nœuds. En particulier, le régime "dense" $\alpha_n \sim 1$ est le plus souvent trivial à analyser, tandis que le régime "parcimonieux" $\alpha_n \sim \frac{1}{n}$ est particulièrement complexe [1]. Dans [6], Lei et Rinaldo donnent des garanties fortes pour l'algorithme SC (sur matrice d'adjacence) dans le régime *relativement parcimonieux* $\alpha_n \sim \frac{\log n}{n}$, qui ne peuvent en particulier pas être obtenues dans le régime parcimonieux [1]. Leur technique de preuve est adaptée par la suite par Pensky et Zhang au cas dynamique [9], pour lequel une nouvelle quantité clé est la *régularité temporelle* ε_n : plus ε_n est faible, moins les communautés changent au cours du temps. Les auteurs montrent que, pour une adaptation simple de l'algorithme SC avec une matrice d'adjacence lissée, les garanties de Lei et Rinaldo peuvent être améliorées lorsque $\varepsilon_n = o\left(\frac{1}{\log n}\right)$. Néanmoins, ces travaux n'établissent pas de lien entre la parcimonie et la régularité temporelle.

Dans [4], nous montrons qu'il est effectivement possible d'obtenir des garanties établissant un lien entre α_n et ε_n. En particulier, pour une régularité suffisante $\varepsilon_n \sim \frac{1}{\log^2 n}$, il est possible d'obtenir les mêmes garanties que [9] *dans le cas parcimonieux* $\alpha_n \sim \frac{1}{n}$. Nous étendons également les résultats au SC sur Laplacien normalisé, qui est bien plus couramment utilisé en pratique, ainsi qu'à un DSBM sous forme de modèle de Markov à états cachés (HMM) introduit dans [12].

2 Modèles et notations

Un graphe à n nœud est représenté par une matrice d'adjacence symmétrique $A \in \{0,1\}^{n\times n}$, où $a_{ij} = a_{ji} = 1$ indique la présence d'une arête entre les nœuds i et j. La matrice des degrés est une matrice diagonale $D(A) = \operatorname{diag}\left(\left(\sum_j a_{ij}\right)_{i=1}^{n}\right)$, et le Laplacien normalisé du graphe est défini comme $L(A) = D(A)^{-\frac{1}{2}} A D(A)^{-\frac{1}{2}}$ (en supposant les degrés non-nuls par simplicité). La norme spectrale $\|\cdot\|$ est définie par $\|A\| = \sup_{\|x\|_2=1} \|Ax\|_2$.

SBM. Un SBM à K communautés est paramétré par une matrice d'appartenance $\Theta \in \{0,1\}^{n\times K}$ contenant un unique 1 sur chaque ligne, tel que $\Theta_{ik} = 1$ si le nœud i appartient à la communauté k, ainsi que par une matrice de probabilités de connection $B \in [0,1]^{K\times K}$. Les arêtes sont alors tirées indépendemment selon des lois de Bernoulli : pour le nœud i dans la communauté k et le nœud $j \neq i$ dans la communauté ℓ, on tire $a_{ij} \sim \mathrm{Ber}(b_{k\ell})$. On définit $P = \Theta B \Theta^\top$ la matrice contenant les probabilités de connection entre nœuds.

La matrice B contient typiquement des termes élevés sur sa diagonale et faible en dehors, pour représenter les probabilités de connection respectivement au sein d'une communauté et entre communautés. Une quantité cruciale est l'évolution de ces termes avec n, pour cela nous supposons que $B = \alpha_n B_0$, pour un **facteur de parcimonie** $\alpha_n \in (0,1)$. Par simplicité nous supposerons que B_0 contient 1 sur sa diagonale, et $\tau \in (0,1)$ ailleurs (des résultats plus généraux sont donnés dans [4]).

DSBM. Le DSBM est paramétré par des communautés $\Theta_1, \ldots, \Theta_t$ évoluant en fonction du temps, et une matrice B définie comme ci-dessus. Étant donnés les Θ_q, les matrices d'adjacence $A_q \sim \mathrm{SBM}(\Theta_q, B)$ sont tirées indépendemment selon un SBM classique. Par simplicité, la matrice B ainsi que n et K n'évoluent pas en fonction du temps. Nous considérons alors deux modèles possibles pour les Θ_q.

Dans le DSBM dit *déterministe*, noté D-DSBM, les Θ_q ne sont pas des variables aléatoires. Nous supposons alors qu'au plus $n\varepsilon_n$ nœuds changent de communauté entre deux étapes temporelles, pour un **facteur de régularité** ε_n (plus ε_n est faible, plus le modèle est régulier). Nous supposons également que la taille des communautés est toujours comprise entre $n_{\min}$ et $n_{\max}$ à chaque étape, et définissons $\bar{n}_{\min} \overset{\text{def.}}{=} n_{\min} + \tau n$, $\bar{n}_{\max} \overset{\text{def.}}{=} n_{\max} + \tau n$, et $\mu_B \overset{\text{def.}}{=} \frac{\bar{n}_{\max}}{\bar{n}_{\min}}$. On notera que, lorsque la taille des communautés est de l'ordre de $\frac{n}{K}$ et $\tau \sim \frac{1}{K}$, alors $\bar{n}_{\min}, \bar{n}_{\max} \sim \frac{n}{K}$. Pour un τ fixe en revanche, $\bar{n}_{\min}, \bar{n}_{\max} \sim n$.

Dans le deuxième modèle, noté M-DSBM, les Θ_q suivent un modèle de Markov selon lequel chaque nœud a, entre chaque étape temporelle, une probabilité $(1-\varepsilon_n)$ de rester dans la même communauté et une probabilité uniforme $\frac{\varepsilon_n}{K-1}$ d'aller dans n'importe quelle autre communauté, indépendemment des autres nœuds. Les A_q étant indépendantes conditionellement aux Θ_q, le M-DSBM est un HMM.

3 Partitionnement spectral

L'algorithme SC fonctionne de la manière suivante : étant donné une matrice W, qui est ici soit la matrice d'adjacence A soit le Laplacien $L(A)$, on construit une matrice $U \in \mathbb{R}^{n\times K}$ formée par les K vecteurs propres correspondant au K valeurs propres dominantes de W, et on applique l'algorithme des k-moyennes [7] sur les lignes de U pour obtenir $\hat{\Theta}$. S'il est connu que le problème exact des k-moyennes est NP-complet, il est possible d'approcher son objectif à un facteur multiplicatif $(1+\delta)$ près [5], ce que l'on considère ici.

Dans le cadre dynamique, nous nous plaçons à un instant t et considérons que notre

but est d'estimer Θ_t le plus précisément possible en exploitant les données observées $A_0, \dots, A_t$ (un but légèrement différent serait d'estimer tous les Θ_q simultanément). Une méthode simple [2, 11, 10] est d'appliquer l'algorithme SC à une version *temporellement lissée* de la matrice d'adjacence notée A_t^{smooth}, ou au Laplacien normalisé $L(A_t^{\text{smooth}})$. Dans la littérature, deux variantes ont principalement été étudiées :

$$A_t^{\text{unif}} = \frac{1}{r}\sum_{k=0}^{r-1} A_{t-k}, \qquad A_t^{\text{exp}} = (1-\lambda)A_{t-1}^{\text{exp}} + \lambda A_t \tag{1}$$

L'estimateur A_t^{unif} "uniforme" [9] est une simple moyenne sur les r dernières valeurs de A_t, tandis que l'estimateur A_t^{exp} "exponentiel" [2, 11] est défini récursivement avec un "facteur d'oubli" $\lambda \in (0,1)$, et initialisé à $A_0^{\text{exp}} = A_0$. Nous supposerons dans la suite que t est suffisamment grand pour que l'effet de l'initialisation soit négligeable (voir [4]).

Suivant [6], nous mesurons la qualité d'un partitionnement par l'erreur suivante : $E(\hat{\Theta}, \Theta) \overset{\text{def.}}{=} n^{-1} \min_{Q \in \mathcal{P}_K} \|\hat{\Theta}Q - \Theta\|_0$, où $\mathcal{P}_K$ est l'ensemble des matrices de permutation de $\{1, \dots, K\}$, et $\|\cdot\|_0$ compte le nombre d'éléments non-nuls. Il est possible de relier la performance de l'algorithme SC à une certaine concentration en *norme spectrale*.

Lemme 1 (Adapté de [6])**.** *Soit un SBM (Θ, B) avec des communautés de taille comprise entre $n_{\min}$ et $n_{\max}$ et $B = \alpha_n B_0$ définie comme ci-dessus. Soit $\hat{P}$ un estimateur de P (resp. $\hat{L}$ un estimateur de $L(P)$), et $\hat{\Theta}_P$ le résultat de l'algorithme SC appliqué à $\hat{P}$ (resp. $\hat{\Theta}_L$ le résultat de l'algorithme appliqué à $\hat{L}$). On a alors :*

$$E(\hat{\Theta}_P, \Theta) \lesssim \frac{(1+\delta)n_{\max}K}{nn_{\min}^2\alpha_n^2(1-\tau)^2}\|\hat{P} - P\|^2, \quad E(\hat{\Theta}_L, \Theta) \lesssim \frac{(1+\delta)n_{\max}\bar{n}_{\max}^2K}{nn_{\min}^2(1-\tau)^2}\|\hat{L} - L(P)\|^2 \tag{2}$$

4 Concentration de la matrice d'adjacence

Dans le cadre du SBM classique, Lei et Rinaldo [6] montrent que, dans le cas relativement parcimonieux $\alpha_n \gtrsim \frac{\log n}{n}$, avec forte probabilité la matrice d'adjacence concentre en $\|A - P\| \lesssim \sqrt{n\alpha_n}$. En particulier, dans le cas de communautés de tailles similaires, d'après le Lemme 1 l'erreur $E(\hat{\Theta}, \Theta)$ tend vers 0 lorsque $K = o(\sqrt{\log n})$.

Dans [9], Pensky et Zhang étudient le D-DSBM, toujours dans le cas relativement parcimonieux $\alpha_n \gtrsim \frac{\log n}{n}$. Ils définissent un facteur $\rho_n = \min(1, \sqrt{\alpha_n n \varepsilon_n})$ et montrent que, pour $r \sim \frac{1}{\rho_n}$, l'estimateur uniforme concentre en $\|A_t^{\text{unif}} - P_t\| \lesssim \sqrt{n\alpha_n\rho_n}$, ce qui améliore le résultat par rapport au cas statique dès lors que $\varepsilon_n = o(1/(n\alpha_n))$. Notre contribution principale est similaire, mais lie crucialement les facteurs de parcimonie et régularité.

Théorème 1. *On se place dans le cas du D-DSBM, pour t suffisamment grand. Soit $\rho_n \overset{\text{def.}}{=} \min(1, \sqrt{n\alpha_n\bar{n}_{\max}})$. On considère l'estimateur $A_t^{\text{smooth}} = A_t^{\text{unif}}$ avec $r \sim \frac{1}{\rho_n}$ ou $A_t^{\text{smooth}} = A_t^{\text{exp}}$ avec $\lambda \sim \rho_n$. Si*

$$\frac{\alpha_n}{\rho_n} \gtrsim \frac{\log n}{n}, \tag{3}$$

alors on a le résultat suivant : pour tout $\nu > 0$, il existe une constante C_ν telle que, avec probabilité $1 - n^{-\nu}$,

$$\|A_t^{\text{smooth}} - P_t\| \leqslant C_\nu \sqrt{n\alpha_n \rho_n} \tag{4}$$

Par rapport à [9], on remarque que le facteur ρ_n est légèrement amélioré en remplaçant n par $\bar{n}_{\max}$. Plus important, la condition (3) améliore l'hypothèse de relative parcimonie : lorsque le modèle est suffisamment régulier (ε_n petit), α_n peut également être plus faible. Par exemple, $\varepsilon_n \sim \frac{1}{\log^2 n}$ suffit pour atteindre le modèle parcimonieux $\alpha_n \sim \frac{1}{n}$.

On notera qu'il existe une conjecture sur le cas parcimonieux et $\varepsilon_n \sim 1$ avec une analyse inspirée de la physique statistique [3], similaire au cas parcimonieux statique [1]. Néanmoins, ces analyses ne s'appliquent pas au SC traditionnel ou au cas $K > 2$, et il est impossible d'obtenir les garanties fortes de consistance présentées ci-dessus.

La proposition ci-dessous montre que le M-DSBM donne le même résultat que le D-DSBM, mais que la régularité ne peut pas être trop forte dans ce cas. Néanmoins, la borne inférieure obtenue sur ε_n est très faible et n'affecte pas le cas $\varepsilon_n \sim \frac{1}{\log^2 n}$.

Proposition 1. *Le résultat du Théorème 1 est encore valide dans le cas du M-DSBM avec probabilité jointe $1 - n^{-\nu}$ sur les A_q et Θ_q, en remplaçant $\bar{n}_{\max}$ par n et sous la condition que $\varepsilon_n \gtrsim \sqrt{\frac{\log n}{n}}$.*

5 Concentration du Laplacien normalisé

La concentration spectrale du Laplacien normalisé a été moins abordée dans la littérature. À notre connaissance, la meilleure borne disponible [8] dans le cas SBM relativement parcimonieux est $\|L(A) - L(P)\| \lesssim \sqrt{\frac{\log n}{n\alpha_n}}$, et le cas dynamique n'a pas été traité.

Théorème 2. *On se place dans le cas du D-DSBM, pour t suffisamment grand. Soit $\rho_n = \min(1, \sqrt{n\alpha_n \bar{n}_{\max}})$. On considère l'estimateur $A_t^{\text{smooth}} = A_t^{\text{unif}}$ avec $r \sim \frac{1}{\rho_n}$ ou $A_t^{\text{smooth}} = A_t^{\text{exp}}$ avec $\lambda \sim \rho_n$. On a le résultat suivant : pour tout $\nu > 0$, il existe des constantes C_ν, C'_ν telles que, si*

$$\frac{\alpha_n}{\rho_n} \geqslant C'_\nu \mu_B \frac{\log n}{\bar{n}_{\min}}, \tag{5}$$

alors, avec probabilité $1 - n^{-\nu}$,

$$\|L(A_t^{\text{smooth}}) - L(P_t)\| \leqslant \frac{C_\nu \mu_B}{\bar{n}_{\min}} \sqrt{\frac{n\rho_n}{\alpha_n}} \tag{6}$$

Dans le cas de communautés de taille similaire, combinées avec le Lemme 1, la borne (6) donne un résultat similaire aux garanties obtenues avec la matrice d'adjacence (4) (il est néanmoins connu qu'en pratique, l'algorithme SC donne de meilleures performances avec le Laplacien normalisé). On remarquera également la condition de parcimonie (5)

qui est légèrement plus forte que (3), en particulier, l'ordre des quantifieurs vis-à-vis de la constante ν n'est pas le même. Dans le cas du SBM statique relativement parcimonieux, une conséquence du Théorème 2 est d'obtenir une borne en $\|L(A) - L(P)\| \lesssim \frac{1}{\sqrt{n\alpha_n}}$, ce qui est la meilleure borne à notre connaissance.

Bibliographie

[1] Emmanuel Abbe : Community detection and stochastic block models : recent developments. *J. Mach. Learn. Res.*, pages 1–86, 2018.

[2] Yun Chi, Xiaodan Song, Dengyong Zhou, Koji Hino et Belle L. Tseng : Evolutionary spectral clustering by incorporating temporal smoothness. *In KDD*, pages 153–162, 2007.

[3] Amir Ghasemian, Pan Zhang, Aaron Clauset, Cristopher Moore et Leto Peel : Detectability thresholds and optimal algorithms for community structure in dynamic networks. *Phys. Rev. X*, 6(3):1–9, 2016.

[4] Nicolas Keriven et Samuel Vaiter : Sparse and Smooth : improved guarantees for Spectral Clustering in the Dynamic Stochastic Block Model. *ArXiv preprint arXiv :2002.02892*, 2020.

[5] Amit Kumar, Yogish Sabharwal et Sandeep Sen : A simple linear time $(1 + \varepsilon)$-approximation algorithm for k-means clustering in any dimensions. *In FOCS*, pages 454–462, 2004.

[6] Jing Lei et Alessandro Rinaldo : Consistency of spectral clustering in stochastic block models. *Ann. Stat.*, 43(1):215–237, 2015.

[7] Stuart P. Lloyd : Least squares quantization in pcm. *IEEE Trans. Inf. Theory*, 28(2):129–137, 1982.

[8] Roberto Imbuzeiro Oliveira : Concentration of the adjacency matrix and of the laplacian in random graphs with independent edges, 2009. arXiv :0911.0600.

[9] Marianna Pensky et Teng Zhang : Spectral clustering in the dynamic stochastic block model. *Electron. J. Stat.*, 13(1):678–709, 2019.

[10] Kevin S. Xu : Stochastic block transition models for dynamic networks. *In AISTAT*, pages 1–23, 2015.

[11] Kevin S Xu, Mark Kliger et Alfred O Hero Iii : Evolutionary spectral clustering with adaptative forgetting factor. *In ICASSP*, pages 2174–2177, 2010.

[12] Tianbao Yang, Yun Chi, Shenghuo Zhu, Yihong Gong et Rong Jin : Detecting communities and their evolutions in dynamic social networks - a bayesian approach. *Mach. Learn.*, 82(2):157–189, 2011.

DeepLTRS: A Deep Latent Recommender System based on User Ratings and Comments

Dingge Liang [1] & Marco Corneli [1,2] & Charles Bouveyron [1] & Pierre Latouche [3]

[1] *Université Côte d'Azur, Inria, Cnrs, Laboratoire J.A. Dieudonné, Maasai research team*
[2] *Université Côte d'Azur, Maison de la Modélisation, de la Simulation et des Interactions*
[3] *Université de Paris, Laboratoire MAP5*

Abstract. We introduce the deep latent recommender system (deepLTRS) for providing users with high quality recommendations based on other user ratings but also comments. Our approach extends the standard variational auto encoder architecture associated with deep latent variable models in order to handle both ordinal entries and texts made by users about products. DeepLTRS assumes a latent representation of both users and products, allowing a natural visualisation of the positioning of users in relation to products. A desirable feature of our approach is its ability to forecast the most likely words that would be used by an user to review a new product. Numerical experiments on simulated data sets demonstrate that deepLTRS outperforms existing models, in particular in the context of extreme sparsity.

Keywords. recommender system, variational auto encoder, deep latent variable model

1 Introduction

With the development of information technology and the Internet industry, information overload has become a challenge for people in processing information. For users, how to quickly and accurately locate the content they need in the exponentially growing resources is very important and extremely challenging. For merchants, how to present the right items to users in time to promote transaction volume and economic growth is also a very difficult matter. Recommender systems are information filtering systems that can learn user interests based on user profile or historical behavior records, and predict the user rating or preference for a given item. In the last ten years, such systems have changed the way merchants communicate with users and strengthened the interactivity between users.

Due to the extreme sparsity of data sets, the task of completing a matrix of rating is challenging and many authors have proposed different strategies to tackle this problem. Introduced by Gopalan et al. (2013), Hierarchical Poisson Factorization (HPF) assumes that the matirx of user ratings follows a Poisson distribution with latent user preferences and latent item attributes as model parameters. Compared to HPF whose sparse model (absence of a rating) and response model (rating values) are tightly coupled, Hierarchical Compound Poisson Factorization (HCPF, Basbug and Engelhardt (2016)) allows to choose

the most appropriate response model from addictive exponential dispersion model family and better captures the relationship between sparsity and response. More recently, Coupled Compound Poisson Factorization (CCPF, Basbug and Engelhardt (2017)) proposed a more general framework capable of selecting an arbitrary data-generating model among different mixture models, matrix factorization models and linear regression models.

Unfortunately, the aforementioned models only considered user ratings and ignored the importance of reviews. In the paper of McAuley and Leskovec (2013), the Hidden Factors and Hidden Topics (HFT) model combined latent rating factors with latent review topics. It can predict top ten words that users are most likely to use in comments. However, when a large amount of user ratings are missing, the performance of the prediction turns out to be limited.

Combining the advantages of dealing with extreme sparseness and considering the impact of user reviews on ratings, we proposed the deep latent recommender system (deepLTRS). It associated the standard variational auto encoder (VAE) architecture introduced by Kingma and Welling (2013) with deep latent variable models, handling both ordinal rating entries and comment texts made by users about products. The particularity of the neural network makes the inference method more flexible, and by assuming latent representations of both users and products, our model can better capture the relationships between them.

In this article, we first introduce the principles and developments of deepLTRS, and then demonstrate the results of numerical experiments by implementing our model on simulated data sets.

2 Deep lower rank matrix factorization

Consider an $M \times P$ data matrix Y of ordinal data involving a relation between the i-th row and the j-th column (e.g. a customer reviewing a product). One or more entries in this matrix would be missing (i.e. not observed) and we might be interested into inferring the missing value(s) from the observed ones. The following generative model is introduced

$$Y_{ij} = f_\gamma(< R_i, C_j >) + \epsilon_{ij}, \tag{1}$$

where $R_i \in \mathbb{R}^D$ is the latent representation of the i-th row of Y (denoted by Y_i) and C_j is the latent representation of the j-th column of Y (denoted by Y^j), $D << \min\{M, P\}$. The function f_γ is called *decoder*, which depends on a set of parameters γ. We can assume that it is a neural network as long as f_γ is continuous. The residuals ϵ_{ij} are supposed i.i.d. normally distributed random variables, with zero mean and unknown variance η^2.

In the following, R_i and C_j are seen as two independent random variables, such that

$$R_i \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, I_D),\ \forall i; \quad C_j \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, I_D),\ \forall j \tag{2}$$

$R \in \mathbb{R}^{M \times D}$ denotes the matrix whose i-th row is R_i and $C \in \mathbb{R}^{P \times D}$ the matrix whose j-th row is C_j. Since R and C are random matrices, we might be interested into solving the following maximization problem

$$\max_{\gamma, \eta^2} \log p(Y|\gamma, \eta^2), \tag{3}$$

where

$$p(Y|\gamma, \eta^2) = \int_{(R,C)} p(Y, R, C|\gamma, \eta^2) dR dC$$

is the observed data log-likelihood. Solving this problem should allow us to predict any entry Y_{ij} based on some values of R_i and C_j, since a maximum likelihood (ML) estimate of the decoder would be available. However, computing the most likely posterior value of R_i or C_j with the given data is not straightforward, and the maximization problem in Eq.(3) is difficult to derive directly. Therefore, we propose a variational strategy to replace the observed data log-likelihood by a lower bound

$$\log p(Y|\gamma, \eta^2) \geq \mathbb{E}_{q(R,C)} \left[\log \frac{p(Y, R, C|\gamma, \eta^2)}{q(R, C)} \right], \tag{4}$$

for all $q(\cdot, \cdot)$, joint distribution over the pair (R, C). Relying on a mean field approximation, it is also assumed that

$$q(R, C) = \prod_i q(R_i) \prod_j q(C_j), \tag{5}$$

$$q(R_i) = g(R_i; \mu_i^R := h_{1,\phi}(Y_i), S_i^R := h_{2,\phi}(Y_i)), \tag{6}$$

and

$$q(C_j) = g(\mu_j^C := l_{1,\iota}(Y^j), S_j^C := l_{2,\iota}(Y^j)), \tag{7}$$

where $g(\cdot, \mu, \Sigma)$ denotes the multivariate Gaussian distribution density function corresponding to the distribution $\mathcal{N}(\mu, \Sigma)$. The two functions $h_\phi : \mathbb{R}^P \to \mathbb{R}^{2 \times D}$ and $l_\iota : \mathbb{R}^M \to \mathbb{R}^{2 \times D}$ are the *encoders* parametrized by ϕ and ι, respectively.

The *deep latent variable model* (DLVM, Rezende et al. (2014)) described so far could be fitted to some data by using a Monte Carlo variational EM algorithm. Then the idea would be using the variational maximum at posteriori (VMAP) estimates $\hat{\mu}_i^R$ and $\hat{\mu}_j^C$ to replace R_i and C_j in Eq. 1. Thus, the ML estimate $\hat{\gamma}$ could be employed to predict Y_{ij}. This prediction would be particularly useful in case the true Y_{ij} is missing (not observed). Instead, we assume that more information is available in the form of *text*. For instance, when a customer rates a product, a *review* is usually available. In the next section, we will extend the DLVM detailed so far to account for this additional information.

3 A text based recommender system

By considering all the available reviews, it is possible to gather all the different vocables employed by the users into a *dictionary*, whose size is V. Thence, with a slight abuse of notation, we denote $W^{(i,j)} \in \mathbb{N}^V$ the review by the i-th user of the j-th product. The v-th entry of $W^{(i,j)}$, denoted by $W_v^{(i,j)}$, is the number of times that a word v appears into the corresponding review.

Following the Latent Dirichlet Allocation (LDA) by Blei et al. (2003), each document $W^{(i,j)}$ follows a mixture distribution over a set of K latent topics. The topic proportions in the document $W^{(i,j)}$ are denoted by θ_{ij}, a vector in lying in the $K-1$ simplex. Being β_{vk} the probability that vocable v is extracted from the k-th topic, with $\sum_{k=1}^{K} \beta_{kv} = 1$, LDA assumes that

$$p(W^{(i,j)}|\theta_{ij}) \sim \text{Multinomial}(L_{ij}, \beta\theta_{ij}),$$

where L_{ij} is a parameter accounting for the number of words in the review $W^{(i,j)}$ and supposed $\beta \in \mathbb{R}^{V\times K}$ is a matrix whose entry (v,k) is β_{vk}. Moreover, conditionally to θ_{ij}, all the reviews $\{W^{(i,j)}\}$ are independent random vectors.

In order to relate this generative model for text to the one outlined in Section 2, we further assume that the topic proportion as sampled as follows

$$\theta_{ij} = \sigma\left(\frac{R_i + C_j}{2}\right), \tag{8}$$

where $\sigma(\cdot)$ denotes the softmax function and R_i, C_j are the very same random vectors appearing in Eq. 1. This assumption replaces the one according to which the parameters θ_{ij} would follow independent Dirichlet distributions, as in LDA. The interpretation of the equation above is clear: the topic proportions used in the review $W^{(i,j)}$ are a function of the mid point between the embedding vectors of the user i and the product j respectively, in the latent space $\mathbb{R}^D$.

Finally, given a pair (R_i, C_j), it is assumed that Y_{ij} and $W^{(i,j)}$ are independent and we describe a framework in which the dependence between them is completely captured by the latent embedding vectors R_i and C_j.

A graphical representation of the generative model described so far can be seen in Figure 1. All the above assumptions allow us to define a lower bound of the data log-likelihood of the observed data (Y, W)

$$\begin{aligned}
\log p(Y, W|\eta^2, \gamma, \beta) &\geq \mathbb{E}_{q(R,C)}\left[\frac{\log p(Y, W, R, C|\eta^2, \gamma, \beta)}{q(R,C)}\right] \\
&= \mathbb{E}_{q(R,C)}\left[\frac{\log p(W, R, C|\beta)}{q(R,C)} + \frac{\log p(Y, R, C|\eta^2, \gamma)}{q(R,C)}\right] \\
&= \mathbb{E}_{q(R,C)}\left[\log p(W|R, C, \beta)\right] + \mathbb{E}_{q(R,C)}\left[\log p(Y|R, C, \eta^2, \gamma)\right] \\
&+ 2D_{KL}(q(R,C)||p(R,C))
\end{aligned} \tag{9}$$

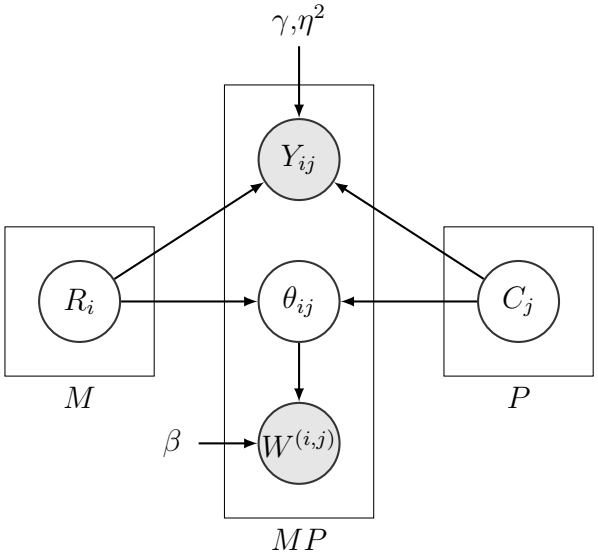


Figure 1: Graphical representation of the generative model.

where $D_{KL}(q(\cdot)||p(\cdot)))$ denotes the Kullback-Leibler divergence between the variational posterior distribution of the latent random matrices and their prior distribution. The above inequality applies to every joint distribution over the pair (R, C). In the following, the mean field approximation in Eq. 5 is maintained while, in order to account for the additional textual information, Eqs. 6-7 are modified as follows

$$q(R_i) = g(R_i; \mu_i^R := h_{1,\phi}(Y_i, W^{(i,\cdot)}), S_i^R := h_{2,\phi}(Y_i, W^{(i,\cdot)})) \tag{10}$$

and

$$q(C_j) = g(\mu_j^C := l_{1,\iota}(Y^j, W^{(\cdot,j)}), S_j^C := l_{2,\iota}(Y^j, W^{(\cdot,j)})), \tag{11}$$

where $W^{(i,\cdot)} := \sum_j W^{(i,j)}$ corresponds to a document concatenating all the reviews written by user i and, similarly $W^{(\cdot,j)} := \sum_i W^{(i,j)}$ corresponds to all the reviews about the j-th product. Notice also that the encoders h_ϕ, l_ι are now functions being defined on $\mathbb{R}^{P+V}$ and $\mathbb{R}^{M+V}$, respectively.

4 Experiments on simulated data

An ordinal data matrix Y with $M = 750$ rows and $P = 600$ columns is simulated in order to study the robustness of deepLTRS regarding the sparsity. Our model is also compared to several state-of-the-art competitors. Figure 2 presents the RMSE results on test data.

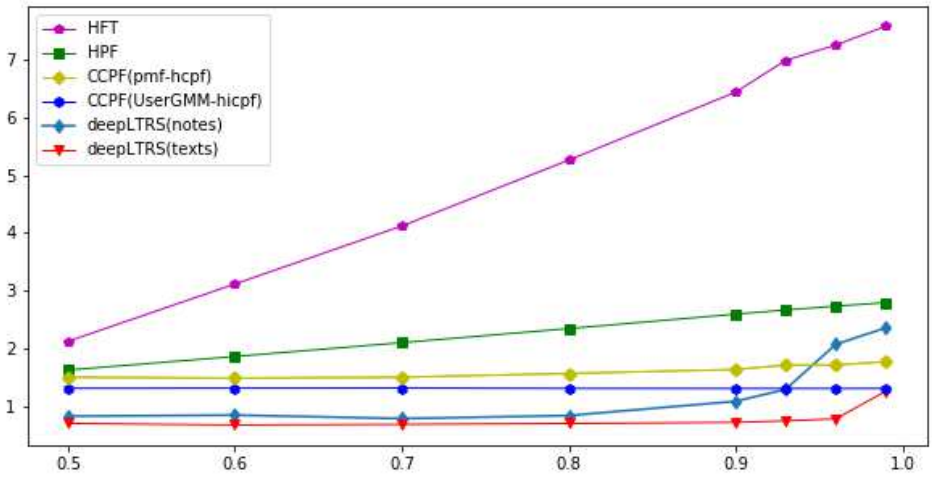


Figure 2: Test RMSE of models: HFT, HPF, CCPF and deepLTRS with different sparsity.

This work has been supported by the French government, through the 3IA Côte d'Azur Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002.

Bibliographie

Gopalan, P., Hofman, J.M. and Blei, D.M., (2013), *Scalable Recommendation with Poisson Factorization*, *arXiv preprint arXiv:1311.1704*.

Basbug, M.E, and Engelhardt, B.E., (2016), *Hierarchical Compound Poisson Factorization*, *Proceedings of the International Conference on international conference on Machine Learning*, pp.17951803.

McAuley, J.J, and Leskovect, J., (2013), *Hidden factors and hidden topics: understanding rating dimensions with review text*, *RecSys'13*, DOI:10.1145/2507157.2507163.

Kingma, D.P. and Welling, M., (2003). *Auto-encoding Variational Bayes*, *arXiv preprint arXiv:1312.6114*.

Rezende, D.J., Mohamed, S. and Wierstra, D., (2014). *Stochastic Backpropagation and Approximate Inference in Deep Generative Models*, *arXiv preprint arXiv:1401.4082*.

Blei, D.M., Ng, A.Y. and Jordan, M.I., (2003). *Latent Dirichlet Allocation*, *The Journal of Machine Learning Research*, 3, 993-1022.

Strivastava, A. and Sutton, C., (2017). *Autoencoding Variational Inference for Topic Models*, *arXiv preprint arXiv:1703.01488*.

A hidden semi-Markov model for segmenting environmental toroidal data

Francesco Lagona [1,3] & Antonello Maruotti [2,3]

[1] *University of Roma Tre, via Chiabrera 199 00145 Rome (Italy) - francesco.lagona@uniroma3.it*
[2] *LUMSA University, Via della Traspontina, 21 - 00193 Rome (Italy) - a.maruotti@lumsa.it*
[3] *Dept. of Mathematics, University of Bergen, Allégaten 41, 5007 Bergen (Norway)*

Abstract. Toroidal time series are temporal sequences of bivariate angular observations that often arise in environmental and ecological studies. A hidden semi-Markov model is proposed for segmenting these data according to a finite number of latent classes, associated toroidal densities. The model conveniently integrates circular correlation, multimodality and temporal auto-correlation. A computationally efficient EM algorithm is proposed for parameter estimation. The proposal is illustrated on a time series of wind and sea wave directions.

Keywords. hidden semi-Markov model, EM algorithm, model-based clustering, toroidal data

1 Introduction

Bivariate sequences of angles are often referred to as toroidal time series, because the pair of two angles can be represented as a point on a torus. These data often arise in environmental and ecological studies. Examples include time series of wind and wave directions [8], time series of wind mean directions and directions of the maximum gust observed each day [2] and time series of turning angles in studies of animal movement [10].

The analysis of toroidal time series is complicated by the difficulties in modeling the dependence between angular measurements over time [7]. An additional complication is given by the multimodality of the marginal distribution of the data, because environmental toroidal data are observed under time-varying heterogeneous conditions.

This paper introduces a toroidal hidden semi-Markov model (HSMM) that simultaneously accounts for dependence across circular measurements, temporal auto-correlation, multimodality and latent time-varying heterogeneity. Under this model, the distribution of toroidal data is approximated by a mixture of toroidal densities, whose parameters depend on the evolution of a latent semi-Markov process. While the toroidal density

accommodates dependence between two circular variables, a mixture of toroidal densities allows for multimodality and, finally, a latent semi-Markov process accounts for temporal correlation and, simultaneously, for time-varying heterogeneity.

Our proposal extends previous approaches that are based on toroidal hidden Markov models [9, 1]. Under a toroidal hidden Markov model, the data are approximated by a mixture of toroidal densities, whose parameters depend on the evolution of a latent, first-order Markov chain with a finite number of states. The sojourn times of each state of a Markov chain are distributed according a geometric distribution. Hence the most likely dwell time for every state of a hidden Markov model with underlying first-order Markov chain is 1. Our proposal relaxes this restrictive assumption by replacing the latent Markov chain with a latent semi-Markov model, allowing for sojourn times that are not necessarily geometrically distributed.

2 A hidden semi-Markov model for toroidal data

Let $\mathbf{z} = (x, y)$ be a pair of angles, $x, y \in [0, 2\pi)$. Moreover, let $f(x; \alpha)$ and $f(y; \beta)$ be two circular densities, respectively known up to the parameters α and β. Further, let $F(x; \alpha)$ and $F(y; \beta)$ be the two cumulative distribution functions of x and y, defined with respect to a fixed, although arbitrary, origin. Finally, let $g(u; \gamma), u \in [0, 2\pi)$ be a parametric circular density, known up to a parameter γ. Then,

$$f_q(\mathbf{z}; \theta) = 2\pi g\left(2\pi\left(F(x; \alpha) - qF(y; \beta)\right)\right) f(x; \alpha) f(y; \beta)) \quad q = \pm 1 \tag{1}$$

is a parametric toroidal density with support $[0, 2\pi)^2$, known up to the parameter vector $\theta = (\alpha, \beta, \gamma)$, having the marginal densities $f(x; \alpha)$ and $f(y; \beta)$ [3]. Equation (1) is a typical example of a copula-based construction of a bivariate density, obtained by decoupling the margins from the joint distribution. When the binding density g is the uniform circular distribution, say $g(x) = (2\pi)^{-1}$, then equation (1) reduces to the product of the marginal densities. Otherwise, the dependence between x and y is captured by the concentration of g: when g is highly concentrated, the dependence is high; when g is more diffuse, dependence is low. Finally, the constant $q = \pm 1$ determines whether the dependence between x and y is positive ($q = 1$) or negative ($q = -1$).

The proposed hidden semi-Markov model can be described as a dynamic mixture of copula-based toroidal densities. To illustrate, let $\mathbf{z} = (\mathbf{z}_t, t = 1, \ldots, T)$, $\mathbf{z}_t = (x_t, y_t)$, $x_t, y_t \in [0, 2\pi)$, be a toroidal time series. We assume that the distribution of the data is driven by the evolution of an unobserved semi-Markov process with K states, which represents (time-varying) latent classes and can be specified as a sequence $\mathbf{u} = (\mathbf{u}_t, t = 1, \ldots, T)$ of multinomial variables $\mathbf{u}_t = (u_{t1} \ldots u_{tK})$ with one trial and K classes, whose binary components represent class membership at time t. The joint distribution $p(\mathbf{u}; \pi)$ of the chain is fully known up to a parameter π that includes K initial probabilities $\pi_k = P(u_{1k} = 1), k = 1, \ldots, K, \sum_k \pi_k = 1$, $K^2 - K$ transition probabilities $\pi_{hk} = P(u_{tk} =$

$1|u_{t-1,h} = 1), h, k = 1, \ldots, K, \sum_k \pi_{hk} = 1,\ h \neq k$ (whereas $\pi_{kk} = 0,\ k = 1 \ldots K$), and, finally, p parameters of the dwell time distributions of each state.

The specification of the HSMM is completed by assuming that the observations are conditionally independent, given a realization of the semi-Markov process. As a result, the conditional distribution of the observed process, given the latent process, takes the form of a product density, say

$$f(\mathbf{z}|\mathbf{u}; \theta_1, \ldots, \theta_K) = \prod_{t=1}^{T} \prod_{k=1}^{K} f(\mathbf{z}_t; \theta_k)^{u_{tk}}, \tag{2}$$

where $f(\mathbf{z}; \theta_k)$, $k = 1, \ldots, K$ are the K cylindrical densities defined by (1) and known up to a vector of parameters θ_k.

The likelihood function of the model is therefore obtained by integrating the joint density of the observed data and the unobserved class memberships with respect to the segmentation $\mathbf{u}$, namely

$$L(\pi, \theta; \mathbf{z}) = \sum_{\mathbf{u}} p(\mathbf{u}; \pi) f(\mathbf{z}|\mathbf{u}; \theta_1, \ldots, \theta_K). \tag{3}$$

By computing the maximum likelihood estimate $\hat{\theta}$ [11, chapter 12], the toroidal time series can be segmented according to the posterior probabilities of class membership

$$\hat{\pi}_{tk} = P(u_{tk} = 1 \mid \mathbf{z}; \hat{\theta}), \tag{4}$$

based on $\hat{\theta}$. More precisely, the observation at time t can be allocated to class $k^\star$ if $\hat{\pi}_{tk^\star} \geq \hat{\pi}_{th}$, for each $h = 1 \ldots K$ (maximum a posteriori, MAP, allocation).

When the dwell distribution of each latent state is geometric, the model reduces to a hidden Markov model that ignores alternative dwell time distribution. If, additionally, the transition probability matrix of the model has equal rows, the model reduces to a mixture model where observations are clustered by ignoring the information redundancy that is due to temporal correlation.

3 An application to marine data

The proposed methods have been implemented to segment a time series of $T = 1326$ semi-hourly wind and wave directions, taken in wintertime by the buoy of Ancona, which is located in the Adriatic Sea at about 30 km from the coast. Figure 1 displays the scatter plot of the data. Point coordinates indicate the direction (in radians) from which winds blow and waves travel. For simplicity, these bivariate observations are plotted on the plane, although data points are actually on a torus. The interpretation of these data is not easy. While in the ocean wind and wave directions are strongly correlated, this is not necessarily true in the Adriatic Sea, due to the orography of the basin and the

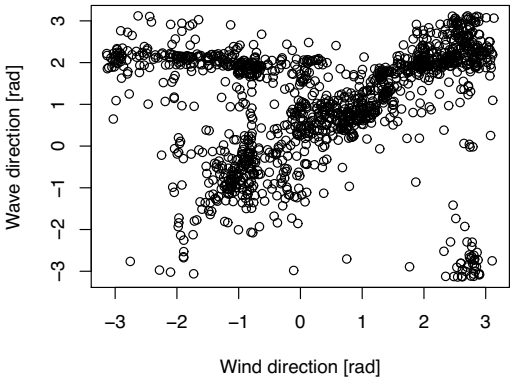


Figure 1: Wave directions and heights, as observed by the buoy of Ancona in wintertime ($-\pi$, $-pi/2$, 0, $\pi/2$ respectively indicate South, West, North, East). For simplicity, the data are plotted on the plane, although they are points on the torus $[-pi/2, \pi/2)^2$.

location of the buoy. Coastal winds generate synchronized waves only when the waves travel unobstructed, that is, either northwesterly or southeasterly, along the major axis of the basin. When western and south-western winds blow from the coast, waves are not synchronized with wind and travel along the major axis of the Adriatic basin from SE to NW. This explains the clusters shown in Figure 1 and suggests the occurrence of two latent wind–wave regimes. Accordingly, a HSMM with two states have been estimated from these data.

The proposed HMM requires a parametric specification of the toroidal density (1), which reduces to the choice of the binding density g and the choice of the marginal densities $f(x;\alpha)$ and $f(y;\beta)$ that respectively model the marginal distribution of the wind and wave direction. However, depending on the choice of the binding density, the density (1) can be multimodal [4]. Using multimodal densities in segmentation and classification problems, such as the one motivating this paper, may unnecessarily complicate the interpretation of the results. Unimodal densities can however be obtained by using the wrapped Cauchy as a binding density g [4].

Accordingly, for this study, the binding density has been specified as a centered wrapped Cauchy

$$g(u;\gamma) = \frac{1}{2\pi}\frac{1-\gamma^2}{1+\gamma^2-2\gamma\cos(u)} \quad u \in [0, 2\pi).$$

This circular density depends on a single concentration parameter $\gamma \in [0, 1)$ and reduces to the uniform circular density when $\gamma = 0$.

Wrapped Cauchy densities that include additional location parameters α_1 and β_1 have

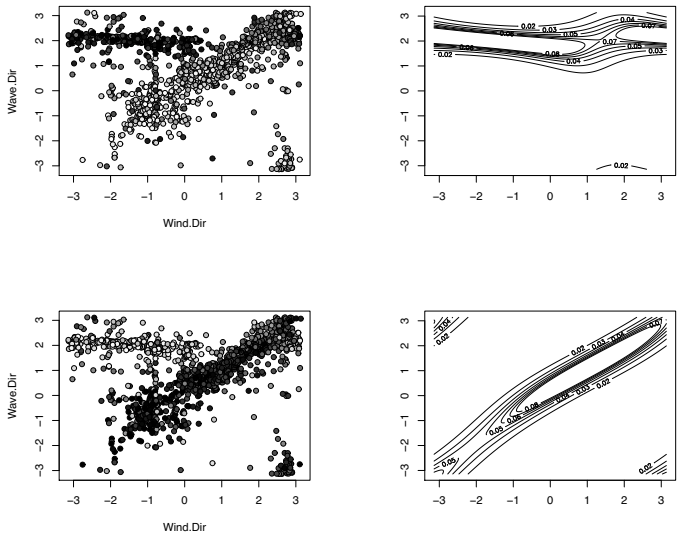


Figure 2: Segmentation of a time series of wind and wave directions. Left: observations colored with grey levels according to the estimated membership probabilities of each class (black indicates a probability equal to 1). Right: contour plot of state-specific toroidal densities.

been instead exploited to model the marginal distributions of wind and wave direction, say

$$f(x;\alpha) = \frac{1}{2\pi}\frac{1-\alpha_2^2}{1+\alpha_2^2-2\alpha_2\cos(y-\alpha_1)} \quad x \in [0,2\pi) \tag{5}$$

$$f(y;\beta) = \frac{1}{2\pi}\frac{1-\beta_2^2}{1+\beta_2^2-2\beta_2\cos(y-\beta_1)} \quad y \in [0,2\pi) \tag{6}$$

The proposed toroidal density is therefore obtained by taking a wrapped Cauchy density that binds wrapped Cauchy marginals, a model known as the bivariate wrapped Cauchy model [5].

Figure 2 shows the shapes of the two state-specific toroidal distributions and the segmented observations. The model successfully segment the observations according to to clusters, and offers a clear-cut indication of the distribution of the data under each regime. Under state 1, wind and wave directions are essentially independent, because coastal winds do not generate waves. Under state 2, winds blows along the major axis of the Adriatic basin and their directions are highly correlated with the directions of the wave that they generate.

Overall, the model describes the plasticity of the wind–wave interaction in the Adriatic Sea, indicating that the joint distribution of wind and wave data changes under different environmental regimes. Regime switching changes not only the modal directions and con-

centrations around these modes but also, and more interestingly, the correlation structure of the data. As a result, on the one side, the (marginal) weak correlation between wind and wave directions is explained by the presence of coastal winds (component 1). On the other side, the model indicates that the wind direction is an accurate predictor of the wave direction only under a specific regime (state 2).

References

[1] Bulla J, Lagona F, Maruotti A, Picone M (2012) A Multivariate Hidden Markov Model for the Identification of Sea Regimes from Incomplete Skewed and Circular Time Series, Journal of Agricultural, Biological, and Environmental Statistics, 17: 544-567

[2] Coles S (1998) Inference for circular distributions and processes, Statistics and Computing, 8: 105-113.

[3] Johnson RA, Wehrly TE (1978) Some angular-linear distributions and related regression models. Journal of the American Statistical Association 73: 602-606.

[4] Jones MC, Pewsey A, Kato S (2015). On a class of circulas: copulas for circular distributions. Annals of the Institute of Statistical Mathematics 67: 843-862.

[5] Kato S, Pewsey A (2015) A Möbius transformation-induced distribution on the torus, Biometrika, 102: 359-370

[6] Lagona F (2019) Copula-based segmentation of cylindrical time series, Statistical and Probability Letters, 144: 16-22.

[7] Lagona F (2018) Correlated cylindrical data. In: C. Ley and T. Verdebout (Eds) Applied Directional Statistics: Modern Methods and Case Studies, Chapman & Hall/CRC: New York, 45-59.

[8] Lagona F, Picone M, Maruotti A, Cosoli S (2014) A hidden Markov approach to the analysis of space-time environmental data with linear and circular components, Stochastic Environmental Research and Risk Assessment 29: 397-409.

[9] Lagona F, Picone M (2013) Maximum likelihood estimation of bivariate circular hidden Markov models from incomplete data, Journal of Statistical Computation and Simulation, 83: 1223-1237

[10] Mastrantonio G (2018) The joint projected normal and skew-normal: A distribution for poly-cylindrical data, Journal of Multivariate Analysis, 165: 14-26.

[11] Zucchini W, Macdonald IL and Langrock R (2016) Hidden Markov models for time series, Chapman and Hall, Boca Raton FL (US)

Convergence d'un score d'ensemble en ligne : étude empirique

Benoît Lalloué [1,3,*], Jean-Marie Monnez [1,3,†], Éliane Albuisson [2,4,5,‡]

[1] *Université de Lorraine, CNRS, Inria*, IECL**, F-54000 Nancy, France*
**Inria, Project-Team BIGS*
***Institut Elie Cartan de Lorraine, Vandoeuvre-lès-Nancy, France*
[2] *Université de Lorraine, CNRS, IECL**, F-54000 Nancy, France*
[3] *Inserm U1116, Centre d'Investigation Clinique Plurithématique 1433, Université de Lorraine, Nancy, France*
[4] *BIOBASE, Pôle S2R, CHRU de Nancy, Vandoeuvre-lès-Nancy, France*
[5] *Faculté de Médecine, InSciDenS, Vandoeuvre-lès-Nancy, France*
∗ *benoit.lalloue@univ-lorraine.fr;* † *jean-marie.monnez@univ-lorraine.fr;*
‡ *eliane.albuisson@univ-lorraine.fr*
Financement : Programme Investissement d'Avenir ANR-15-RHU-0004

Résumé. Dans un contexte en ligne où des données arrivent de façon continue, on souhaite actualiser les paramètres d'un score "batch" construit à l'aide d'une méthode d'ensemble. On utilise pour cela des processus d'approximation stochastique, dont la convergence a été établie théoriquement par les auteurs, permettant d'actualiser les estimations des paramètres lors de la prise en compte de nouvelles observations sans avoir à conserver toutes les données obtenues précédemment. Nous étudions ici empiriquement la convergence du score en ligne vers le score "batch", en utilisant différents jeux de données à partir desquels on simule des flux de données et différents types de processus.

Mots-clés. Apprentissage pour les données massives, approximation stochastique, médecine, méthode d'ensemble, score en ligne.

Abstract. In an online setting, where data arrives continuously, we want to update the parameters of a "batch" score constructed with an ensemble method. To do so, we use stochastic approximation processes, the convergence of which has been theoretically established by the authors, so that parameter estimates can be updated when new observations are taken into account without the need to store all the data obtained previously. Here we study empirically the convergence of the online score to the "batch" score, using different datasets from which data streams are simulated and using different types of processes.

Keywords. Learning for big data, stochastic approximation, medicine, ensemble method, online score.

1 Introduction

Afin d'établir un score d'événement en ligne actualisé lors de l'arrivée de nouvelles données d'apprentissage sans avoir à stocker l'ensemble des données obtenues, nous avons entrepris une étude en plusieurs étapes.

Pour cela, deux classifieurs ont été utilisés, l'analyse discriminante linéaire et la régression logistique, ainsi qu'une méthode de construction d'un score d'ensemble qui a été définie dans Duarte et al (2018a). Les méthodes d'ensemble, en construisant une collection de prédicteurs "de base" (en faisant varier l'échantillon initial, les variables sélectionnées, la méthode de régression ou de classification utilisée...) puis en agrégeant leurs prédictions, permettent souvent d'obtenir de meilleurs résultats que les prédicteurs individuels (si ceux-ci sont relativement bons et suffisamment différents, Genuer et Poggi 2017).

Nous avons tout d'abord défini et démontré la convergence de plusieurs types de processus d'approximation stochastique pour actualiser en ligne les paramètres d'une fonction de régression linéaire (Duarte *et al.* 2018b) ou logistique (Lalloué *et al.* 2019b) et montré l'intérêt d'utiliser des données standardisées en ligne plutôt que des données brutes, en particulier pour éviter une explosion numérique.

Le principe général de construction d'un score en ligne a ensuite été présenté dans Lalloué *et al.* (2019a).

Pour terminer cette étude, nous avons implémenté en R la construction de ce score en ligne et en avons testé empiriquement la convergence sur plusieurs jeux de données, en utilisant pour chaque classifieur plusieurs processus d'approximation stochastique et en comparant la précision des estimations obtenues. Nous présentons ici certains de ces résultats empiriques.

2 Expérimentations

2.1 Données et score "batch" de référence

Cinq jeux de données ont été utilisés : quatre disponibles sur Internet et un dérivé de l'étude EPHESUS (Pitt 2003), tous déjà utilisés pour tester les performances de processus de gradient stochastique (Duarte *et al.* 2018b, Lalloué *et al.* 2019b). `Twonorm` et `Ringnorm` sont des jeux de données simulées avec des variables homogènes (Breiman 1996). `Quantum` contient des données observées réelles, sans valeurs aberrantes et dont la plupart des variables sont sur la même échelle. `Adult2` et `HOSPHF30D` contiennent des données observées réelles, avec des valeurs aberrantes et des variables de différents types et sur différentes échelles. La table 1 résume ces données. Les détails des pré-traitements sont donnés dans Lalloué *et al.* (2019b)

Pour chaque jeu de données, on construit un score d'ensemble de référence en utilisant la méthode définie dans Duarte *et al.* (2018a) avec les paramètres suivants :

Table 1: Description of the datasets.

Dataset name	N_a	N	p_a	p	**Source**
Twonorm	7400	7400	20	20	www.cs.toronto.edu/ delve/data/datasets.html
Ringnorm	7400	7400	20	20	www.cs.toronto.edu/ delve/data/datasets.html
Quantum	50000	15798	78	12	dérivé de www.osmot.cs.cornell.edu/kddcup
Adult2	45222	45222	14	38	dérivé de www.cs.toronto.edu/ delve/data/datasets.html
HOSPHF30D	21382	21382	29	13	dérivé de EPHESUS study

N_a: nombre d'observations disponibles; N: nombre d'observations sélectionnées; p_a: nombre de paramètres disponibles; p: nombre de paramètres sélectionnés.

- Deux règles de classification sont utilisées : analyse discriminante linéaire (LDA) et régression logistique (LR).
- 100 échantillons bootstrap sont générés pour les deux règles.
- Toutes les variables disponibles sont incluses (des expérimentations avec des modalités de sélection et des nombres différents de variables tirées au sort sont en cours).
- Pour chaque règle de classification, les 100 prédicteurs associés sont agrégés par moyenne arithmétique puis les coefficients sont normalisés de manière à ce que le score varie entre 0 et 100.
- L'agrégation entre les deux scores synthétiques S_1 (LDA) et S_2 (LR) est faite par combinaison convexe : $S = \lambda S_1 + (1-\lambda)S_2$ avec ici $\lambda = 0.5$ (ultérieurement, une valeur optimale de λ pourra être déterminée).

Le score ainsi obtenu pour chaque jeu de données est utilisé comme "gold standard" pour évaluer la convergence des processus en ligne testés.

A partir de chaque jeu de données, un flux de données est simulé en effectuant à chaque étape un tirage au hasard avec remise de 100 nouvelles observations. Les scores en ligne sont alors construits et mis à jour à partir de ces flux.

2.2 Processus

Types de processus

Les processus stochastiques (X_n) utilisés sont de trois types différents :

- gradient stochastique "classique" (notation C$___$). À l'étape n, card $I_n = m_n$ observations (Z_j, S_j) sont prises en compte et on calcule récursivement :

$$X_{n+1} = X_n - a_n \frac{1}{m_n} \sum_{j \in I_n} Z_j \left(h(Z_j' X_n) - S_j \right);$$

avec Z_j vecteur des variables explicatives ; $S_j \in \{0,1\}$; $h(u) = u$ pour la LDA, $h(u) = \frac{e^u}{1+e^u}$ pour la LR.

- gradient stochastique "moyennisé" (notation A␣␣␣) : $\overline{X}_{n+1} = \frac{1}{n+1}\sum_{i=1}^{n+1} X_i$.
- uniquement dans le cas de la LDA : processus prenant en compte à chaque étape toutes les observations (Z_j, S_j) jusqu'à cette étape, $j \in I_1 \cup ... \cup I_n$ (mention finale "all")(Duarte *et al.* 2018b).

Dans tous les cas, les variables explicatives sont standardisées en ligne (notation ␣S␣␣) : le principe et l'intérêt pratique de cette méthode pour éviter des explosions numériques ont été montrés dans Duarte *et al.* (2018b) et dans Lalloué *et al.* (2019b). En effet, pour certains jeux de données (`Adult2`, `HOSPHF30D`) les processus avec données brutes conduisent à une explosion numérique, contrairement à ceux avec données standardisées en ligne.

Choix du pas

Le pas a_n peut être :

- continûment décroissant : $a_n = \frac{c}{(b+n^\alpha)}$ (notation ␣␣␣V).
- constant : $a_n = 1/p$ (avec p le nombre de variables explicatives) (notation ␣␣␣C).
- constant par paliers (Bach 2014) : $a_n = \frac{c}{(b+\lfloor \frac{n}{\tau} \rfloor)^\alpha}$ ($\lfloor . \rfloor$ est la partie entière, τ la taille des paliers) (notation ␣␣␣P).

On prend dans tous les cas $\alpha = 2/3$, $b = 1$ et $c = 1$.

Notation des processus

Dans un couple de processus, le premier est celui utilisé pour la LDA, le second celui pour la LR.

Par exemple, AS100Call_AS100P200 désigne le couple formé pour la LDA d'un processus moyennisé (A) avec données standardisées en ligne (S), 100 nouvelles observations par étape (100), à pas constant(C) et prenant en compte toutes les observations jusqu'à l'étape en cours (all) ; et pour la LR d'un processus moyennisé (A) avec données standardisées en ligne (S), 100 nouvelles observations par étape (100) et à pas constant par paliers de taille 200 (P200).

Processus testés

Les six couples de processus testés font partie de ceux ayant eu les meilleures performances lors des études dédiées à la LDA en ligne (Duarte *et al.* 2018b) et à la régression logistique en ligne (Lalloué *et al.* 2019b), ou sont des processus "classiques" habituellement utilisés (mis à part la standardisation en ligne des données).

On utilise 100 nouvelles observations par étape. Chaque processus a été appliqué sur chacun des flux issus des jeux de données, avec un nombre d'observations utilisées de $100N$, correspondant à N itérations.

Critère de convergence

On a utilisé comme critère de convergence la différence relative des normes $\frac{\|\theta^b - \hat{\theta}_N\|}{\|\theta^b\|}$ entre le vecteur θ^b des coefficients obtenus pour le score "batch" et le vecteur $\hat{\theta}_N$ des coefficients estimés par un processus après N itérations. On considère qu'il y a eu convergence lorsque la valeur de ce critère est inférieure au seuil arbitraire de 0.05.

2.3 Résultats

Trois résultats sont comparés pour chaque couple de processus : la valeur du critère sur les coefficients normalisés (modifiés pour que le score varie entre 0 et 100) et standardisés (divisés par l'écart-type de la variable associée) pour le score synthétique S_1 obtenu par l'agrégation des LDA, celle pour le score synthétique S_2 obtenu par agrégation des régressions logistiques, et celle pour le score final S (table 2).

Table 2: Différences relatives des normes après $100N$ observations utilisées

Processus		**Twonorm**	**Ringnorm**	**Quantum**	**Adult2**	**HOSPHF30D**
CS100V _CS100V	S_1	0.0010*	0.0020*	0.0073*	0.0076*	0.0165*
	S_2	0.0033*	0.0009*	0.0168*	0.1002	0.0566
	S	**0.0015***	**0.0014***	**0.0083***	**0.0414***	**0.0289***
AS100P50 _AS100P50	S_1	0.0006*	0.0007*	0.0027*	2.7560	0.0176*
	S_2	0.0006*	0.0007*	0.0032*	0.0346*	0.0203*
	S	**0.0005***	**0.0007***	**0.0029***	**1.6968**	**0.0192***
AS100C _AS100P200	S_1	0.0006*	0.0007*	0.0028*	0.0066*	0.0165*
	S_2	0.0007*	0.0007*	0.0033*	0.0069*	0.0206*
	S	**0.0006***	**0.0007***	**0.0030***	**0.0067***	**0.0190***
CS100Vall _CS100V	S_1	0.0005*	0.0006*	0.0033*	0.0287*	0.0153*
	S_2	0.0033*	0.0009*	0.0168*	0.1002	0.0566
	S	**0.0017***	**0.0007***	**0.0090***	**0.0281***	**0.0290***
AS100P50all _AS100P50	S_1	0.0006*	0.0007*	0.0046*	0.0100*	0.0060*
	S_2	0.0006*	0.0007*	0.0032*	0.0346*	0.0203*
	S	**0.0005***	**0.0007***	**0.0039***	**0.0193***	**0.0147***
AS100Call _AS100P200	S_1	0.0006*	0.0007*	0.0046*	0.0153*	0.0060*
	S_2	0.0007*	0.0007*	0.0033*	0.0069*	0.0206*
	S	**0.0005***	**0.0007***	**0.0039***	**0.0120***	**0.0149***

* marque les valeurs du critère < 0.05.

Première abréviation : processus pour la LDA; Deuxième : processus pour la régression logistique.

Type de processus : C pour SGD classique, A pour ASGD.

Données : R pour brutes, S pour standardisées en ligne (*1er nombre* : nombre de nouvelles observations par étape).

Pas : V pour variable, C pour constant, P pour constant par palier (*2e nombre* : taille des paliers).

On constate que, pour tous les couples de processus testés, le score en ligne final S est très proche du score de référence "batch" sur quatre des cinq jeux de données testés. Seul le couple de processus AS100P50_AS100P50 appliqué au jeu de données `Adult` ne converge pas en N itérations. Dans la plupart des cas, la valeur du critère pour le score final S est comprise entre celles des deux scores intermédiaires. Ceci conduit, pour deux couples de processus (CS100V_CS100V et CS100Vall_CS100V) appliqués à `Adult2` et `HOSPHF30D`, à

une convergence du score final alors que le score intermédiaire S_2 n'a pas encore convergé. Si l'on range les couples de processus du plus performant au moins performant pour chaque jeu de données puis qu'on calcule le rang moyen sur l'ensemble des jeux de données, le meilleur couple de processus est AS100P50all_AS100P50.

3 Conclusion

Nous avons mis au point un programme de construction d'un score en ligne en associant des processus avec données standardisées en ligne dont la convergence a déjà été établie théoriquement. On observe empiriquement la convergence de ce score d'ensemble en ligne vers le score "batch" avec plusieurs processus, ainsi que la supériorité de certains choix : processus moyennisés et pas constant par paliers notamment. On a inclus ici toutes les variables disponibles, le score est ainsi obtenu par bagging et agrégation de deux méthodes. D'autres expérimentations ont été effectuées avec des modalités de sélection aléatoire des variables.

Bibliographie

Bach, F. (2014). Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *Journal of Machine Learning Research* 15, pp. 595-627.

Breiman, L. (1996). Bias, variance, and arcing classifiers. *Technical Report 460*, Departement of Statistics, University of California, Berkeley.

Duarte, K., Monnez J.-M. and Albuisson E. (2018a). Methodology for Constructing a Short-Term Event Risk Score in Heart Failure Patients. *Appl Math*, 09(08):954-74.

Duarte, K., Monnez, J.-M. and Albuisson, E. (2018b). Sequential linear regression with online standardized data, *PloS One*, 13 (1) e0191186.

Genuer, R. and Poggi, J.-M. (2017). Arbres CART et Forêts aléatoires, Importance et sélection de variables, *hal-01387654*.

Lalloué, B., Monnez, J.-M and Albuisson, E. (2019a). Actualisation en ligne d'un score d'ensemble. *51e Journées de Statistique. hal-02152352.*

Lalloué, B., Monnez, J.-M and Albuisson, E. (2019b). Streaming constrained binary logistic regression with online standardized data, *hal-02156324*.

Oza, N.C. and Russell, S.J. (2001). Online Bagging and Boosting. In: *Proceedings of the Eighth International Workshop on Artificial Intelligence and Statistics, AISTATS 2001.*

Pitt, B., Remme, W., Zannad, F., Neaton, J., Martinez, F., Roniker, B., Bittman, R., Hurley, S., Kleiman, S., and Gatlin, M. (2003). Eplerenone, a selective aldosterone blocker, in patients with left ventricular dysfunction after myocardial infarction. *New England Journal of Medicine*, 348, pp. 1309-1321.

Construction of a copula estimator through recursive partitioning of the unit hypercube

Oskar Laverny [1] & Véronique Maume-Deschamps [2] & Esterina Masiello [3] & Didier Rullière [4]

[1] *Université Claude Bernard Lyon 1, Insitut Camille Jordan UMR 5208, France, et SCOR SE, France, oskar.laverny@math.univ-lyon1.fr*
[2] *Université Claude Bernard Lyon 1, Insitut Camille Jordan UMR 5208, France, veronique.maume@univ-lyon1.fr*
[3] *Université Claude Bernard Lyon 1, Insitut Camille Jordan UMR 5208, France, esterina.masiello@univ-lyon1.fr*
[4] *Université Claude Bernard Lyon 1, Laboratoire SAF EA 2429, France, didier.rulliere@univ-lyon1.fr*

Résumé. Nous construisons un estimateur de copule linéaire par morceau, flexible et consistant. Pour cela, nous nous appuyons sur la formalisation des copules patchwork ainsi que sur les divers estimateurs constants par morceaux de densités multivariés existants dans la littérature. Si les copules checkerboards imposent une partition, notre estimateur la construit à partir de l'échantillon disponible, en minimisant une distance choisie sur l'espace des copules. De plus, si l'ajout des contraintes de marges rend parfois certains estimateurs non-paramétriques inutilisables, notre estimateur n'estime que la structure de dépendance et garantit l'uniformité des marges. Des raffinements, comme la réduction de dimension localisée ou le bagging, sont développés, analysés et testés sur des données simulées.

Mots-clés. Estimation de copule, statistiques non-paramétriques, erreur quadratique intégrée, réduction de dimension localisée, dépendance, arbres d'estimation de densité, agrégation.

Abstract. We construct a flexible, consistent, piecewise linear estimator for a copula, leveraging the patchwork copula formalization and various piecewise constant density estimators. While the patchwork structure imposes the grid, our estimator is data-driven and constructs the grid recursively from the data, minimizing a chosen distance on the copula space. Furthermore, while the addition of the copula constraints makes the available solutions for density estimation unusable, our estimator is only concerned with dependence and guarantees the uniformity of margins. Refinements such as localised dimension reduction and bagging are developed, analyzed, and tested through applications on simulated data.

Keywords. Copula estimation, nonparametric statistics, integrated square error, localised dimension reduction, dependence, density estimation tree, bagging.

1 Introduction

Although the estimation of copulas is a wide-treated subject, most performant estimators avaliable in the litterature are based on restricted, parametric estimation: vine copulas (see Nagler and Czado (2016)) and graphical models (see Li et al. (2018)) for exemple are potential solutions but under restrictive assumptions. Classical density estimators such as kernels or wavelets do not satisfy marginal copula constraints. There also exists several tree-structured piecewise constant density estimators, but they do not always lead to proper copulas when applied on pseudo-observations or true copula samples.

We propose a non-parametric, piecewise constant, copula density estimator inspired by the patchwork framework of Durante et al. (2015) and the density estimation trees from Ram and Gray (2011). We build a recursive estimation procedure for this class, and assess its consistency. Our estimator is, under mild conditions, a proper copula. It provides a fast evaluation on new data points, can be rapidly simulated, and could be used for compression purposes.

This overview is organised as follows: Section 2 describes the piecewise linear copula class, Section 3 introduces the estimation procedure and the consistency result, and Section 4 finaly deals with numerical aspects.

2 Piecewise linear copulas

Let $X = (X_1, ..., X_d)$ be a d-dimensional random vector. We are here interested in the dependence structure between components of X. The copula, whose formalisation is due to Sklar (1959), is a key concept to the study of dependence between random variables. Suppose now that X has marginal distribution functions (d.f.) $F_1, ..., F_d$ and copula C with density c.

Let $\mathbb{I} = [0,1]^d$ be the unit hypercube, $\mathcal{L}$ be a finite partition of $\mathbb{I}$ consisting of (non-degenerated) hyperrectangles called leaves, and p be a vector of non-negative weights summing to one, indexed by these leaves. The piecewise linear copula with parameters $(p, \mathcal{L})$ is defined by its d.f. $C_{p,\mathcal{L}}$ and density $c_{p,\mathcal{L}}$ as:

$$\forall u \in \mathbb{I},\ C_{p,\mathcal{L}}(u) = \sum_{\ell \in \mathcal{L}} p_\ell \lambda_\ell(u) \text{ and } c_{p,\mathcal{L}}(u) = \sum_{\ell \in \mathcal{L}} \frac{p_\ell}{\lambda(\ell)} \mathbf{1}_{u \in \ell} \tag{1}$$

where λ denotes the Lebesgue measure of a set and $\lambda_\ell(u) = \lambda(\ell)^{-1}\lambda([0,u] \cap \ell)$. Under some conditions on $(p, \mathcal{L})$ that we will analyze, $C_{p,\mathcal{L}}$ will be a copula. Note that if for all $\ell \in \mathcal{L}$, $p_\ell = \lambda(\ell)$, then $C_{p,\mathcal{L}}$ is the independence copula, since leaves are hyperrectangular.

The above formulation of the piecewise linear copula allows to obtain closed-form expressions for classical quantities of interest in copula modeling, e.g. Kendall's tau and Spearman's rho, which can be useful for validation purposes.

3 The CORT estimator

Suppose that we have a dataset $(u_i)_{i\in 1:n}$ of observations in $\mathbb{I}$, from a copula C with density c. We seek parameters $(p, \mathcal{L})$ of $c^{(n)}_{p,\mathcal{L}}$, an approximation of c in the piecewise linear copula class. We adopt a two stage estimation procedure, considering first the partition $\mathcal{L}$ to be known. Following Ram and Gray (2011), we use an Integrated Square Error (ISE) loss to build the weights p knowing $\mathcal{L}$. The quantity to be minimised writes:

$$\|c_{p,\mathcal{L}} - c\|_2^2 = \int_{\mathbb{I}} (c_{p,\mathcal{L}}(u) - c(u))^2 du = \|c_{p,\mathcal{L}}\|_2^2 - 2\langle c_{p,\mathcal{L}}, c\rangle + \|c\|_2^2 \tag{2}$$

By a simple Monte-Carlo plug-in of the empirical d.f. in the cross-product, we obtain that weights p are the unique solution of the quadratic program with objective $p'Ap - 2p'A\mathrm{f}$, where A and f are given by, (denoting $|\mathcal{L}|$ the cardinal of $\mathcal{L}$,)

$$A = \left(\lambda(\ell)^{-1}\mathbf{1}_{\ell=k}\right)_{\ell\in\mathcal{L},k\in\mathcal{L}} \qquad \text{(size } |\mathcal{L}| \times |\mathcal{L}|\text{)}$$

$$\mathrm{f} = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{u_i\in\ell}\right)_{\ell\in\mathcal{L}} \qquad \text{(size } |\mathcal{L}|\text{)}$$

We denote $\langle .,.\rangle_{\mathcal{L}}$ the scalar product on $\mathbb{R}^{|\mathcal{L}|}$ given by $\langle x, y\rangle_{\mathcal{L}} = x'Ay$. The associated norm and distance are denoted $\|.\|^2_{\mathcal{L}}$ and $d_{\mathcal{L}}$. In particular, we have $\|c_{p,\mathcal{L}}\|_2^2 = \|p\|^2_{\mathcal{L}}$ and $\frac{1}{n}\sum_{i=1}^{n} c_{p,\mathcal{L}}(u_i) = \langle p, \mathrm{f}\rangle_{\mathcal{L}}$.

One can then show that the global minimiser of the Monte-Carlo plug-in in the ISE, without more constraints, is f. In order to obtain a proper copula, however, there are marginal uniformity constraints, which write:

$$\forall u \in \mathbb{I}, \forall j \in \{1, ..., d\}, \sum_{\ell\in\mathcal{L}} p_\ell \lambda_{\ell_j}(u_j) = u_j. \tag{3}$$

Denoting $\mathcal{C}$ the space defined by these constraints, we show that $\mathcal{C}$ is an affine half-space of $\mathbb{R}^{|\mathcal{L}|}$, and in particular a convex, closed, non-empty subset of $[0, 1]^{|\mathcal{L}|}$. By convexity we have the existence and the unicity of a solution. This leads to $p = P_{\mathcal{L},\mathcal{C}}(\mathrm{f})$ where $P_{\mathcal{L},S}$ denotes the orthogonal projection on a set S with respect to $\langle .,.\rangle_{\mathcal{L}}$.

From these results, we construct the partition $\mathcal{L}$ recursively, by splitting a leaf on a breakpoint of dimension k, $2 \le k \le d$ chosen to minimise the ISE. Rescaling the new

leaves to $\mathbb{I}$ allows to start over and split again, until a proper stopping condition is reached (no more points, independence, or no further reduction in the loss). A localised dimension reduction procedure was included by testing for independence through a Monte-Carlo resampling of a specific test statistic, allowing a selection a splitting dimensions. A summary of the procedure is given in Algorithm 1.

Algorithm 1: CORT estimation

Data: Observed ranks $\boldsymbol{u}_1, .., \boldsymbol{u}_n \in \mathbb{I}$
Result: Parameters p and $\mathcal{L}$ of the estimated piecewise linear copula
Initialize the tree by $\mathcal{L} = \{\mathbb{I}\}$ and $p_{\mathcal{L}} = \{1\}$.
while *there exist leaves* $\ell \in \mathcal{L}$ *that are still splittable* **do**
  foreach $\ell \in \mathcal{L}$ *that is splittable* **do**
    Select some splitting dimensions $D \subset \{1, ..., d\}$ through the localised dimension reduction procedure.
    if $D^* \neq \emptyset$ **then**
      Find a breakpoint x minimizing the surrogate loss.
      Split the leaf ℓ on this breakpoint and dimensions.
    end
  end
end
Compute weights p through the quadratic program detailed above.
return p *and* $\mathcal{L}$

The resulting estimator is called CORT for Copula Recursive Tree. Note that since constraints takes degree of freedoms, splits are done in a higher dimension than in the density estimation procedure from Ram and Gray (2011). Furthermore, the quadratic program for the weigths has no closed-form solution and is therefore only computed once at the end.

We extend a consistency result from Ram and Gray (2011) which was related to $f^{(n)}_{\mathrm{f},\mathcal{L}}$, the density estimator without the marginal constraints. Under certain conditions they showed, based on a generalisation by Vapnik-Chervonenkis of the Glivenko-Cantelli Theorem, that $\|f^{(n)}_{\mathrm{f},\mathcal{L}} - f\|_2^2 \to 0$, a.s. We provide the following extension:

Theorem 1 (Consistency of the CORT estimator) *For c the density of the true copula, assuming the diameter of the leaves goes to 0 as n goes to ∞, the estimator $c^{(n)}_{p_{\mathcal{L}}}$ is consistent, i.e:*

$$\mathbb{P}\left(\lim_{n\mapsto+\infty} \|c^{(n)}_{p,\mathcal{L}} - c\|_2^2 = 0\right) = 1$$

The detailed proof is given in Laverny et al. (2020). It mainly uses the projection $P_{\mathcal{L},\mathcal{C}}$ on the normed space $(\mathbb{R}^n, \|.\|^2_{\mathcal{L}})$.

4 Numerical aspects

Bagging and cross validation procedures can be used in density estimation: using out-of-bag samples, Wu, Hou, and Yang (2017) poposed definitions of an out-of-bag ISE and an out-of-bag Kullback-Leibler divergence. We extended these principle to provide other out-of-bags metrics, based on L2 distances between densities or distribution functions. These metrics allow to check the performance of the copula recursive trees alone, and to check the improvement provided by the forest principle.

The current implementation of this algorithm is avaliable on CRAN, as an R package[1]. Simulations studies are avalibales as vignettes of this package, including datasets, and deeper analysis can be found in Laverny et al. (2020). Overall, we concluded on several datasets that the model performs as expected.

Bibliography

F. Durante, J. Fernández-Sánchez, J. J. Quesada-Molina and M. Úbeda-Flores (2015). *Convergence results for patchwork copulas.* In: European Journal of Operational Research 247.2, pp. 525-531.

P. Ram and A. G. Gray (2011). *Density estimation trees.* In: Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining, pp. 627-635.

K. Wu, W. Hou, and H. Yang (2017). *Density estimation via the random forest method.* In: Communications in Statistics-Theory and Methods, 47(4), pp. 877-889

A. Sklar (1959). *Fonctions de répartition à n dimension et leurs marges.* In: Université Paris 8.3.2 pp. 1-3.

T. Nagler and C. Czado. (2016). *Evading the Curse of Dimensionality in Nonparametric Density Estimation with Simplified Vine Copulas.* In: Journal of Multivariate Analysis 151, pp. 69–89. issn: 0047259X.

Y. Li, X. Liu, and F. Liu. (2018) *PANDA: AdaPtive Noisy Data Augmentation for Regularization of Undirected Graphical Models*

O. Laverny, V.Maume-Deschamps, E. Masiello, and D. Rullière. (2020) *Dependence structure estimation using Copula Recursive Trees.* arXiv preprint arXiv:2005.02912

[1]See https://cran.r-project.org/web/packages/CORT/index.html

Estimation des paramètres d'un modèle de culture à partir de données de plein champ et de données de plateforme de phénotypage

Jean-Benoist Leger [1] & Estelle Kuhn [2] & Boris Parent [3] & François Tardieu [4] Claude Welcker [5]

[1] *UTC, CNRS, UMR 7253 Heudiasyc, Compiègne, France*
[2] *Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France, estelle.kuhn@inrae.fr*
[3] *INRAE, LEPSE, 2 Place Pierre Viala, 34000 Montpellier, boris.parent@inrae.fr*
[4] *INRAE, LEPSE, 2 Place Pierre Viala, 34000 Montpellier, francois.tardieu@inrae.fr*
[5] *INRAE, LEPSE, 2 Place Pierre Viala, 34000 Montpellier, claude.welcker@inrae.fr*

Résumé. Les modèles de culture élaborés par des écophysiologistes décrivent les processus de développement d'une plante. Ils permettent en particulier de rendre compte des différences de comportement de plusieurs variétés dans différents environnements, dues aux interactions génotype-environnement. Pour les utiliser à des fins prédictives, il est nécessaire de calibrer auparavant leurs paramètres. Nous considérons le modèle de culture APSIM et proposons un modèle joint bayésien à effets mixtes dans lequel nous inférons la valeur des paramètres inconnus à partir de données issues d'expérience de plein champ et mesurées en plateforme de phénotypage. Nous choisissons des lois a priori informatives pour intégrer les connaissances d'expert et implémentons un algorithme de type Gibbs hybride pour simuler la loi a posteriori. Les résultats obtenus sur données simulées et réelles mettent en évidence le gain obtenu sur la précision des estimations en utilisant les données issues de plateforme de phénotypage en sus des données du champ.

Mots-clés. Modèle de culture, données hétérogènes, modèles à effets mixtes, modèle bayésien, algorithme Gibbs hybride.

Abstract. Crop models were developed by ecophysiologists to describe plant development. They allow in particular to report difference existing between several genotypes in several environments, due to genotype by environment interaction. It is first necessary to calibrate these models to use them for prediction purpose. We consider the crop model APSIM and present a joint bayesian model with mixed effects. We infer models parameter values from data collected in the field and in phenotyping platform. Prior distribution are chosen in order to integrate expert knowledge. We implement an hybrid Gibbs algorithm to simulate the posterior distribution. Results obtained from simulated and real data highlight clearly the advantage of using phenotyping platform data in addition to field data.

Keywords. crop model, heterogeneous data, mixed effects models, bayesian model, Gibbs hybrid algorithm.

1 Introduction

1.1 Contexte de l'amélioration des plantes

Un des enjeux actuels en sciences du végétal vise à mieux comprendre les mécanismes impliqués dans le développement des plantes et leurs réponses aux conditions environnementales. Ces mécanismes diffèrent d'une espèce à l'autre, chacune ayant des phases de développement spécifiques. Au sein d'une même espèce, les différents processus qui se succèdent au cours de la croissance ont lieu de façon différente selon la variété considérée : ils se réalisent plus ou moins rapidement, à des périodes plus ou moins précoces, donnant lieu à une importante variabilité de comportements. Mieux comprendre comment cette variabilité dans les processus de croissance est reliée au génotype caractérisant la variété est un objectif essentiel en amélioration des plantes.

De plus, une forte interaction existe entre la variété considérée et l'environnement, incluant non seulement les aspects météorologiques mais également la composition du sol, les intrants, etc. Ainsi, une même variété va évoluer différemment dans différents environnements et différentes variétés vont se comporter différemment dans un même environnement (cf. Millet et al. (2016)). Mieux comprendre ces interactions génotype-environnement-conduite de culture est un levier important pour fournir de meilleures recommandations dans le choix des variétés selon l'environnement, en particulier dans un contexte de changement climatique fort, incluant de plus en plus d'événements climatiques extrêmes (cf. Millet et al. (2019)).

La modélisation mathématique est un outil extrêmement pertinent pour mieux comprendre, quantifier et prédire ces interactions. Des modèles linéaires ont tout d'abord été utilisés, donnant lieu à un cadre relativement limité du point de vue de la modélisation des effets génotypiques et environnementaux. Plus récemment, des modèles descriptifs des mécanismes de croissance des plantes, appelés modèles de culture, ont été développés par des écophysiologistes des plantes, comme par exemple le modèle APSIM (cf. Keating et al (2003)). Ces modèles dynamiques rendent compte des processus qui interviennent lors du développement de la plante. Ils utilisent en entrée des variables environnementales et des paramètres dépendant du génotype et fournissent en sortie des caractères plus ou moins intégrés de la plante comme par exemple la date de floraison, le rendement, la biomasse au cours du temps. Ces modèles descriptifs permettent également de modéliser les interactions génotype-environnement-conduite de culture. Utiliser à des fins prédictives, ils sont un outil efficace pour prédire ces interactions. Cependant un grand nombre de paramètres de ces modèles sont généralement inconnus et doivent être calibrés. La valeur des paramètres peut être ajustée manuellement en comparant les sorties du modèle à des données. Cette approche requiert néanmoins beaucoup de temps, d'autant plus si le nombre de paramètres est important. Une approche plus rapide consiste à ajuster un modèle statistique basé sur le modèle de culture à partir des données disponibles, en inférant la valeur des paramètres via un estimateur statistique. Ce type d'approche reste néanmoins

difficile à mettre en oeuvre du fait de la complexité du modèle de culture et demande la mise en place de méthodes numériques efficaces. Des premières approches ont été proposées (cf. Cooper et al (2016)). Cependant elles ne permettent de traiter qu'un nombre réduit de données et d'estimer un petit nombre de paramètres.

1.2 Le modèle de culture APSIM

Nous considérons le modèle de culture Agricultural Production Systems sIMulator (APSIM) avec le module maïs et une extension du module feuille réalisée par l'unité INRAE LEPSE (Lacube et al. (2017)). Il s'agit d'un modèle à pas de temps journalier qui simule un couvert constitué de plantes moyennes en mettant à jour un vecteur de descripteurs de la plante qui évolue au cours du temps. Les entrées de ce modèle sont des covariables météorologiques, des covariables descriptives du sol et de la conduite de culture. Il fournit en sortie la date de floraison et les composantes du rendement. Certains paramètres ont un sens physique, comme par exemple l'efficacité d'interception lumineuse, d'autres ne sont pas interprétables. Par ailleurs, certains processus sont communs à toute l'espèce, les paramètres associés auront une valeur commune à tous les génotypes de cette espèce, tandis que d'autres processus sont variables génétiquement, les paramètres associés auront des valeurs dépendant du génotype, comme par exemple le nombre final de feuilles.

1.3 Les données de sources hétérogènes

Nous disposons d'un riche jeu de données issu du projet DROPS European Project incluant un panel de diversité composé de 230 génotypes hybrides observés en plein champ dans 13 conditions environnementales. Une condition environnementale est définie par un lieu et une année. Les lieux sont répartis en Europe du nord au sud, rendant compte de conditions climatiques très contrastées. Les années considérées varient de 2012 à 2013. Les données comprennent la date de floraison, les composantes du rendement (nombre de grains, poids d'un grain) et les conditions environnementales.

De plus, des expériences auxiliaires complémentaires ont été réalisées sur la plateforme INRAE PhénoArch à Montpellier (Cabrera-Bosquet et al. (2016)). Ces expériences ont permis d'obtenir des données supplémentaires sur des paramètres mécanistes intervenant dans le modèle de culture. Ainsi, des mesures de quantité telle que le nombre final de feuilles observées ont été effectuées en plateforme, apportant des informations complémentaires au jeu de données obtenu en plein champ.

L'objectif est d'utiliser les deux sources d'information champ et plateforme dans la procédure d'inférence statistique des paramètres du modèle de culture.

2 Modélisation statistique

2.1 Modélisation de la variabilité génotypique

Nous disposons pour chaque génotype du panel DROPS de mesures répétées du rendement, du nombre de grains et de la date de floraison dans M conditions environnementales. Nous considérons un modèle statistique à effets mixtes (cf. Pinheiro et Bates (2000)) basé sur le modèle de culture APSIM qui permet de prendre en compte simultanément les variabilités inter-génotype et intra-génotype. On note Y_{gm} la mesure vectorielle dans la condition expérimentale m pour le génotype g et on modélise pour tout g et tout m :

$$\log Y_{gm} = \log G(e_m, \beta_g, \gamma_g) + \varepsilon_{gm} \tag{1}$$

où G représente la sortie du modèle de culture APSIM, e_m le vecteur de variables descriptives de l'environnement m, β_g le vecteur des paramètres du génotype g non mesurées en plateforme, γ_g le vecteur des paramètres du génotype g mesurées en plateforme, ε_{gm} un terme d'erreur supposé gaussien centré de variance diagonale $(\sigma_1^2, \sigma_2^2, \sigma_3^2)$. On suppose que les vecteurs d'effets aléatoires (β_g) et (γ_g) sont indépendants identiquement distribués de loi gaussienne d'espérance $\bar{\beta}$ resp. $\bar{\gamma}$, et de matrice de covariance diagonale Σ_β, resp. Σ_γ.

2.2 Modélisation des données de plateforme de phénotypage

Nous considérons un modèle joint pour intégrer les données issues de la plateforme à l'inférence des paramètres du modèle de culture. Pour cela, nous modélisons les mesures des paramètres effectuées en plateforme de phénotypage via un modèle linéaire en fonction de la valeur des paramètres du modèle de culture APSIM. On note Z_g le vecteur de taille p des mesures de paramètres du génotype g. Pour $1 \leq l \leq p$, on a :

$$Z_{g,l} = \mu_l + \zeta_l \gamma_{g,l} + \eta_{g,l} \tag{2}$$

où μ et ζ sont des vecteurs inconnus de $\mathbb{R}^p$ et $\eta_{g,l}$ un terme d'erreur résiduel supposé gaussien centré de variance τ_l^2.

3 Inférence bayésienne du modèle joint

Du fait de la complexité du modèle de culture APSIM et du grand nombre de paramètres à estimer, nous faisons le choix d'une approche bayésienne qui va permettre de régulariser la procédure d'estimation. Nous souhaitons choisir des lois *a priori* uniformes pour les paramètres du modèle de culture qui prennent leurs valeurs dans des intervalles bornés. Toutefois, pour des raisons computationnelles, nous avons effectué une reparamétrisation du modèle, et les nouveaux paramètres sont à valeurs réelles. Nous choisissons pour ces paramètres des lois *a priori* normales, telles que leurs transformées par la reparamétrisation

inverse soient les plus proches au sens de la divergence de Kullback-Leibler des lois uniformes de départ. Pour les paramètres μ et ζ du modèle des données issues de la plateforme, nous choisissons des lois *a priori* normales centrées sur la valeur attendue, 0 pour l'ordonnée à l'origine et 1 pour la pente. Nous fixons des lois inverse gamma qui sont conjuguées pour les lois *a priori* des paramètres de variances des bruits.

Nous appliquons un algorithme de Monte Carlo Markov Chain de type Gibbs hybride (cf. Carlin et Louis (2008)) pour générer une chaine de Markov qui sous des hypothèses de régularité du modèle est ergodique et a pour loi stationnaire la loi a posteriori. A partir des réalisations de cet algorithme, nous construisons des estimateurs empiriques des lois *a posteriori* des paramètres du modèle.

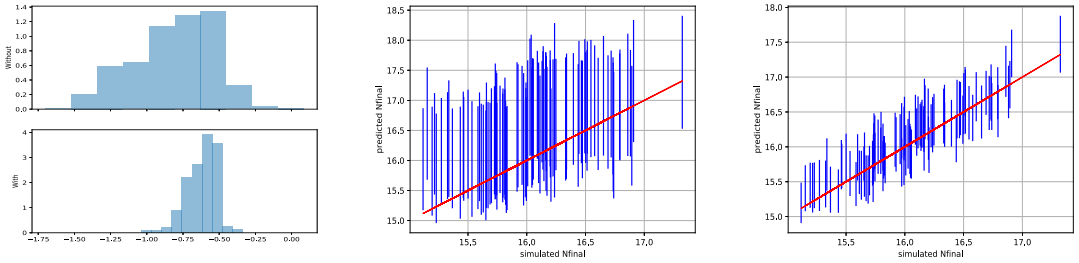


Figure 1: Histogramme de la loi a posteriori de Nfinal sans (gauche, haut) et avec (gauche, bas) les données plateforme supplémentaires ; Simulations versus prédictions via intervalles de crédibilité à 90% pour Nfinal sans (centre) et avec (droite) les données plateforme supplémentaires.

4 Expériences numériques

Nous effectuons une étude de simulation en considérant les 13 environnements réels du jeu de données DROPS et les valeurs des paramètres proches de ceux du génotype de référence *B*73. Nous simulons 100 génotypes. Nous estimons les trois paramètres du modèle correspondants au nombre final de feuilles (noté Nfinal), au premier ligulochrone et au poids moyen potentiel d'un grain, les autres étant fixés à la valeur de référence. Nous mettons en évidence que les estimateurs obtenus à partir des données issues du champ et de la plateforme dans le modèle joint sont plus précis que les estimateurs obtenus à partir des seules données issues du champ dans le modèle initial (cf Figures 1 et 2 gauche).

Nous ajustons ensuite le modèle proposé aux données réelles. Les prédictions obtenues à partir du modèle avec les paramètres estimés à partir des données issues du champ et de la plateforme sont meilleures que celles obtenues avec les paramètres estimés à partir des seules données champ (cf Figure 2 droite).

Ce travail a été financé par le projet AMAIZING ANR-10-BTBR-01. Les auteurs remercient la plateforme MIGALE, INRAE, 2020, Migale bioinformatics Facility pour les moyens de calcul et les capacités de stockage.

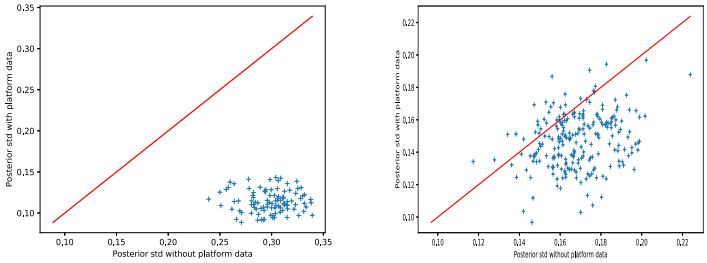


Figure 2: Ecart-types de la distribution a posteriori de Nfinal obtenus sans (abscisse) et avec (ordonnée) les données plateforme supplémentaires en simulation (gauche) et sur données réelles (droite).

Bibliographie

Cabrera-Bosquet, L., et al., (2016), High-throughput estimation of incident light, light interception and radiation-use efficiency of thousands of plants in a phenotyping platform. *New Phytologist*, 212, (1), 269-281.

Carlin, B.P. and Louis, T. (2008), Bayesian methods for data analysis, *Chapman and Hall/CRC*.

Cooper, M. and Technow, F. and Messina, C. and Gho, C and Totir, L. R. (2016), Use of crop growth models with whole-genome prediction: application to a maize multienvironment trial, *Crop Science*, 56, (5), 2141–2156.

Keating, B. and Carberry, P. and Hammer, G. and Probert, M. and Robertson, M. and Holzworth, D. and Huth, N.and Hargreaves, J. and *et al.*, (2003), An overview of APSIM, a model designed for farming systems simulation, *European journal of agronomy*, 18, (3-4), 267–288.

Lacube, S., et al., (2017) Distinct controls of leaf widening and elongation by light and evaporative demand in maize, *Plant Cell and Environment*, 40, (9), 2017-2028.

Millet E., Welcker C, Kruijer W, Negro S, Coupel-Ledru A, et al., (2016), Genome-wide analysis of yield in Europe: allelic effects vary with drought and heat scenarios, *Plant Physiol*, 172, 749-764.

Millet, E. and Kruijer, W. and Coupel-Ledru, A. and Prado, S.A. and Cabrera-Bosquet, L. and Lacube, S. and Charcosset, A. and Welcker, C. and van Eeuwijk, F. and Tardieu, F., (2019), Genomic prediction of maize yield across European environmental conditions, *Nature genetics*, 51, (6), 952–956.

Pinheiro, J.C. and Bates D.M. (2000), Mixed-Effects Models in S and S-PLUS, *Springer*.

Deconvolution with unknown noise distribution

Luc Lehéricy [1] & Élisabeth Gassiat [2] & Sylvain Le Corff [3]

[1] *Université Côte d'Azur, CNRS, Laboratoire J.A. Dieudonné, 06108 Nice, France, luc.lehericy@univ-cotedazur.fr.*

[2] *Université Paris-Saclay, CNRS, Laboratoire de mathématiques d'Orsay, 91405, Orsay, France, elisabeth.gassiat@universite-paris-saclay.fr.*

[3] *Samovar, Télécom SudParis, Département CITI, TIPIC, Institut Polytechnique de Paris, France, sylvain.le_corff@telecom-sudparis.eu.*

Résumé. Le problème de déconvolution se formule ainsi : étant donné des observations $\mathbf{Y}_i, i = 1, \dots, n$ i.i.d. qui s'écrivent

$$\mathbf{Y}_i = \mathbf{X}_i + \varepsilon_i,$$

où $\mathbf{X}_i$ est le signal et ε_i du bruit, comment retrouver la loi du signal à partir de l'échantillon $(\mathbf{Y}_1, \dots, \mathbf{Y}_n)$? Ce problème fait l'objet d'une abondante littérature lorsque la loi du bruit est connue. Toutefois, lorsque cette loi est inconnue, ce modèle n'est en général plus identifiable. Dans ce travail, nous montrons que si le signal et le bruit peuvent être séparés en deux composantes $\mathbf{X} = (X^{(1)}, X^{(2)})$ et $\varepsilon = (\varepsilon^{(1)}, \varepsilon^{(2)})$ telles que la queue de distribution du signal est assez légère et que les composantes du bruit sont indépendantes, alors sous une hypothèse sur les zéros de la fonction génératrice des moments du couple $(X^{(1)}, X^{(2)})$, le modèle est identifiable. Nous introduisons également un estimateur non paramétrique et montrons qu'il est consistant pour la topologie de la convergence en loi.

Mots-clés. estimation non paramétrique, déconvolution, identifiabilité, modèle à variable latente.

Abstract. In deconvolution problem, i.i.d. observations $\mathbf{Y}_i, i = 1, \dots, n$, are generated by

$$\mathbf{Y}_i = \mathbf{X}_i + \varepsilon_i,$$

where $\mathbf{X}_i$ is the signal and ε_i is the noise. The objective is to infer the distribution of the latent data based on $(\mathbf{Y}_1, \dots, \mathbf{Y}_n)$. There is a wide range of literature on density deconvolution when the distribution of the noise ε_i is assumed to be known. However, when the distribution of the noise is also unknown, this model can not be identified in full generality. We show that when the signal and noise can be split in two components $\mathbf{X} = (X^{(1)}, X^{(2)})$ and $\varepsilon = (\varepsilon^{(1)}, \varepsilon^{(2)})$ such that the tails of the signal distribution is light enough and the noise components are independent, the model is identifiable under an assumption on the zeros of the moment generating function of the couple $(X^{(1)}, X^{(2)})$. We also introduce a nonparametric estimator and show its consistency with respect to the weak convergence topology.

Keywords. nonparametric estimation, deconvolution, identifiability, latent data model.

1 Introduction

Le problème de déconvolution a été l'objet d'un important travail lorsque la distribution du bruit est supposée connue, voir par exemple [Devroye, 1989], [Liu and Taylor, 1989], [Stefanski and Carroll, 1990], pour les premières méthodes de déconvolution non paramétriques, [Carroll and Hall, 1988] et [Fan, 1991] pour les vitesses minimax, ou encore [Dedecker et al., 2015] et ses références pour un travail récent. Toutefois, tous ces travaux supposent la loi du bruit connu. Et en effet, si la loi du bruit est inconnue, le modèle de déconvolution n'est en général pas identifiable, et *a fortiori* l'estimation de la loi du signal est impossible.

L'enjeu du présent travail est de montrer que dans le cas particulier où les observations sont générées par paires $(Y^{(1)}, Y^{(2)}) = (X^{(1)}, X^{(2)}) + (\varepsilon^{(1)}, \varepsilon^{(2)})$, il est possible, sous une hypothèse non paramétrique faible sur le signal $\mathbf{X}$ et sans aucune hypothèse sur le bruit ε (en dehors de l'indépendance de $\varepsilon^{(1)}$ et $\varepsilon^{(2)}$), de retrouver la loi du signal à partir de la loi des observations.

Une fois ce résultat d'identifiabilité établi, nous l'utilisons pour montrer la consistance de deux estimateurs non paramétriques et illustrons leurs performances sur des données simulées.

Cette soumission est issue de [Gassiat et al., 2020]. Nous ne décrivons pas ici l'estimateur du maximum de vraisemblance introduit dans l'article, qui est également consistant.

2 Résultats

2.1 Identifiabilité

Soient $\mathbf{X} = (X^{(1)}, X^{(2)})$ et $\varepsilon = (\varepsilon^{(1)}, \varepsilon^{(2)})$ des variables aléatoires telles que pour tout $i \in \{1, 2\}$, $X^{(i)}$ et $\varepsilon^{(i)}$ sont à valeurs dans $\mathbb{R}^{d_i}$, ε est indépendant de $\mathbf{X}$ et $\varepsilon^{(1)}$ est indépendant de $\varepsilon^{(2)}$. Pour $i \in \{1, 2\}$, soit

$$Y^{(i)} = X^{(i)} + \varepsilon^{(i)} .$$

Soit $\mathbb{P}_{R,P}$ la loi de $\mathbf{Y} = (Y^{(1)}, Y^{(2)})$ lorsque $\mathbf{X}$ a pour loi R et pour tout $i \in \{1, 2\}$, $\varepsilon^{(i)}$ a pour loi P_i, où $P = (P_1, P_2)$. Notons R_1 la loi marginale de X_1 et R_2 la loi marginale de X_2. Pour tout $\rho > 0$ et tout $d \geqslant 1$, soit $\mathcal{M}_\rho^d$ l'ensemble des mesures μ sur $\mathbb{R}^d$ telles qu'il existe $A, B > 0$ tels que pour tout $\lambda \in \mathbb{R}^d$,

$$\int \exp\left(\lambda^\top x\right) \mu(\mathrm{d}x) \leqslant A \exp\left(B \|\lambda\|^\rho\right) ,$$

où on note $\|\lambda\|$ la norme euclidienne du vecteur λ. Si R est une mesure de probabilité sur $\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ telle que $R_1 \in \mathcal{M}_\rho^{d_1}$ et $R_2 \in \mathcal{M}_\rho^{d_2}$, alors l'extension Φ_R sur $\mathbb{C}^{d_1} \times \mathbb{C}^{d_2}$ de la

fonction caractéristique de R, définie par

$$\begin{aligned}\Phi_R : \mathbb{C}^{d_1} \times \mathbb{C}^{d_2} &\longrightarrow \mathbb{C} \\ (z_1, z_2) &\longmapsto \int \exp\left(iz_1^\top x_1 + iz_2^\top x_2\right) R(\mathrm{d}x_1, \mathrm{d}x_2)\,,\end{aligned}$$

est une fonction analytique.

H1 Pour tout $z_0 \in \mathbb{C}^{d_1}$, $z \mapsto \Phi_R(z_0, z)$ n'est pas la fonction nulle et pour tout $z_0 \in \mathbb{C}^{d_2}$, $z \mapsto \Phi_R(z, z_0)$ n'est pas la fonction nulle.

Théorème 1. *Soient R et $\widetilde{R}$ des mesures de probabilité sur $\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ qui vérifient H1. Supposons également qu'il existe $\rho < 2$ tel que R_1 et $\widetilde{R}_1$ sont dans $\mathcal{M}_\rho^{d_1}$ et R_2 et $\widetilde{R}_2$ sont dans $\mathcal{M}_\rho^{d_2}$. Alors $\mathbb{P}_{R,P} = \mathbb{P}_{\widetilde{R},\widetilde{P}}$ implique $R = \widetilde{R}$ et $P = \widetilde{P}$ à translation près.*

Ce théorème ne fait absolument aucune hypothèse sur la loi du bruit. La condition $\rho < 2$ implique notamment que la queue de distribution X est plus légère qu'une queue gaussienne. La condition « à translation près » est inévitable : il est toujours possible d'ajouter une constante à $\mathbf{X}$ et de la retrancher à ε sans changer la loi des observations.

2.2 Consistance

Dans cette section, nous énonçons la consistance d'un estimateur des moindres carrés. Nous accolons une étoile $\star$ aux vrais paramètres du modèle.

Soit $\mathcal{S}$ un voisinage compact de 0 dans $\mathbb{R}^{d_1+d_2}$. Soient $\Phi_{P_1^\star}$ et $\Phi_{P_2^\star}$ les fonctions caractéristiques de ε_1 et ε_2. Pour toute mesure de probabilité R sur $\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$, soit

$$M(R) = \int_{\mathcal{S}} \left|\Phi_{R^\star}(t_1,t_2)\Phi_R(t_1,0)\Phi_R(0,t_2) - \Phi_R(t_1,t_2)\Phi_{R^\star}(t_1,0)\Phi_{R^\star}(0,t_2)\right|^2 \\ |\Phi_{P_1^\star}(t_1)\Phi_{P_2^\star}(t_2)|^2 \mathrm{d}t_1\mathrm{d}t_2\,.$$

Cette quantité apparaît dans la preuve du Théorème 1, qui assure que $M(R) = 0$ si et seulement si $R = R^\star$ (à translation près). Étant donné un estimateur $\widehat{\Phi}_n$ de la fonction caractéristique de (Y_1, Y_2), posons

$$M_n(R) = \int_{\mathcal{S}} \left|\widehat{\Phi}_n(t_1,t_2)\Phi_R(t_1,0)\Phi_R(0,t_2) - \Phi_R(t_1,t_2)\widehat{\Phi}_n(t_1,0)\widehat{\Phi}_n(0,t_2)\right|^2 \mathrm{d}t_1\mathrm{d}t_2\,.$$

Soit $\mathcal{R}$ un ensemble de mesures de probabilité sur $\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ telles qu'il existe $\rho < 2$ tel que pour tout $R \in \mathcal{R}$, les distributions marginales de R sont dans $\mathcal{M}_\rho$ et R vérifie H1. On prend comme estimateur $\widehat{R}_n$ un élément de $\mathcal{R}$ qui vérifie

$$M_n(\widehat{R}_n) = \inf_{R \in \mathcal{R}} M_n(R).$$

Théorème 2. *Supposons que $\mathcal{R}$ est compact pour la topologie de la convergence en loi, que $R^\star \in \mathcal{R}$ et que*

$$\sup_{(t_1,t_2)\in\mathcal{S}} |\widehat{\Phi}_n(t_1,t_2) - \Phi_{R^\star}(t_1,t_2)\Phi_{P_1^\star}(t_1)\Phi_{P_2^\star}(t_2)| = o_{\mathbb{P}_{R^\star,P^\star}}(1)\,. \tag{1}$$

Soit d une distance qui métrise la convergence en loi sur $\mathcal{R}$. Alors $d(\widehat{R}_n, \mathcal{R}^\star)$ tend vers 0 en probabilité sous $\mathbb{P}_{R^\star,P^\star}$, où $\mathcal{R}^\star$ est l'ensemble des $R \in \mathcal{R}$ égaux à $R^\star$ à translation près.

2.3 Simulations

Les simulations sont fondées sur le modèle suivant : $Z_0 \sim U(0, 2\pi)$ et pour tout $k \geqslant 1$,

$$Z_k = Z_{k-1} + \sigma_x \varepsilon_k\,, \quad X_k = \cos(Z_k) \quad \text{et} \quad Y_k = X_k + \sigma_y \eta_k\,,$$

où $(\sigma_x, \sigma_y) \in \mathbb{R}_+^* \times \mathbb{R}_+^*$ et où $(\varepsilon_k, \eta_k)_{k\geqslant 1}$ sont des v.a. gaussiennes centrées réduites i.i.d. et indépendantes de Z_0. Nous avons utilisé $(\sigma_x, \sigma_y) = (0.1, 0.1)$ et échantillonné $n = 100\,000$ observations. Les estimateurs se fondent sur les couples $((Y_k, Y_{k+1}))_{1\leqslant k\leqslant n}$.

La Figure 1 illustre qualitativement la convergence de l'estimateur ci-dessus et de l'estimateur du maximum de vraisemblance de [Gassiat et al., 2020]. Plus de simulations peuvent être trouvées dans [Gassiat et al., 2020, Sections 4 et C].

Notez que même si les variables aléatoires $(\mathbf{Y}_k)_{k\geqslant 1} = ((Y_k, Y_{k+1}))_{k\geqslant 1}$ ne sont pas i.i.d., l'équation (1) est vérifiée en prenant $\widehat{\Phi}_n(t) = \frac{1}{n}\sum_{k=1}^n e^{it_1^\top Y_k + it_2^\top Y_{k+1}}$ car $(X_k)_{k\geqslant 1}$ est une chaîne de Markov fortement ergodique. La moyenne empirique se comporte alors essentiellement comme dans le cas i.i.d., pour lequel (1) est vérifiée dès que ε admet un moment d'ordre 2, voir par exemple la procédure suivie dans la Proposition 13 de [De Castro et al., 2016]. Le choix de $\mathcal{R}$ comme l'ensemble des lois à densité constante par morceaux sur un découpage fixé ne permet par contre pas de satisfaire l'hypothèse $R^\star \in \mathcal{R}$. En pratique, il faut ajuster la taille du découpage avec n pour retrouver la vraie densité.

Références

[Carroll and Hall, 1988] Carroll, R. and Hall, P. (1988). Optimal rates of convergence for deconvolving a density. *J. Amer. Statist. Assoc.*, 83(404) :1184–1186.

[De Castro et al., 2016] De Castro, Y., Gassiat, E., and Lacour, C. (2016). Minimax adaptive estimation of nonparametric hidden Markov models. *J. Mach. Learn. Res.*, 17(111) :1–43.

[Dedecker et al., 2015] Dedecker, J., Fischer, A., and Michel, B. (2015). Improved rates for Wasserstein deconvolution with ordinary smooth error in dimension one. *Electron. J. Stat.*, 9(1) :234–265.

[Devroye, 1989] Devroye, L. (1989). Consistent deconvolution in density estimation. *Canad. J. Statist.*, 17(2) :235–239.

[Fan, 1991] Fan, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *Ann. Statist.*, 19(3) :1257–1272.

[Gassiat et al., 2020] Gassiat, É., Le Corff, S., and Lehéricy, L. (2020). Identifiability and consistent estimation of nonparametric translation hidden Markov models with general state space. *Journal of Machine Learning Research*, 21(115) :1–40.

[Liu and Taylor, 1989] Liu, M. C. and Taylor, R. L. (1989). A consistent nonparametric density estimator for the deconvolution problem. *Canad. J. Statist.*, 17(4) :427–438.

[Stefanski and Carroll, 1990] Stefanski, L. and Carroll, R. J. (1990). Deconvoluting kernel density estimators. *Statistics*, 21(2) :169–184.

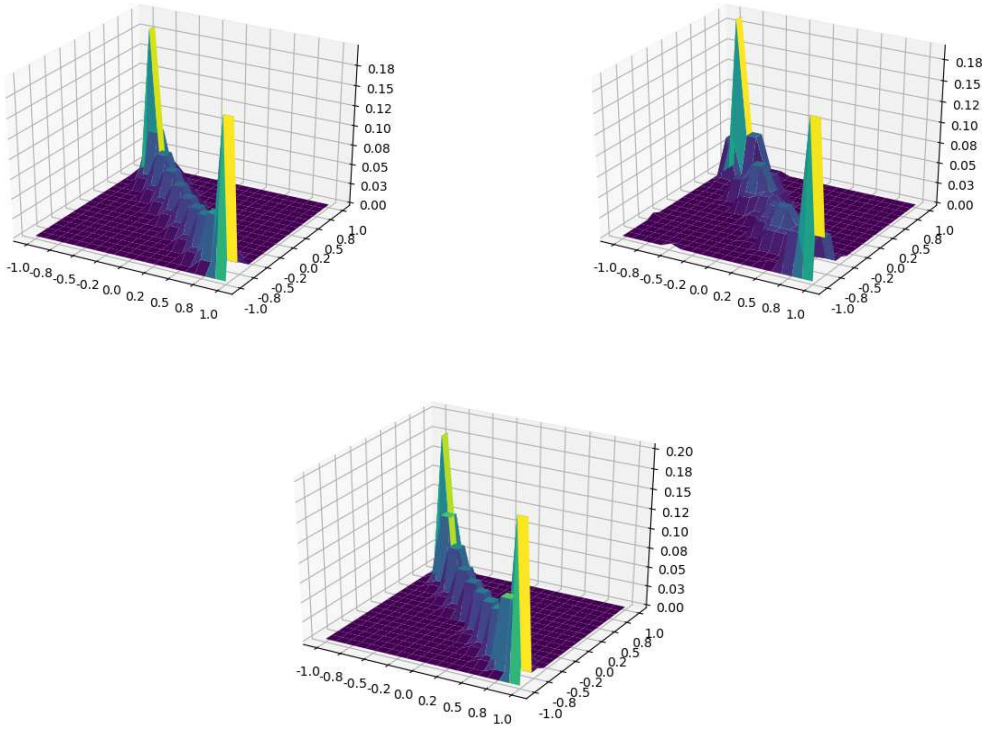


FIGURE 1 – Approximation par des fonctions constantes par morceaux de la densité de $(X^{(1)}, X^{(2)})$. En haut à gauche : estimateur oracle (lorsque les états $\mathbf{X}$ sont observés). En haut à droite : estimateur de la Section 2.2. En bas : estimateur du maximum de vraisemblance (voir [Gassiat et al., 2020]).

Modèle de régression multi-tâche par processus gaussiens avec moyenne informée

Arthur LEROY[*]
Servane GEY[†]
Pierre LATOUCHE[‡]
Benjamin GUEDJ[§]

Résumé : La régression par processus gaussien est un outil classique de l'apprentissage supervisé qui présente l'avantage de fournir un cadre probabiliste permettant une quantification de l'incertitude des prédictions. L'apprentissage dans de tels modèles se concentre généralement sur l'estimation des hyper-paramètres du noyau associé à la structure de covariance, et non sur la moyenne a priori du processus. Ainsi, la qualité de prédiction peut se retrouver fortement affectée dès lors que l'on s'éloigne des points d'observation avec une moyenne a priori non pertinente. Ce travail propose l'utilisation d'un modèle multi-tâche, dont les données de plusieurs individus sont supposées partager une structure commune. Cette structure est modélisée par un processus moyen, commun à tous les individus, permettant une prediction plus pertinente même lorsque les observations d'un nouvel individu sont peu nombreuses ou mal reparties sur l'espace des entrées. L'apprentissage des hyper-paramètres du modèle est effectué par un algorithme EM parallélisable. Une étape de calcul de la loi a posteriori du processus moyen est ensuite effectuée pour intégrer l'information de tous les individus des données d'apprentissage. Enfin, après marginalisation, l'étape de prédiction est analogue au cas classique de la régression par processus gaussien. Les prédictions ainsi obtenues pour un nouvel individu sont centrées sur des valeurs informées par les autres individus, avec prise en compte des deux sources d'incertitude distinctes. Une étude de simulations illustre les bénéfices et les coûts calculatoires d'une telle approche.

Mots clefs. Processus Gaussiens, Apprentissage multi-tâche, Algorithme EM, ...

Abstract. Gaussian process regression is a common tool of supervised learning that provides a convenient probabilistic framework, leading to predictions associated to uncertainty quantification. The learning step in such a model generaly focuses on hyper-parameters estimation of the kernel associated to the covariance structure, rather than the prior mean of the process. Therefore, the quality of prediction might severely decrease, with an unappropriate prior mean, as we move away from observation points. This work presents a multi-task model, where data come from several individuals supposed to share some structure altogether. We model this structure through a mean process, common to all individuals, leading to more reliable predictions even though a new individual is observed on few or sparse input locations. An EM algorithm, that can be parallelized, is used to learn the model hyper-parameters. An additional step integrates knowledge from all individuals in the training dataset through the computation of the mean process posterior. Finally, after marginalization, the prediction step is analogous to the classic gaussian process regression. Such predictions for a new individual are centered on values informed by the other individuals, with uncertainty coming from the two distinct sources. A simulation study illustrates the advantages and computational costs of such an approach.

Keywords. Gaussian Processes, Multi-task learning, EM algorithm, ...

Contexte

L'apprentissage par processus gaussiens (GP, pour 'Gaussian Processes' en anglais) a fait l'objet de nombreuses études durant ces deux dernières décennies, dont l'ouvrage de référence sur le sujet est Rasmussen and Williams (2006). Cette approche non-paramétrique permet d'estimer une fonction d'apprentissage dans

[*]MAP5 - Université de Paris, arthur.leroy@parisdescartes.fr
[†]MAP5 - Université de Paris, servane.gey@parisdescartes.fr
[‡]MAP5 - Université de Paris, pierre.latouche@parisdescartes.fr
[§]MODAL - INRIA / University College London, benjamin.guedj@inria.fr

un cadre probabiliste en posant une hypothèse a priori sur la forme de cette fonction. La simplicité de la méthode, ainsi que la possibilité d'obtenir des intervalles de crédibilité pour les prédictions, ont participé à son succès. Néanmoins, son coût calculatoire en $\mathcal{O}(N^3)$ en complique l'utilisation pour des grands jeux de données.

Une littérature très riche se concentre sur la problématique de donner de bonnes approximations lorsque le nombre N d'observations d'un processus est trop grand pour permettre l'inférence exacte. Dans le cas de l'étude de plusieurs tâches, ou individus, ayant chacun un nombre d'observations raisonnable, une littérature existe au croisement entre l'apprentissage multi-tâche et les GPs.

Un des premiers papiers, Yu, Tresp, and Schwaighofer (2005), qui développe un modèle bayésien hiérarchique, définit des paramètres de moyenne et de covariance communs entre les individus et estimés par un algorithme EM. Ensuite, Bonilla, Chai, and Williams (2008) se concentrent sur le partage de la structure de covariance entre les différents individus, alors que dans J. Q. Shi and Wang (2008), les auteurs proposent un premier modèle dans lequel il existe une fonction moyenne commune, à estimer, pour chaque GP individuel. Cette approche fait l'objet de plusieurs articles, ainsi que d'un package R associé, Jian Qing Shi and Cheng (2014). Pour estimer cette fonction moyenne, les auteurs proposent d'utiliser une décomposition dans une base de B-splines. Cela constitue donc une approche paramétrique déterministe, ignorant l'incertitude liée à l'estimation de cette fonction moyenne. L'objectif de notre travail réside donc dans la définition d'une extension de cette idée, en modélisation la moyenne a priori également par un GP, qui sera alors estimé à l'aide de tous les individus. Cette approche présente l'avantage d'offrir un cadre non-paramétrique global et une prise en compte de l'incertitude liée au processus moyen commun. Le coût calculatoire à payer en contrepartie reste raisonnablement du même ordre tant que l'union des temps d'observations (les entrées sont souvent assimilées au temps dans la littérature) de tous les individus ne grandit pas outre mesure.

A noter que l'article Yang et al. (2016) présente des idées du même ordre dans un cadre un peu différent, en définissant un modèle hiérarchique et un algorithme MCMC associé. Le coût calculatoire de cette procédure pouvant être rapidement trop important, nous tentons également dans notre approche d'utiliser les propriétés agréables des GPs pour calculer exactement les log-vraisemblances utilisées dans l'apprentissage, qui est effectué à l'aide d'un algorithme EM.

Quelques notations

- M le nombre d'individus,
- N_i le nombre d'observations pour l'individu i,

On dispose d'un échantillon $\{(\mathbf{y}_i, \mathbf{t}_i)\}_{i=1,\dots,M}$, tel que:

- $\mathbf{t}_i = (t^1, \dots, t^{N_i})^T$ le vecteur des temps de l'individu i,
- $\mathbf{y}_i = y_i(\mathbf{t}_i) = (y_i(t^1), \dots, y_i(t^{N_i}))^T$ le vecteur des observations de l'individu i,
- $\mathbf{t} = \bigcup_i \mathbf{t}_i$ l'union de tous les points de temps observés,
- $N = card(\mathbf{t})$, le nombre total de points de temps distincts,
- K_{θ_0} un noyau d'hyper-paramètres θ_0,
- m_0 une fonction réelle donnée comme moyenne a priori pour le processus moyen, noté μ_0,
- $(\Sigma_{\theta_i})_i$ un ensemble de noyaux de même forme et d'hyper-paramètres respectifs $(\theta_i)_i$,
- $\sigma_i^2 \in \mathbb{R}$, $\forall i$,
- $\Theta = \{\theta_0, (\theta_i)_i, (\sigma_i^2)_i\}$ le vecteur des hyper-paramètres du modèle.

Modèle et hypothèses

On pose le modèle suivant:

$$\forall i, \forall t, \quad y_i(t) = \mu_0(t) + f_i(t) + \epsilon_i(t),$$

- $\mu_0(\cdot) \sim GP(m_0(\cdot), K_{\theta_0}(\cdot, \cdot))$,

- $f_i(\cdot) \sim GP(0, \Sigma_{\theta_i}(\cdot,\cdot))$, $(f_i)_i \perp\!\!\!\perp$,
- $\epsilon_i(t) \sim \mathcal{N}(0, \sigma_i^2)$, $(\epsilon_i)_i \perp\!\!\!\perp, \forall t \in \mathbb{R}$,
- $\mu_0 \perp\!\!\!\perp (f_i)_i$.

On note $\Psi_i(\cdot,\cdot) = \Sigma_{\theta_i}(\cdot,\cdot) + \sigma_i^2 I_d$, et on en déduit:

$$y_i(\cdot)|\mu_0 \sim GP(\mu_0(\cdot), \Psi_i(\cdot,\cdot)), \quad (y_i|\mu_0)_i \perp\!\!\!\perp$$

Si on applique cela à notre échantillon, on a la loi a priori suivante:

$$y_i(\mathbf{t}_i)|\mu_0(\mathbf{t}_i) \sim \mathcal{N}(\mu_0(\mathbf{t}_i), \Psi_i(\mathbf{t}_i, \mathbf{t}_i)),$$

avec

$$\Psi_i(t_k, t_l) = cov(y_i(t_k), y_i(t_l)), \quad \forall t_k, t_l \in \mathbf{t}_i.$$

L'apprentissage

Dans un modèle de régression GP, il est généralement nécessaire d'apprendre un nombre limité d'hyper-paramètres, qui caractérisent le noyau associé à la fonction de covariance. Dans notre cadre, puisque les observations sont supposées être la somme de deux GPs indépendants, il faut également apprendre les hyper-paramètres θ_0 du noyau K_{θ_0} du processus moyen. De plus, nous verrons dans l'étape de prédiction que ce processus moyen est estimé grâce aux observations issues de nos échantillons d'apprentissage. Il est donc aussi nécessaire d'apprendre les (θ_i, σ_i^2), pour tout $i = 1, \dots, M$. Il est important de noter que le processus moyen μ_0 est par définition commun à tous les individus, et son estimation est dépendante des hyper-paramètres Θ. Une procédure classique dans ce contexte est l'utilisation d'un algorithme EM, qui procède en alternance au calcul de la loi de μ_0 avec Θ fixé, puis à l'estimation des hyper-paramètres par maximisation de log-vraisemblance (qui fait intervenir $p(\mu_0)$). Généralement, un tel algorithme itératif converge vers des maxima locaux en peu d'itérations, et différents choix d'initialisations peuvent aider à trouver un maximum global.

Etape E:

$$\begin{aligned} p(\mu_0(\mathbf{t})|(\mathbf{y}_i)_i, \Theta) &\propto \overbrace{p((\mathbf{y}_i)_i|\mu_0(\mathbf{t}), (\sigma_i^2)_i, (\theta_{\theta_i})_i))}^{\prod_{i=1}^{M} \mathcal{N}(\mu_0(\mathbf{t}), \Psi_i)} \overbrace{p(\mu_0(\mathbf{t})|\theta_0)}^{\mathcal{N}(m_0, K_{\theta_0})} \\ &= \mathcal{N}(\hat{m}_0(\mathbf{t}), \hat{K}), \end{aligned}$$

avec:

- $\hat{K} = \left((K_{\theta_0})^{-1} + \sum_{i=1}^{M} (\Psi_i)^{-1} \right)^{-1}$,
- $\hat{m}_0(\mathbf{t}) = \hat{K} \left(K_{\theta_0}^{-1} m_0 + \sum_{i=1}^{M} (\Psi_i)^{-1} \mathbf{y}_i \right)$.

NB: les matrices $(\Psi_i)_i$ et vecteurs $\boldsymbol{y}_i$ sont en fait ici complétés avec des 0 pour être de dimension N, tout comme $\mu_0(\boldsymbol{t})$. Il s'agit de détails techniques que nous omettons par soucis de concision.

Etape M:

$$\hat{\Theta} = arg \max_{\Theta} \mathbb{E}_{\mu_0} \left[log\ p(\mu_0(\mathbf{t}), (\mathbf{y}_i)_i, \Theta) \right]$$

Il est intéressant de noter que cette maximisation, relativement compliquée à première vue, peut se découper assez simplement en $M+1$ maximisations par indépendance des individus entre eux. Ce qui a pour avantage de devoir optimiser à chaque fois une fonction avec un faible nombre de paramètres, et de pouvoir effectuer les calculs indépendamment en parrallèle les uns des autres.

La prédiction

L'objectif réside dans la prédiction de la quantité $p(y_*(t)|y_*(\mathbf{t}_*),(\mathbf{y}_i)_i)$, définissant la loi d'un nouvel individu d'indice $*$ en un temps non observé t, en connaissant les observations de ce nouvel individu et les observations de tous les autres individus.

Calcul de la loi a posteriori de μ_0

Une fois les hyper-paramètres du modèle appris, il est nécessaire d'effectuer le calcul de la loi à posteriori du processus moyen μ_0, pour les temps auxquels on souhaite obtenir une prédiction. Il s'agit d'une étape spécifique à notre procédure, qui n'existe pas dans une régression GP classique, et qui permet d'obtenir une moyenne informée pour notre prédiction finale.

Par concision, on note $\mathbf{t}_+ = (t, \mathbf{t}_*)$ le vecteur de tous les temps observés, auquel on ajoute le temps à prédire.

$$\begin{aligned} p(y_*(t), y_*(\mathbf{t}_*)|(\mathbf{y}_i)_i) &= \int p(y_*(t), y_*(\mathbf{t}_*), \mu_0(\mathbf{t}_+)|(\mathbf{y}_i)_i)\ d\mu_0(\mathbf{t}_+) \\ &= \int p(y_*(t), y_*(\mathbf{t}_*)|(\mathbf{y}_i)_i, \mu_0(\mathbf{t}_+))\ p(\mu_0(\mathbf{t}_+)|(\mathbf{y}_i)_i)\ d\mu_0(\mathbf{t}_+) \\ &\underbrace{=}_{(\mathbf{y}_i)_i|\mu_0 \perp\!\!\!\perp} \int p(y_*(t), y_*(\mathbf{t}_*)|\mu_0(\mathbf{t}_+))\ p(\mu_0(\mathbf{t}_+)|(\mathbf{y}_i)_i)\ d\mu_0(\mathbf{t}_+). \end{aligned}$$

Or, par définition du modèle, on a:

$$p(y_*(t), y_*(\mathbf{t}_*)|\mu_0(\mathbf{t}_+)) = \mathcal{N}\left(\begin{bmatrix}\mu_0(t)\\ \mu_0(\mathbf{t}_*)\end{bmatrix}, \begin{pmatrix}\Psi_*(t,t) & \Psi_*(t,\mathbf{t}_*)\\ \Psi_*(\mathbf{t}_*,t) & \Psi_*(\mathbf{t}_*,\mathbf{t}_*)\end{pmatrix}\right).$$

De plus, pendant l'étape E de l'entrainement, on a vu que:

$$p(\mu_0(\mathbf{t}_+)|(\mathbf{y}_i)_i) = \mathcal{N}(\hat{m}_0(\mathbf{t}_+), \hat{K}(\mathbf{t}_+,\mathbf{t}_+)) = \mathcal{N}\left(\begin{bmatrix}\hat{m}_0(t)\\ \hat{m}_0(\mathbf{t}_*)\end{bmatrix}, \begin{pmatrix}\hat{K}(t,t) & \hat{K}(t,\mathbf{t}_*)\\ \hat{K}(\mathbf{t}_*,t) & \hat{K}(\mathbf{t}_*,\mathbf{t}_*)\end{pmatrix}\right).$$

Avec $\hat{m}_0$ et $\hat{K}$ définis comme dans l'étape E, on a finalement:

$$p(y_*(t), y_*(\mathbf{t}_*)|(\mathbf{y}_i)_i) = \mathcal{N}(\hat{m}_0(\mathbf{t}_+), \Gamma_*(\mathbf{t}_+,\mathbf{t}_+)) = \mathcal{N}\left(\begin{bmatrix}\hat{m}_0(t)\\ \hat{m}_0(\mathbf{t}_*)\end{bmatrix}, \begin{pmatrix}\Gamma_*(t,t) & \Gamma_*(t,\mathbf{t}_*)\\ \Gamma_*(\mathbf{t}_*,t) & \Gamma_*(\mathbf{t}_*,\mathbf{t}_*)\end{pmatrix}\right),$$

avec $\Gamma_* = \Psi_* + \hat{K}$.

La prédiction

En utilisant la formule habituelle de prédiction GP, on obtient la loi a posteriori suivante:

$$p(y_*(t)|y_*(t_*),(\mathbf{y}_i)_i) = \mathcal{N}(m_*, v_*),$$

avec :

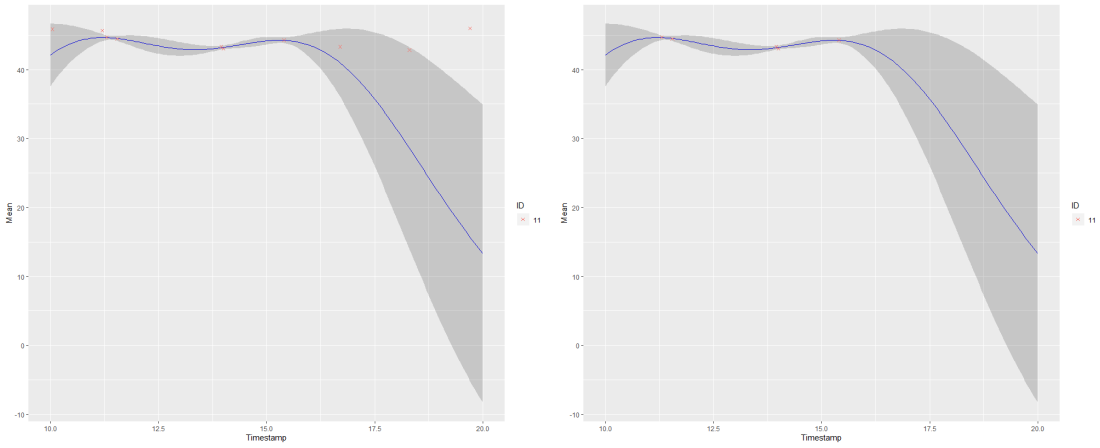
- $m_* = \hat{m}_0(t) + \Gamma_*(t, \mathbf{t}_*)\Gamma_*(\mathbf{t}_*, \mathbf{t}_*)^{-1}(y_*(t_*) - \hat{m}_0(\mathbf{t}_*))$,
- $v_* = \Gamma_*(t, t) - \Gamma_*(t, \mathbf{t}_*)\Gamma_*(\mathbf{t}_*, \mathbf{t}_*)^{-1}\Gamma_*(\mathbf{t}_*, t)$.

Visualisation graphique

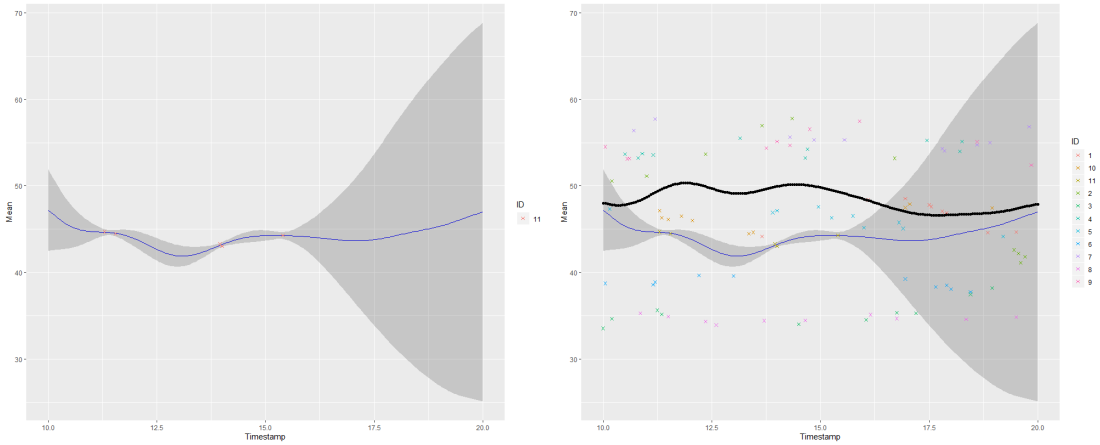
Une étude de simulation a été effectuée pour comparer une approche classique de modélisation par un unique GP avec la procédure décrite précédemment. Des données ont été simulées pour 11 individus, en tirant pour 10 temps d'observation et des hyper-paramètres spécifiques à chacun, des observations issues de GPs de même fonction moyenne (constante égale à 45) et de même forme de fonction de covariance (noyau exponentiel).

Le processus moyen μ_0 de notre procédure a été estimé à l'aide des 10 premiers individus et la prédiction s'effectue sur le 11ème individu, pour lequel on observe seulement 4 temps tirés aléatoirement. Dans les deux cas, la moyenne a priori (m_0) a été fixée à 0, afin de simuler l'absence d'information préalable. Sur les graphiques ci-dessous, on peut comparer les résultats de cette prediction sur une grille de 200 points de temps, pour chacune des deux approches.

Pour des raisons de concision, nous omettons ici les résultats plus détaillés concernant l'évaluation des erreurs, la comparaison avec d'autres algorithmes, les temps de calculs et la performance sur des cas pathologiques. Ces détails seront évoqués lors de la présentation orale.



La figure ci-dessus montre le résultat de la prédiction avec un modèle GP classique, qui a été entrainé seulement 4 observations représentées ici par les croix rouges de la figure de droite. Sur la figure de gauche, les croix rouges représentent la totalité des observations simulées. On voit que la prédiction reste fiable lorsque l'on reste proche des données observées, mais plongent rapidement vers la valeur a priori 0 dès que l'on s'en éloigne.



Les figures ci-dessus montrent le résultat de la prédiction avec le modèle décrit précédemment, qui a été entrainé sur la totalité des observations des 10 premiers individus, et sur les 4 mêmes observations du 11ème individu. La figure de gauche illustre cette prédiction dans le même contexte que pour un seul GP, mais qui cette fois reste bien meilleur, même lorsque l'on s'éloigne des points d'observation. Ceci s'explique à l'aide du graphique de droite, qui montre la totalité des observations utilisées (en couleur), et la valeur du processus moyen (points noirs) pour tous les temps de la grille de prédiction. En effet, lorsque l'on s'éloigne des temps observation, la prédiction se rapproche de la moyenne a priori, qui est cette fois informée par les autres individus.

Travaux en cours

Dans le cas où les courbes observées sont supposées issues de GPs avec des processus moyens différents, il est également possible de poser un modèle qui intègre une partie *clustering* au sein de la méthode décrite précédemment. Dans ce contexte, on suppose qu'il existe un nombre k de clusters dont chacun est caracterisé par un processus moyen μ_k spécifique. Il est alors possible d'écrire une procédure d'apprentissage qui ne sera plus un EM classique, mais un EM variationnel (une approche déjà utilisée dans le contexte GP dans Titsias (2009)) compte tenu des relations de dépendance entre les variables latentes. La prédiction quant à elle se fait de manière relativement naturelle par la suite. Ce travail est actuellement en cours d'implémentation.

Références

Bonilla, Edwin V, Kian M. Chai, and Christopher Williams. 2008. "Multi-Task Gaussian Process Prediction." In *Advances in Neural Information Processing Systems 20*, edited by J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, 153–60. Curran Associates, Inc.

Rasmussen, Carl Edward, and Christopher K. I. Williams. 2006. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. Cambridge, Mass: MIT Press.

Shi, J. Q., and B. Wang. 2008. "Curve Prediction and Clustering with Mixtures of Gaussian Process Functional Regression Models." *Statistics and Computing* 18 (3): 267–83. doi:10.1007/s11222-008-9055-1.

Shi, Jian Qing, and Yafeng Cheng. 2014. "Gaussian Process Function Data Analysis R Package 'GPFDA'," 33.

Titsias, Michalis K. 2009. "Variational Learning of Inducing Variables in Sparse Gaussian Processes." *AISTATS*, 8.

Yang, Jingjing, Hongxiao Zhu, Taeryon Choi, and Dennis D. Cox. 2016. "Smoothing and MeanCovariance Estimation of Functional Data with a Bayesian Hierarchical Model." *Bayesian Analysis* 11 (3): 649–70. doi:10.1214/15-BA967.

Yu, Kai, Volker Tresp, and Anton Schwaighofer. 2005. "Learning Gaussian Processes from Multiple Tasks." In *Proceedings of the 22Nd International Conference on Machine Learning*, 1012–9. ICML '05. Bonn, Germany: ACM. doi:10.1145/1102351.1102479.

Optimal quantization of the mean measure and application to clustering of measures

Clément Levrard [1] & Martin Royer [2] & Frédéric Chazal[3]

[1] *Université de Paris, 8 place Aurélie Nemours, 75013 Paris, clement.levrard@lpsm.paris*
[2,3] *INRIA Saclay, 1 rue Honoré D'Estienne d'Orves, 91120 Palaiseau, martin.royer@inria.fr, frederic.chazal@inria.fr*

Abstract. We consider the case where data consists of i.i.d. point sets, that are thought of as discrete measures. Our motivations are twofold. First we intend to approximate the mean of the measure-generating process (the mean measure, corresponding to the intensity function in a point process framework) with a finite measure supported by k points. To this aim we introduce two quantization algorithms and assess their optimality with respect to the expected distortion.

Second, we build a vectorization of the discrete measures data based on the previously obtained approximation of the mean measure. In a nutshell, the measures are mapped in a k-dimensional Euclidean space, and we prove that this mapping preserves the clusters structure, if any. An application of this scheme is given for clustering measure data that are generated by a mixture of persistence diagrams from different shapes. We show that our vectorization technique allows classical Euclidean clustering algorithms to recover the true clusters, with high probability. Further numerical insights are given by Royer (2019), for the closely related ATOL algorithm. This illustrates the relevance of our approach, especially in the case where data consists of a great number of discrete measure, each with many points.

Résumé. Nous nous intéressons au cas où les observations sont sous la forme d'ensembles de points, ou plus généralement de mesures discrètes et poursuivons deux objectifs. Premièrement nous voulons construire une approximation à support réduit de la mesure moyenne du processus générant ces mesures (correspondant à l'intensité dans un cadre de processus ponctuels). Pour cela nous introduisons deux algorithmes de type quantification et prouvons qu'ils fournissent des résultats optimaux au sens de la distorsion moyenne.

Dans un second temps, à partir de l'approximation de la mesure moyenne obtenue en première partie, nous construisons une vectorisation de nos mesures, c'est-à-dire une fonction qui représente chaque mesure par un point dans un espace Euclidien de dimension k, et prouvons que cette vectorisation préserve les clusters de mesures, si tant est qu'il y en ait. Dans le cas où les mesures sont des diagrammes de persistance générés par des formes de natures différentes, nous montrons qu'avec grande probabilité un clustering basé sur la vectorisation des diagrammes de persistance permet une classification non-supervisée quasi-exacte. Ce type de procédure est à la base de l'algorithme ATOL, Royer (2019), dont les performances numériques mettrons en lumière la pertinence d'une telle approche,

notamment dans le cas où l'on observe beaucoup de mesures discrètes comportant chacune beaucoup de points.

Mots-clés. Clustering, Analyse Topologique de Données

1 Quantification de la mesure moyenne

On suppose que l'on observe $X_1, \dots, X_n$, n mesures discrètes i.i.d. de support inclus dans la boule Euclidienne de rayon R, $\mathcal{B}(0, R)$, et on notera X une variable aléatoire de même loi que X_1. On peut penser par exemple au cas où on observe n diagrammes de persistance, ou encore la réalisation de n processus spatiaux. Notre premier objectif est d'approcher la **mesure moyenne**, $\mathbb{E}(X)$, définie, pour un borélien A, par

$$\mathbb{E}(X)(A) = \mathbb{E}(X(A)).$$

Si possible, nous voulons approcher $\mathbb{E}(X)$ par une mesure discrète de support réduit, c'est-à-dire par une mesure de la forme $P_{\mathbf{c}} = \sum_{j=1}^{k} \mu_j \delta_{c_j}$, où $\mathbf{c} = (c_1, \dots, c_k)$ est appelé un dictionnaire, formé de k mots c_j. La qualité d'une telle représentation est mesurée par la distance de Wasserstein (au carré) entre $\mathbb{E}(X)$ et $P_{\mathbf{c}}$, appelée aussi **distorsion** de $\mathbf{c}$:

$$R(\mathbf{c}) = W_2^2(\mathbb{E}(X), P_{\mathbf{c}}) = \mathbb{E}(X)(du) \bullet \min_{j=1,\dots,k} \|u - c_j\|^2 = \sum_{j=1}^{k} \mathbb{E}(X)(du) \bullet \|u - c_j\|^2 \mathbb{1}_{W_j(\mathbf{c})}(u),$$

où l'intégrale de f par rapport à la mesure Y est notée par $Y(du) \bullet f(u)$, et $(W_1(\mathbf{c}), \dots, W_k(\mathbf{c}))$ est une partition de Voronoi associée à $\mathbf{c}$ (avec attribution des frontières arbitraire). N'ayant accès qu'à $X_1, \dots, X_n$, nous proposons deux algorithmes essayant de minimiser $W_2^2(\bar{X}_n, P_{\mathbf{c}})$, où $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$. Le premier est une généralisation de l'algorithme de Lloyd (1982).

Algorithme batch

Input : $X_1, \dots, X_n$, et k.

Initialisation : Choix de $c_1^{(0)}, \dots, c_k^{(0)}$ au hasard dans le support de $\bar{X}_n$.

Actualisation : Tant que $\mathbf{c}^{(t+1)} \neq \mathbf{c}^{(t)}$:

- Pour $j = 1, \dots, k$,

$$c_j^{(t+1)} = \frac{1}{\bar{X}_n(W_j(\mathbf{c}^{(t)}))} \bar{X}_n(du) \bullet \left[u \mathbb{1}_{W_j(\mathbf{c}^{(t)})}(u) \right],$$

Sortie : $\mathbf{c}^{(T)}$, dictionnaire de la dernière itération.

Cet algorithme utilise toutes les mesures de l'échantillon à chaque passe, et peut donc s'avérer coûteux si le nombre de mesures est grand. Dans ce dernier cas on peut donner un équivalent mini-batch de l'Algorithme MacQueen (1967).

Algorithme mini-batch

INPUT : $X_1, \dots, X_n$ divisés en mini-batches $(B_1, \dots, B_T)$ de tailles $(n_1, \dots, n_T)$, k.

INITIALISATION : Choix de $c_1^{(0)}, \dots, c_k^{(0)}$ au hasard dans le support de $\bar{X}_n$.

ACTUALISATION : Pour $t = 0, \dots, T-1$:

– Pour $j = 1, \dots, k$,

$$c_j^{(t+1)} = c_j^{(t)} - \frac{1}{(t+1)\bar{X}_{B_{t+1}}(W_j(\mathbf{c}^{(t)}))} \bar{X}_{B_{t+1}}(du) \bullet \left[(c_j^{(t)} - u) \mathbb{1}_{W_j(\mathbf{c}^{(t)})}(u) \right],$$

SORTIE : $\mathbf{c}^{(T)}$, dictionnaire de la dernière itération.

Cet algorithme n'utilise chaque mesure de l'échantillon qu'une seule fois, et peut donc s'avérer beaucoup plus rapide que l'algorithme batch. Ces deux algorithmes fournissent des mesures à k points optimales, dès lors que la mesure moyenne $\mathbb{E}(X)$ satisfait une condition de marge au sens de Levrard (2015).

Théorème 1 *Si $\mathbb{E}(X)$ satisfait une condition de marge de rayon r_0 (au sens de Levrard (2015)), alors il existe une constante R_0 (dépendante de r_0) telle que, dès lors que l'initialisation $\mathbf{c}^{(0)}$ est R_0-proche d'un dictionnaire optimal,*

$$\mathbb{E}\left(R(\mathbf{c}_{Batch}^{(T)}) - R^* \right) \leq \frac{C}{n},$$

et $T \leq \log(n)$. De plus, on a aussi, pour l'algorithme mini-batch

$$\mathbb{E}\left(R(\mathbf{c}_{Minibatch}^{(T)}) - R^* \right) \leq \frac{C \log(n)}{n},$$

si les tailles de mini-batchs sont de l'ordre de $c \log(n)$.

Ce résultat est à comparer avec la vitesse optimale $1/n$ obtenue dans Levrard (2018) en quantification d'une distribution de probabilité sous conditions de marge. Il atteste donc de l'optimalité des algorithmes proposés (à un facteur $\log(n)$ près pour la version mini-batch), et dans le cas batch donne aussi une borne sur le nombre d'itérations nécessaire à l'obtention d'un dictionnaire optimal. A noter que ce résultat est aussi valable pour les algorithmes utilisés en quantification "classique" d'une mesure de probabilité (dans le cas où on observe un n-échantillon de points et non plus de mesure).

2 Clustering de mesures

2.1 Vectorisation basée sur la mesure moyenne

Comme annoncé, notre objectif secondaire est de trouver une représentation des mesures de départ $X_1, \dots, X_n$ via des points dans un espace Euclidien qui conserve une éventuelle structure en clusters. Le schéma de vectorisation adopté est le suivant: pour un noyau Ψ, un facteur d'échelle $\sigma > 0$ et un dictionnaire $\mathbf{c}$, la vectorisation d'une mesure finie X est donnée par

$$v_{\mathbf{c},\sigma}(X) = (X(du) \bullet \psi(\|u - c_1\|/\sigma), \dots, X(du) \bullet \psi(\|u - c_k\|/\sigma)),$$

l'idée étant de représenter une mesure X par la masse qu'elle donne à l'échelle σ autour de chacun des centres $c_1, \dots, c_k$. Deux noyaux présentent un intérêt particulier, $\psi_0 : x \mapsto (1 - ((x-1) \vee 0)) \vee 0$, qui présente le plus d'avantages théoriques, et $\psi_{AT} : x \mapsto \exp(-x)$, utilisé par l'algorithme ATOL Royer (2019) dans un contexte d'analyse topologique de données. Bien sûr, le choix du dictionnaire de vectorisation peut s'effectuer en quantifiant la mesure moyenne, comme expliqué dans la partie précédente. Bien que l'on puisse donner des garanties générales sur la vectorisation qui en résulte, nous préférons plutôt exposer une application de ces résultats dans un contexte d'analyse topologique de données/classification de formes.

2.2 Application au clustering de formes

Un cas précis où les données peuvent se présenter sous la forme de mesures discrètes est celui des diagrammes de persistance. Chaque mesure X_i est générée comme suit: on tire un label $Z_i \in [\![1, L]\!]$, correspondant à une forme sous-jacente $S^{(Z_i)}$. Sachant que $Z_i = \ell$, on tire N_ℓ points sur la forme $S^{(\ell)}$ et construit le diagramme de persistance X_i associé à la distance à ces N_ℓ points, comme décrit dans Cohen-Steiner (2007). Si le nombre de points N_ℓ est suffisamment grand, X_i sera proche du diagramme de persistance de la forme ℓ, $D^{(\ell)}$, vu comme une mesure discrète sur $\mathbb{R}^2$.

Les formes $S^{(1)}, \dots, S^{(L)}$ sont discriminables à partir de leurs diagrammes de persistance si et seulement si, pour tout $1 \leq \ell_1 < \ell_2 \leq L$ il existe $m_{\ell_1,\ell_2} \in \mathbb{R}^2$ tel que

$$D^{(\ell_1)}(\{m_{\ell_1,\ell_2}\}) \neq D^{(\ell_2)}(\{m_{\ell_1,\ell_2}\}).$$

Dans une telle situation, si k est assez grand, on peut montrer que le dictionnaire obtenu dans la première partie mène à une vectorisation permettant un clustering quasi-parfait des mesures données.

Théorème 2 *Si les formes $S^{(1)}, \dots, S^{(L)}$ sont discriminables à partir de leurs diagrammes de persistance, alors, pour k assez grand, si $\mathbb{E}(X)$ satisfait une condition de marge au sens*

de Levrard (2015), on a, avec grosse probabilité,

$$\begin{array}{rcl} Z_{i_1} = Z_{i_2} & \Rightarrow & \|v_{i_1} - v_{i_2}\|_\infty \leq \frac{1}{4}, \\ Z_{i_1} \neq Z_{i_2} & \Rightarrow & \|v_{i_1} - v_{i_2}\|_\infty \geq \frac{1}{2}, \end{array}$$

où les v_i sont les vectorisations obtenues avec les noyaux Ψ_0 ou Ψ_{AT}, avec un facteur d'échelle compris dans une certaine bande $[\sigma_0, 2\sigma_0]$.

Ce résultat montre qu'un algorithme standard de clustering, type Single Linkage, appliqué à la vectorisation des diagrammes de persistance permet un clustering exact sur l'évènement de probabilité en question. Cette garantie théorique de performance raisonnable en clustering sera illustrée par plusieurs exemples numériques, où la vectorisation donnée par ATOL donnera des résultats satisfaisants, comparé à des techniques de vectorisation plus élaborées et donc plus coûteuses.

Bibliographie

Royer, M., Chazal, F., Levrard, C., Ike, Y., and Umeda, Y. (2019). ATOL: Measure Vectorisation for Automatic Topologically-Oriented Learning, *arXiv e-print*.

Lloyd, S. P. (1982). Least squares quantization in PCM, *IEEE Transactions on Information Theory*, 28, pp.129-137.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. I, pp. 281-297.

Levrard, C. (2015). Nonasymptotic bounds for vector quantization in Hilbert spaces, *The Annals of Statistics*, 43, pp. 592-619.

Levrard, C. (2018). Quantization/Clustering: when and why does k-means work, *Journal de la Société Française de Statistiques*, 159.

Cohen-Steiner, D., Edelsbrunner, H., and Harer, J. (2007). Stability of persistence diagrams, *Discrete Computational Geometry*, 37, pp. 103-120.

Modèles linéaires généralisés hiérarchiques pour l'analyse de la diversité du riz

Lorena León [1,2], Jean Peyhardi [2], Catherine Trottier [2,3]

[1] *CIRAD, UMR AGAP, Montpellier, France.*
yinneth-lorena.leon_velasco@cirad.fr
[2] *IMAG, Univ Montpellier, CNRS, Montpellier, France.*
jean.peyhardi@umontpellier.fr
[3] *Univ Paul Valéry Montpellier 3, Montpellier, France.*
catherine.trottier@umontpellier.fr

Résumé. Les traits de panicule sont parmi les caractéristiques les plus représentatives de la diversité du riz ; leur architecture est pertinente pour la classification biologique des plantes ainsi que pour l'amélioration du riz cultivé. Sur la base des modèles linéaires généralisés conditionnels partitionnés (PCGLMs) proposés récemment par Peyhardi et al. (2016), nous visons à expliquer la classification taxonomique du riz à partir d'un ensemble de caractéristiques phénotypiques. Nous avons considéré deux arbres de partition dont les nœuds regroupent différentes catégories taxonomiques. Ce travail illustre comment les PCGLMs permettent d'extraire des informations plus fines que les modèles classiques, comme le modèle logit mutinomial.

Mots-clés. Variable catégorielle structurée hiérarchiquement, PCGLM, fonction de lien, traits de panicule de riz, diversité du riz.

Abstract. Panicle traits are among the most representative features of rice diversity; their architecture is relevant for the biological classification of plants as well as for the improvement of cultivated rice. Based on the partitioned conditional generalized linear models (PCGLMs) recently proposed by Peyhardi et al. (2016), we aim to explain the taxonomic classification of rice given a collection of phenotypic features. We considered two partition trees whose nodes group different taxonomic categories. This work illustrates how the PCGLMs allow us to extract finer information than the conventional models such as the mutinomial logit model.

Keywords. Hierarchically-structured categorical variable, PCGLM, link function, rice panicle traits rice diversity.

1 Contexte et données

La génétique quantitative repose sur des modèles à effets aléatoires dont la vue fonctionnelle est *phénotype* = f(*génotype*) (Gianola, 2007). Dans ce cadre, il n'y a pas de restriction sur la modélisation du génotype et des structures complexes d'apparentement peuvent être modélisées. Par contre, la modélisation des traits phénotypiques (variables réponses du modèle de régression) est très fortement contrainte et le plus souvent, un unique trait ou un vecteur de traits quantitatifs est pris en compte. Nous développons ici une nouvelle approche de modélisation statistique pour l'analyse de la diversité des plantes qui inverse la vue fonctionnelle du modèle de régression qui sera donc

$$génotype = f(phénotype).$$

L'idée est de pouvoir incorporer sans contrainte des traits phénotypiques hétérogènes (variables qualitatives, quantitatives, ordinales et de comptage) tout en ayant de larges possibilités pour modéliser des familles de génotypes sous forme de hiérarchies de catégories (par exemple, des espèces se subdivisant en groupes de variétés, elles-mêmes ayant différentes origines géographiques).

La base de données phénotypiques de riz compte 752 plantes. Chaque plante est classifiée selon son origine géographique, espèce, sous-espèce et sous population. Pour chaque continent (asie, afrique) une espèce cultivée et une espèce sauvage sont considérées : sativa (asie-cultivée), rufipogon (asie-sauvage), glaberrina (afrique-cultivée) et barthii (afrique-sauvage). Cette base de données comprend comme variables explicatives : longueur du rachis, longueur total, nombre de grains, nombre maximum d'ordre de ramifications, nombre de nœuds et nombre de nœuds dans le rachis.

2 Le modèle

Notre nouvelle approche repose sur les modèles linéaires généralisés hiérarchiques récemment introduits par Peyhardi et al. (2016). Les PCGLMs (Partitioned Conditional Generalized Linear Models) sont des modèles de régression pour réponses catégorielles. Ils sont fondés sur :

- Un arbre de partition qui spécifie la structure hiérarchique des catégories de la réponse.
- Un modèle sur chaque nœud non terminal reliant l'espérance de la réponse et les covariables à travers l'équation :

$$r(\boldsymbol{\pi}) = F(Z\beta) \tag{1}$$

avec $\boldsymbol{\pi} = (\pi_1, ..., \pi_J)$: les probabilités de chaque catégorie. D'après Peyhardi et al. (2015), ce modèle est caractérisé par un triplet (r, F, Z) où :

– r est un rapport des probabilités parmi les 4 suivants,

	Référence	Cumulatif	Séquentiel	Adjacent
$r_j(\pi)$	$\frac{\pi_j}{\pi_j + \pi_J}$	$\pi_1 + ... + \pi_j$	$\frac{\pi_j}{\pi_j + ... + \pi_J}$	$\frac{\pi_j}{\pi_j + \pi_{j+1}}$

– F est une fonction de répartition, par exemple, logistique, normale, Gumbel ou Student.

– Z est une matrice de design, par exemple, complet (sans contrainte) :

$$Z_c = \begin{pmatrix} 1 & & x^t & & \\ & \ddots & & \ddots & \\ & & 1 & & x^t \end{pmatrix}$$

et proportionnel (pente commune) :

$$Z_p = \begin{pmatrix} 1 & & x^t \\ & \ddots & \vdots \\ & & 1 \quad x^t \end{pmatrix}$$

À titre d'exemple, le modèle classique logit multinomial, décrit par les équations :

$$P(Y = j) = \frac{exp(\alpha_j + x^T \delta_j)}{1 + \sum_{k=1}^{J-1} exp(\alpha_k + x^T \delta_k)}, j = 1, \ldots, J-1, \tag{2}$$

est caractérisé par le triplet *(référence, logistique, complet)*. En utilisant la généricité de la spécification (r, F, Z) des GLM pour les réponses catégorielles, il est possible à chaque nœud de spécifier différentes fonctions de lien et d'utiliser différentes variables explicatives. En partant de l'hypothèse que les paramètres sont différents d'un nœud à l'autre, le modèle est facilement estimable puisque la log-vraisemblance du PCGLM est alors égale à la somme des log-vraisemblances de chaque nœud. Ceci donne une flexibilité que l'on ne retrouve pas dans d'autres modèles.

Cependant, cette flexibilité a un coût puisque le nombre de modelés à visiter est très vaste. Nous avons adopté l'heuristique suivant : spécifier l'arbre de partitionnement, puis à chaque nœud séparément choisir d'abord r puis F et enfin sélectionner les variables explicatives.

3 Application

Deux arbres ont été étudiés pour la classification taxonomique d'***Oryza***. Dans le premier arbre (cf. Fig. 1), les données sont d'abord divisées selon le caractère de domestication

(sauvage ou domestique) et ensuite selon l'origine géographique (afrique ou asie). Tandis que dans le deuxième arbre (non montré ici), les divisions sont faites dans l'ordre inverse. L'objectif pour chaque nœud est de trouver le meilleur modèle (r, F, Z) tout en

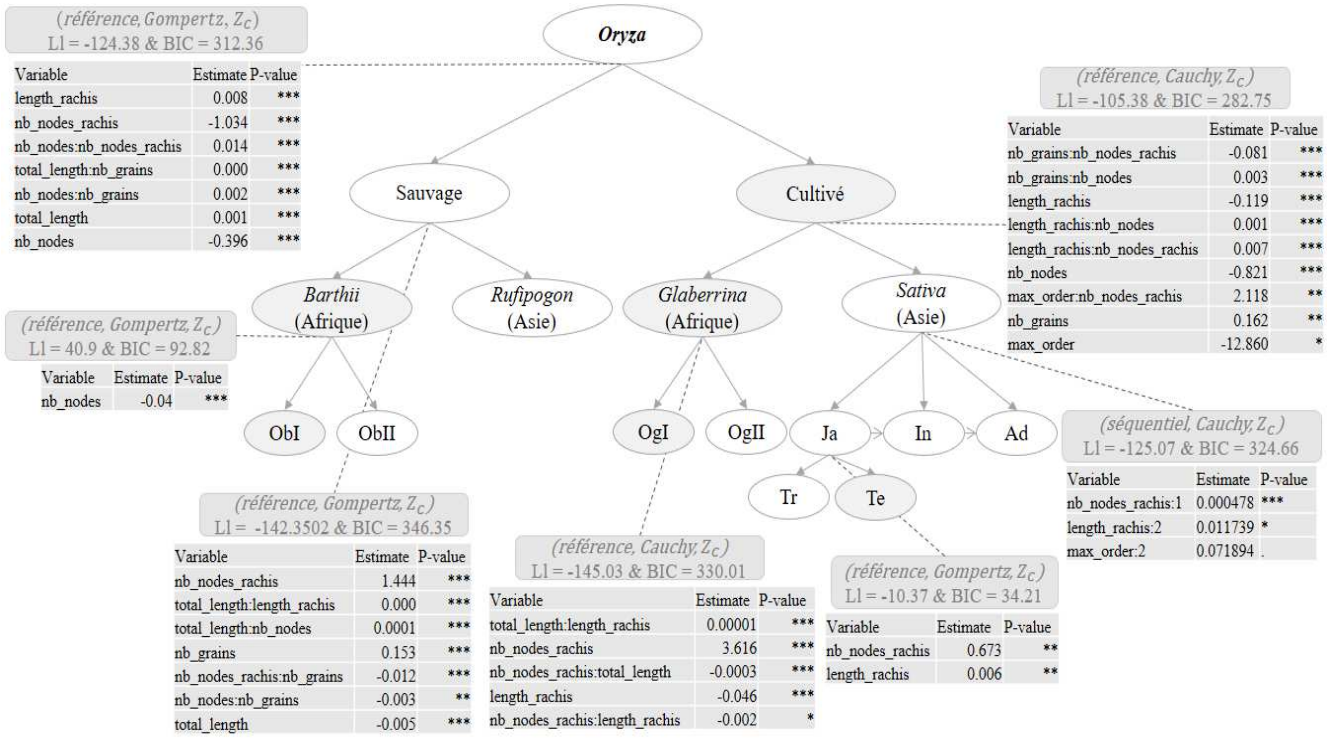


Figure 1: PCGLM pour les données sur la diversité du riz dans lequel les sous-espèces sont Ob : Obar ; Og : Ogla ; Ja : Japonica ; In : Indica ; Ad : Admix (mélange d'Aromatique avec Te) et les variétés Tr : Tropical Japonica et Te : Tempéré Japonica. Les nœuds grises sont la catégorie de référence utilisée pour l'estimation du modèle.

sélectionnant les variables phénotypiques pertinentes. Lorsqu'un nœud possède plus de deux enfants, le choix du rapport revêt un caractère important pour traduire notamment une notion d'ordre entre les catégories. Pour le cas particulier des trois sous-espèces de *Sativa*, Huang et al. (2012) ont présenté un scénario démographique dans lequel *Japonica* a d'abord été domestiquée à partir de l'espèce sauvage *Or-IIIa*, tandis que *Indica* a ensuite été développée à partir de *Or-I* avec l'adoption de nombreux allèles de domestication de *Japonica*. Dans le cas du groupe *Aromatique*, on considère qu'il a été domestiqué peu après que *Japonica* et *Indica* aient déjà existé. Pour prendre en compte cet ordre chronologique, le rapport séquentiel a été proposé pour ce nœud.

Lorsque l'on compare les deux arbres à l'aide du critère BIC, l'arbre de la figure 1 ($BIC = 1757.83$) est préféré au deuxième avec l'ordre inverse ($BIC = 1772.9$). Bien que l'écart ne soit pas si marqué, ces résultats sont concordants avec les études publiées qui

montrent que les différences entre les traits phénotypiques sont plus prononcés entre les groupes sauvage/cultivé qu'entre les groupes asie/afrique. Cependant, des méthodologies telles que des analyses génomiques des populations et les techniques de séquençage des données, démontrent une forte différenciation génétique entre le riz asiatique et africain.

Le nombre de nœuds dans la panicule est l'une des variables les plus discrimantes dans chaque partition de l'arbre. Elle est fortement corrélée à la quantité de grains, c'est-à-dire à la productivité, comme le montre le modèle d'interactions simples (non présenté ici) dans lequel cette variable s'avère être la plus importante pour différencier les types de riz cultivé et sauvage. Au sein des espèces cultivées, les panicules d'origine africaine ont généralement des rachis plus longs, alors que dans ceux d'origine asiatique, on observe une plus grande quantité de grains. Cette corrélation inverse entre les traits de longueur et de comptage a été bien documentée pour le riz (Crowell et al., 2016). Plus particulièrement pour l'espèce *Sativa*, les modèles avec les six permutations des trois sous-populations ont été estimés. Nous avons constaté que l'ordre : 1. *Japonica*, 2. *Indica* et 3. *Admix*, était le plus approprié, en utilisant le modèle *(séquentiel, cauchit, complet)*. Le modèle avec F : Cauchy a montré le meilleur ajustement en comparaison à des autres fonctions de répartition telles que normale, logistique et log-log complémentaire. La distribution de Cauchy a des queues plus lourdes, permettant ainsi des valeurs plus extrêmes que les autres distributions et donc un ajustement alternatif intéressant à considérer.

4 Conclusions et perspectives

Nous avons identifié les effets des traits phénotypiques pour la description et la différenciation des catégories taxonomiques. Étant donné le faible écart entre les BIC de chaque arbre, il n'est pas certain qu'une des partitions considérées soit meilleure que l'autre. Il est constaté qu'une des principales difficultés d'analyser la diversité végétale est que, même si les caractères phénotypiques sont généralement spécifiques à chaque catégorie taxonomique, ils varient également dans différentes conditions environnementales. Cette limitation est évidente dans d'autres études qui ne parviennent à décrire que des différences relativement subtiles pour chacun des niveaux taxonomiques. Dans la plupart des applications, l'arbre de partition n'est pas connu *a priori* et l'objectif serait de construire une méthode statistique pour la recherche de cet arbre. Il est évident que l'ensemble des partitions à tester est très large, compte tenu de toutes les combinatoires possibles pour l'ensemble des données. Pour atténuer le problème de la surcharge de calcul, une méthode heuristique doit être proposée pour traverser cet espace de manière optimale sans qu'il soit nécessaire d'estimer tous les modèles.

Bibliographie

Peyhardi J., Trottier C., et Guédon Y.. (2016), Partitioned Conditional Generalized Linear Models for Categorical Responses, *Statistical Modelling: An International Journal*, 16 (4): 297-321.
Peyhardi J., Trottier C., et Guédon Y.. (2015), A new specification of generalized linear models for categorical responses, *Biometrika*, 102, 889–906.
Gianola D. (2007), Inferences from mixed models in quantitative genetics. *Handbook of Statistical Genetics*, Edition 3, Vol. 1. John Wiley & Sons, Chichester, West Sussex, England, p. 678–717.
Choi, Young J., Platts A., Fuller D., Hsing Y., Wing R., et Purugganan M. (2017). The Rice Paradox: Multiple Origins but Single Domestication in Asian Rice. *Molecular Biology and Evolution.*
Crowell S., Korniliev P., Falcão A., Ismail A., Gregorio G., Mezey J., McCouch S. (2016): Genome-wide association and high-resolution phenotyping link Oryza sativa panicle traits to numerous trait-specific QTL clusters. *Nature communications* p. 10527.
Huang X., Kurata N., Wei X., Wang Z., Wang A., et al. (2012). A map of rice genome variation reveals the origin of cultivated rice. *Nature* 490:497–501

Sparse group variable selection to leverage pleiotropic association in GWAS data

Benoit Liquet [1,2] & Pierre Emmanelle Sugier [1,2] & Matthew Sutton[2],& Anthony Pettitt [2]& Kerrie Mengersen[2] & Thérèse Truong[3]

[1] *LMAP- UMR CNRS 5142, E2S-UPPA, France, benoit.liquet@univ-pau.fr*
[2] *ACEMS, QUT, Australia*
[3] *INSERM U1018, Paris-Sud, France*

Abstract.
Results from genome-wide association studies (GWAS) suggest that complex diseases are often affected by many variants with small effects, known as polygenicity. Bayesian methods provide attractive tools for identifying signal in data where the effects are small but clustered. For example, by incorporating biological pathway membership in the prior they are able to integrate the ideas of gene set enrichment to identify groups of biologically significant genetic variants. Accumulating evidence suggests that genetic variants may affect multiple different complex diseases, a phenomenon known as pleiotropy.

In this work we propose frequentist and Bayesian statistical methods to leverage pleiotropic effects and incorporate prior pathway knowledge to increase statistical power and identify important risk variants. We offer novel feature selection methods for the group variable selection in multi-task regression problem. We develop methods using both penalised likelihood methods and Bayesian spike and slab priors to induce structured sparsity at a pathway, gene or SNP level. We implement Gibbs sampling algorithms for the Bayesian analysis and an alternating direction method of multipliers (ADMM) algorithm for the penalised regression methods. The performance of the proposed approaches are compared to state-of-the-art variable selection strategies on simulated data sets.

The developed statistical approaches is applied for enriching our insights about the genetic mechanisms of thyroid and breast cancer types. The analysed data come from case-control studies including 6631 SNPs from 625 genes from 10 non-overlapping gene pathways. The thyroid cancer data set includes 482 cases and 463 controls. The breast cancer data set includes 1172 cases and 1125 controls.

Keywords. ADMM; Lasso Penalty, Sparsity; Spike and Slab

1 Introduction

Due to the arising of high dimension data, a large number of analyses have been performed in genetic epidemiology. While most of the current results deal with genetics markers associated to a single phenotype, more complex mechanisms still need further studies. Especially pleiotropy defines cases where a single genetic marker have an influence on several disease. In theses studies, genomic data come often from different data set, one per disease. A very particular structure in the data can then be found. From one side, similarly to genetic analysis in general, groups of variables can be given as an a priori information such as genes or pathways. From the other side, the different data sets gives observation sets as an a priori information. Methods dealing with those two type of information should be used. Given the complexity of the structure, most of existing approaches are based on the construction of a test giving p-values based on a null hypothesis rather than building a model that can be interpreted. In this article, an extension of the Partial Least Square is proposed using a particular Lasso penalization.

Recent advances in genetic sequencing techniques have permitted the acquisition of a large number of data which lead to new results in genetic epidemiology. For instance, genome-wide association studies (GWAS) have identified numerous genetic markers linked to multiple phenotypes, suggesting the existence of pleiotropy that occurs when a single variant or gene can influence several phenotype traits (Solovieff et al. 2013). Highlighting pleiotropy provides opportunities for understanding the shared genetic underpinnings among associated diseases. As the genetic effects for most complex traits are small, several methods were proposed to combine results across studies of different phenotypes in order to improve the power of detecting cross-phenotype or pleiotropic associations.

We propose to focus on the case where several GWAS on different traits are available from different independent sources. One way to analyze pleiotropy is to analyze each trait separately in each independent datasets. In order to gain statistical power to detect pleiotropy, we propose models that can combine the different traits into a meta-analysis taking into account that effect in opposite directions may exist across the different traits. In this work we develop frequentist and Bayesian statistical methods to leverage pleiotropic effects and incorporate prior pathway knowledge to increase statistical power and identify important risk variants. We focus on a gene-based approach that combines the association signals from the single nucleotide polymorphisms (SNP) into a signal at the gene level or at the pathway-level.

2 Modelling sparse and grouped associations in many independent datasets

Suppose we have data from K independent datasets, $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2 \cup \cdots \cup \mathcal{D}_K$, where $\mathcal{D}_k = (\{y_{1k}, x_{1k}\}, \ldots, \{y_{n_k k}, x_{n_k k}\})$ and dataset contain $n_1, \ldots, n_K$ samples respectively. The response variable $y_{ik} \in \{0, 1\}$ is the binary phenotype of the ith individual of the kth study and $x_{ik} \in R^p$ is the vector with corresponding p variables of the ith individual of the kth study. These data are assumed to come from a logistic regression model

$$y_{ik} \sim \text{Bernoulli}(g^{-1}(\nu_{ik})) \quad \text{with } \nu_{ik} = x_{ik}^T \beta_{\cdot k}$$

for $k = 1, \ldots, K$, when $g(\cdot)$ is the logit function and $\beta_{\cdot k} \in R^p$ denotes the regression coefficients for the kth study. To simplify further notation, let $\beta_{j\cdot} \in R^K$, $j = 1, \ldots, p$ denote the vector of the K regression coefficients corresponding to the jth SNP over the K datasets. We let β_{jk} denote the regression coefficient for the jth SNP of the kth study.

We assume that the set of SNPs can be partitioned into G groups (gene or pathway level). We define π_g, $g = 1, \ldots, G$ to denote the set of SNPs contained in the gth group and $p_g = \text{card}(\pi_g)$. Finally, we denote the matrix of all regression coefficients as $\mathbf{B} = (\beta_{\cdot 1}, \ldots, \beta_{\cdot K})$.

2.1 Frequentist approach

The negative log likelihood for the combined datasets his

$$-\log(p(\mathcal{D}|\mathbf{B})) = \sum_{k=1}^{K} \sum_{i=1}^{n_k} (y_i x_{ik}^T \beta_{\cdot k} - \log(1 + e^{x_{ik}^T \beta_{\cdot k}}).$$

Using this form of the likelihood, we propose the penalised likelihood estimate

$$\hat{\mathbf{B}} = \underset{\mathbf{B} \in \mathbb{R}^{p \times K}}{\text{argmin}} \left\{ -\log(p(\mathcal{D}|\mathbf{B})) + \lambda_1 \|\mathbf{B}\|_{G_{2,1}} + \lambda_2 \|\mathbf{B}\|_{l_{2,1}} \right\} \tag{1}$$

$$\text{where } \|\mathbf{B}\|_{G_{2,1}} = \sum_{g=1}^{G} \sqrt{n_g} \sqrt{\sum_{i \in \pi_g} \sum_{k=1}^{K} \beta_{ik}^2}$$

$$\text{and } \|\mathbf{B}\|_{l_{2,1}} = \sum_{i=1}^{p} \|\beta_{i\cdot}\|_2 = \sum_{i=1}^{p} \sqrt{\sum_{k=1}^{K} \beta_{ik}^2}$$

where λ_1 and λ_2 are regularisation parameters weighting a $G_{2,1}$-norm penalty $\|\mathbf{B}\|_{G_{2,1}}$ and $l_{2,1}$-norm penalty $\|\mathbf{B}\|_{l_{2,1}}$ respectively. The $G_{2,1}$-norm (Wang et al., 2012) fixes the group structure across studies and encourage sparsity at group level. As important groups

may contain irrelevant SNPs we desire a method which is able to select variables within a group. This is handled by the $l_{2,1}$-norm which allows for more structured sparsity. We propose to fit this model (1) using the alternating direction method of multipliers (ADMM) algorithm (Boyd et al., 2011).

2.2 Bayesian approach

2.2.1 Bayesian group selection

We begin by introducing our model for the analysis by accounting for grouping structure across the studies. Let the vector of binary variables $\gamma = (\gamma_1, \ldots, \gamma_G)^T$ indicate the association status for SNPs where $\gamma_g = 1$ indicates that all p_g SNPs in group g are associated to all K traits for $g = 1, \ldots, G$. We propose the following spike and slab prior for the gth SNP $\mathbf{B}_g \in \mathbb{R}^{p_g \times K}$,

$$\begin{aligned}
\mathbf{B}_g &\sim (1-\gamma_g)\mathcal{MN}_{p_g,K}(\mathbf{0}_{p_g,K}, \Sigma, \tau_g^2 \mathbb{I}_{p_g}) + \gamma_g \delta_0(\mathbf{B}_g) \\
\Sigma &\sim IW(d+K-1, Q), \\
\tau_g^2 &\sim \text{Gamma}\left(\frac{p_g+1}{2}, \frac{\lambda_g^2}{2}\right), \\
\gamma_g &\sim \text{Bernoulli}(\alpha_0) \\
\alpha_0 &\sim \text{Beta}(a, b)
\end{aligned}$$

for $g = 1, \ldots, G$, where $\delta_0(\mathbf{B}_g)$ denotes a point mass at $\mathbf{0} \in \mathbb{R}^{p_g \times K}$. This prior corresponds to the multivariate group LASSO with spike and slab prior proposed in Liquet et al. (2017) where each group contains the coefficients for the SNPs corresponding to group g across all K studies.

2.2.2 Bayesian within-group selection

To enforce both group and within-group sparsity we introduce parameter to address the sparsity at two levels. To model the within group sparsity, define the parameters $\tau_i^{(g)}$ where $\tau_i^{(g)} \geq 0$ for $i = 1, \ldots, p_g$ and $g = 1, \ldots, G$ and define $D_{\tau_g} = diag(\tau_1^{(g)}, \ldots, \tau_{p_g}^{(g)})$. To model the group sparsity define the matrices $\tilde{\mathbf{B}}_g \in \mathbb{R}^{p_g \times K}$ for $g = 1, \ldots, G$. We reparameterise the regression coefficients as

$$\mathbf{B}_g = D_{\tau_g} \tilde{\mathbf{B}}_g.$$

Define the within-group sparse prior as

$$\begin{aligned}
\tilde{\mathbf{B}}_g &\sim (1-\alpha_0)\mathcal{MN}_{p_g,K}(\mathbf{0}_{p_g,K}, \Sigma \otimes \mathbb{I}_{p_g}) + \alpha_0\delta_0(\mathbf{B}_g) \\
\tau_i^{(g)} &\sim (1-\alpha_1)\mathcal{N}^+\left(0, s^2\right) + \alpha_1\delta_0(\tau_i^{(g)}), \\
\Sigma &\sim IW(d, Q), \\
\alpha_0 &\sim \text{Beta}(a_1, a_2) \\
\alpha_1 &\sim \text{Beta}(c_1, c_2) \\
s^2 &\sim \text{InvGamma}(1, t)
\end{aligned}$$

for $g = 1, \ldots, G$ where $\mathcal{N}^+(0, s^2)$ denotes a normal $N(0, s^2)$ distribution truncated below at 0.

Acknowledgment

The authors acknowledge "La Ligue contre le Cancer" for supporting this work on pleiotropy.

Bibliographie

Boyd, S., Parikh, N., Chu, E., Peleato, B. and Eckstein, J. (2011). Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Found. Trends Mach. Learn.*,**3**, 11 – 22.

Liquet, B., Mengersen, K., Pettitt, A.N. and Sutton, M. (2017). Bayesian variables election regression of multivariate responses for group data. *Bayesian Analysis*, **12**, 1039 – 1067.

Solovieff, N., Cotsapas, C., Lee, P.H., Purcell, S.M. and Smoller, J.W. (2013). Pleiotropy in complex traits: challenges and strategie. *Nature Reviews Genetics*, **14(7)**, 483 – 495.

Wang, H., Nie, F., Huang, H., Kim, S., Nho, K., Risacher, S. L., Saykin, A.J. and Shen, L. (2012). Identifying quantitative trait loci via group-sparse multitask regression and feature selection: An imaging genetics study of the ADNI cohort. *Bioinformatics*, **28**, 229 – 237.

Optimisation des parcours patients pour lutter contre l'errance de diagnostic des patients atteints de maladies rares

Frédéric Logé[1,2*], Rémi Besson[3*], Stéphanie Allassonnière[3]

[1] *Campus Innovation Paris, R&D Air Liquide, 78350 Les Loges-en-Josas, France*
[2] *Centre de Mathématiques Appliquées, Polytechnique, 91120 Palaiseau, France*
[3] *Centre de Recherche des Cordeliers, Université de Paris, INSERM, Sorbonne Université, 75006 Paris, France*
frederic.logemunerel@gmail.com ; remi.besson@polytechnique.edu

*Les auteurs ont contribué équitablement à la production de ce travail.

Résumé. Un patient atteint d'une maladie rare en France doit en moyenne attendre deux ans avant d'être diagnostiqué. Cette errance médicale est fortement préjudiciable tant pour le système de santé que pour les patients dont la pathologie peut s'aggraver. Il existe pourtant un réseau performant de centres de référence maladies rares (CRMR), mais les patients ne sont orientés que trop tardivement vers ces structures. Nous considérons une modélisation probabiliste du parcours patient afin de créer un simulateur permettant d'entraîner un système d'alerte détectant les patients en errance et les orientant vers un CRMR tout en considérant les potentiels surcoûts associés à ces décisions. Les premiers résultats obtenus sur données simulées apparaissent prometteurs. Un important travail de mise en relation des données expertes disponibles avec les données de parcours patients reste à faire ainsi que des ajustements sur la modélisation proposée.
Mots-clés. Maladies rares, Parcours patient, Simulation, Graphes, Optimisation

Abstract. A patient suffering from a rare disease in France has to wait an average of two years before being diagnosed. This medical wandering is highly detrimental both for the health system and for patients whose pathology may worsen. There exists an efficient network of Centres of Reference for Rare Diseases (CRMR), but patients are often referred to these structures too late. We are considering a probabilistic modelling of the patient pathway in order to create a simulator that will allow us to create an alert system that detects wandering patients and refers them to a CRMR while considering the potential additional costs associated with these decisions. The first results obtained on simulated data appear promising. An important work of linking the available expert data with the data of patient journey remains to be done as well as adjustments on the proposed modeling.
Keywords. Rare disease, Patient patwhay, Simulation, Graphical models, Optimization

1 Introduction

1.1 Contexte

Le récent rapport Erradiag [1] de l'Alliance Maladies Rares a montré qu'en France un patient atteint d'une maladie rare [2] devra attendre en moyenne 2 ans entre l'apparition des premiers symptômes et le diagnostic de sa pathologie. Pour plus d'un quart d'entre eux cette durée sera de plus de 5 ans. Ce laps de temps, appelé errance de diagnostic, est un fléau pour le système de santé de par les surcoûts économiques (multiplication d'examens et traitements médicaux inutiles) et humain engendrés (aggravation de la pathologie du fait de l'inadéquation de la prise en charge).

Les patients atteints d'une maladie rare sont particulièrement affectés par l'errance de diagnostic du fait du caractère multifactoriel de ces pathologies : une maladie rare touche bien souvent plusieurs organes différents

1. L'errance de diagnostic : Erradiag résultats de l'enquête sur l'errance de diagnostic. 2016. 844 patients interrogés.
2. Maladie rare : maladie dont la prévalence est inférieure à 1/2000 (seuil Européen).

et le diagnostic nécessite donc une approche pluridisciplinaire. Par ailleurs, un praticien n'exerçant pas en centre expert n'observera probablement jamais la plupart des maladies rares au cours de sa carrière. Cependant, bien que les maladies rares soient par définition peu fréquentes, les malades sont eux nombreux puisque l'on estime à 3 millions le nombre de personnes concernées en France et entre 263 et 446 millions dans le monde [Nguengang and al, 2019]. Cela est dû au très grands nombre de maladies rares existantes : pas moins de 9000 sont ainsi recensées dans OrphaNet [3], le portail dédié aux maladies rares.

La France est pionnière dans la lutte contre les maladies rares, lançant dès 2005 le premier Plan National Maladie Rare avec notamment la structuration de centres de référence maladies rares (CRMR). Ces centres pluridisciplinaires ont pour but une meilleure prise en charge des patients atteints de maladies rares. Le rapport Erradiag rappelle que, malgré les progrès enregistrés suite à la création de ces centres, un nombre encore trop important de malades sont orientés très tardivement vers une structure hospitalière. Un quart des patients attendent ainsi plus de 4 ans après les premiers symptômes avant que la recherche du diagnostic ne soit enfin initiée. Une meilleure orientation des patients dans le système de santé est donc nécessaire.

Dans le même temps les données médicales, sous forme de base experte comme Orphanet ou de données de parcours patient (consultations/remboursements) avec l'apparition du dossier médical électronique, se sont structurées et sont désormais plus accessibles pour la recherche rendant possible l'objectif d'amélioration des parcours patient.

1.2 Base expertes et données patients

Il existe un certains nombre de base expertes pour les maladies rares, en particulier OrphaNet. Cette base référence plus de 9000 maladies rares, et leur associe les signes phénotypiques caractéristiques. C'est-à-dire que pour chaque maladie rare nous connaissons les symptômes généralement associés et nous connaissons la probabilité de présenter le symptôme sachant la maladie rare concernée. Cette base est inter-opérable avec HPO [Köhler and al, 2016] qui fournit un vocabulaire unifié sur les signes phénotypiques ainsi que l'ontologie associée.

L'accès à des données de parcours patient est plus difficile puisqu'il s'agit de données personnelles. Le Plan Maladie Rare prévoit le déploiement de la Banque National De Maladies Rares, inter-opérable avec HPO et Orphanet, devant permettre de récolter des données sur les parcours des patients à l'intérieur des centres maladies rares. Cependant ces données ne fournissent pas, ou très partiellement, le parcours médical avant l'entrée dans un centre expert.

Des données de type parcours patients consultation/remboursement (Electronic Health Record dans la littérature ou EHR) sont accessibles en ligne mais les données ne sont pas labelisées avec un identifiant Orpha pour les patients atteints de maladies rares. Certains travaux tels [Colbaugh et al., 2018, Tremblay et al., 2018] contournent cette difficulté en annotant la base EHR à partir d'informations expertes (par exemple si tel médicament a été pris il s'agit probablement de telle maladie rare) mais une telle approche ne peut qu'être limitée à certaines maladies rares.

Le lien entre les bases de données expertes du type OrphaNet et les bases de données de parcours patient n'est pas aisé à opérer. En effet, alors qu'OrphaNet associe des symptômes (identifiants HPO) à des maladies rares (identifiants Orpha), les EHR témoignent du parcours de santé (remboursements médicament, consultation généraliste/spécialiste). Un premier défi est d'associer à un évènement de santé d'EHR un symptôme HPO. [Zhang et al., 2019] propose ainsi une approche de type NLP pour annoter le texte libre d'EHR avec des identifiants HPO. Nous faisons ici l'hypothèse qu'un tel lien est possible et qu'il est ainsi possible de combiner des données expertes avec des données cliniques ce qui est essentiel pour le cas des maladies rares où les données cliniques seules ne suffisent pas mais permettent d'introduire une notion de temporalité absente d'Orphanet.

1.3 Objectif

Notre but ici est de montrer comment nous pouvons améliorer les parcours patients en réduisant l'errance de diagnostic pour les individus atteints de maladies rares. En particulier, nous nous focalisons sur un système d'alerte, lequel suggère à un patient de se rendre en centre expert avec l'information du diagnostic suspecté et des anomalies à rechercher.

3. `https://www.orpha.net/consor/cgi-bin/index.php`. Accessed on [23/02/2020].

Dans la section 2 nous présentons notre modélisation mathématique du problème de prise de décision : envoyer ou non un patient en CRMR. Dans la section 3 nous présentons l'application de notre approche sur des parcours patients simulés, dont le code R est disponible en libre accès. Nous enchaînons ensuite avec la discussion des résultats et une conclusion générale.

2 Modélisation

2.1 Parcours patient

Nous considérons qu'un individu ne peut contracter qu'un seul syndrome à la fois. Ce patient ne suspecte la présence de ce syndrome que via l'apparition de symptômes. Nous notons alors S_i, $i \in \mathbb{N}$, l'ensemble des symptômes observés le i-ème jour suivant l'apparition de la maladie et $H_t = \{S_i; 1 \leq i < t\}$ l'historique de ces observations jusqu'au jour $t-1$. Un exemple de parcours patient est présenté dans la figure 2, graphe de gauche.

2.2 Symptômes

Nous supposons que les symptômes se distribuent en trois groupes :

— Les symptômes latents, qui ne sont pas observables sans examen médical approprié. Exemple : malformations cardiaques ou neurologiques.

— Les symptômes visibles de façon permanente. Exemple : malformation externe.

— Les symptômes récurrents et passagers. Il s'agit de symptômes pouvant apparaître de manière récurrente puis disparaître et apparaître de nouveau. Exemple : migraines, otites.

Pour cette étude nous avons simulé [4] un graphe faisant le lien entre syndromes et symptômes, ainsi que la probabilité d'occurrence d'un symptôme dans la durée, comme représenté figure 1, graphe de gauche. Nous avons supposé que les temps d'occurrence suivaient une loi gaussienne, tronquée à gauche par 0. Pour les symptômes de type chronique, nous avons supposé que les délais entre présence/absence des symptômes suivaient une loi exponentielle.

La base OrphaNet nous fournit la prévalence d'une maladie rare ainsi que la probabilité des signes phénotypiques associés. En combinant connaissances expert et données de parcours patient nous pourrons calibrer des modèles paramétriques du délai d'apparition des symptômes sachant les syndromes.

2.3 Prise de décision : envoi en CRMR

Basé sur l'historique de l'individu, et suite à l'apparition d'au moins un symptôme, nous allons vérifier s'il est pertinent pour le patient de se diriger vers un centre spécialisé en maladie rare.

Nous estimons un prédicteur $\hat{f}$ de la probabilité qu'un individu soit atteint d'une maladie rare. Ce prédicteur prend en entrée l'historique H_t précédemment décrit. A partir de l'apparition du premier symptôme, nous vérifions chaque jour si, oui ou non, cette probabilité dépasse un seuil $\tau, \tau \in [0,1]$, préalablement fixé. Si le seuil est dépassé, le patient est envoyé en CRMR et son syndrome, rare ou non, sera découvert. La sollicitation des médecins spécialisés et les tests réalisés coûteront un certain prix. Si le seuil n'est pas dépassé, le patient ne sera pas envoyé en CRMR et nous considérons pour la modélisation qu'il est envoyé à la fin de la période de temps d'analyse. Nous notons cette procédure π_τ qui à un historique H_t associe la décision binaire $\mathbb{1}\{\hat{f}(H_t) > \tau\}$.

Soit les évènements $E = \{$Le patient a une maladie rare$\}$ et $A = \{$Nous l'avons dirigé vers un centre de

4. Les détails de la simulation sont fournis à l'adresse `github.com/FredericLoge/patientPathway`

maladie rare}. Nous pouvons alors définir un coût pour un parcours patient selon ces deux évènements :

$$\textbf{coût} = \begin{cases} coutErrranceParJour \cdot \text{tempsErrance} & \\ \quad + coutMedecinSpeMR & \text{si } E \cap A \\ coutErrranceParJour \cdot tempsMoyenErrance & \\ \quad + coutMedecinNonSpeMR \cdot nbMoyenMedecinConsultes & \text{si } E \cap \bar{A} \\ coutErrranceParJour \cdot \text{tempsErrance} & \\ \quad + coutMedecinSpeMR & \text{si } \bar{E} \cap A \\ coutErrranceParJour \cdot \text{tempsErrance} & \\ \quad + coutMedecinNonSpeMR & \text{si } \bar{E} \cap \bar{A} \end{cases} \tag{1}$$

où tempsErrance corresponds au temps entre l'observation du premier symptôme et la prise de décision (si prise de décision, sinon dernier jour de la simulation). Les coûts détaillés peuvent être recueillis auprès d'experts du domaine médical.

Notre objectif est de déterminer, pour un prédicteur de maladie rares $\hat{f}$ le seuil τ approprié pour optimiser le coût introduit dans l'équation 1. Formellement, nous cherchons à résoudre

$$\pi^* = \arg \min_{\tau \in [0,1]} \mathbb{E}_{\text{syndrome} \sim \nu} [\textbf{coût} \mid \text{syndrome}] \tag{2}$$

où ν est la probabilité de distribution sur les syndromes.

3 Résultats

Tous les détails sur les résultats et simulations présentées ici peuvent être obtenus à l'adresse `github.com/FredericLoge/patientPathway`. Des graines ont été soigneusement établies dans le code pour s'assurer de sa reproductibilité.

Pour faciliter l'étude, nous avons simulé un graphe symptômes-syndromes ainsi que les probabilités d'occurrence dans le temps des différents symptômes comme représenté dans la figure 1. Quatre syndromes ont été considérés dont un RAS (Rien à Signaler) et un appartenant à la catégorie des maladies rares, le syndrome #1. Un total de dix symptômes a été considéré.

Nous avons généré pour chaque syndrome 100 parcours patient sur une durée de quatre ans, à pas de temps journalier. Ces données ont été utilisées pour calibrer une forêt aléatoire (voir [Hastie et al., 2009] chapitre 15) prédisant la probabilité que le patient suivi ait une maladie rare en se basant sur les symptômes observés et la temporalité de leurs apparitions relativement au premier symptôme. Une fois que le prédicteur est entraîné, nous pouvons estimer la fonction objectif exprimée dans l'équation 2 sur une grille de τ assez fine, comme représenté sur la figure 2. Sur le graphe de gauche, la courbe verte indique la prédiction en fonction du temps, cette prédiction est utilisée pour notre système d'alarme. L'observation temporelle des symptômes 2, 7 et 9 sont également représentés. Sur le graphe de droite on observe le coût moyen en fonction du seuil choisi. Comme attendu, il y a un équilibre à trouver entre envoyer tout le monde en CRMR ($\tau = 0$, extrême-gauche du graphe), et n'envoyer personne ($\tau = 1$, extrême-droite du graphe).

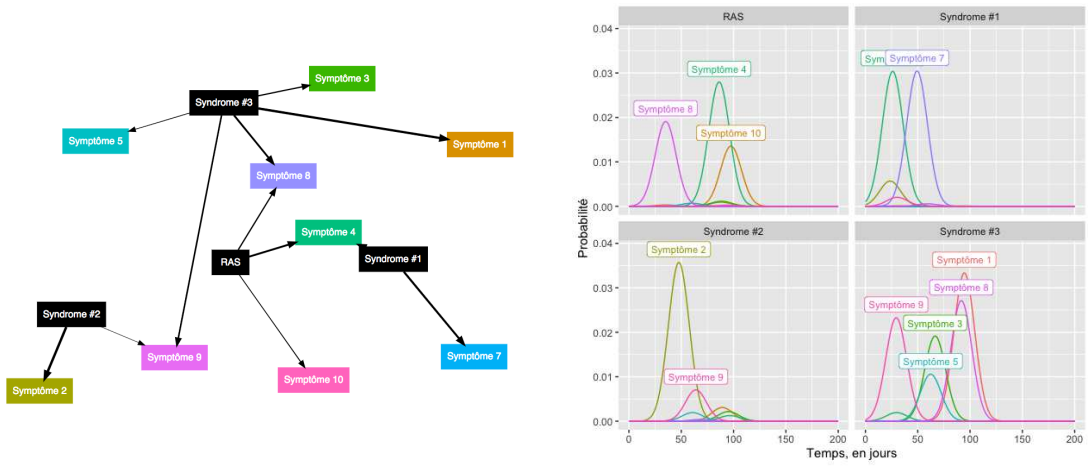


FIGURE 1 – *(Gauche)* Relations entre symptômes syndromes considérés, l'épaisseur du trait indiquant la probabilité qu'un symptôme se déclare lorsque le patient est atteint d'un syndrome. Note : les liens dont la probabilité étant inférieure à 15% ont été exclus pour une meilleure lisibilité. *(Droite)* Probabilité d'occurrence, dans le temps, des différents symptômes conditionnellement au syndrome considéré, le jour 0 étant l'instant où le syndrome se met au place.

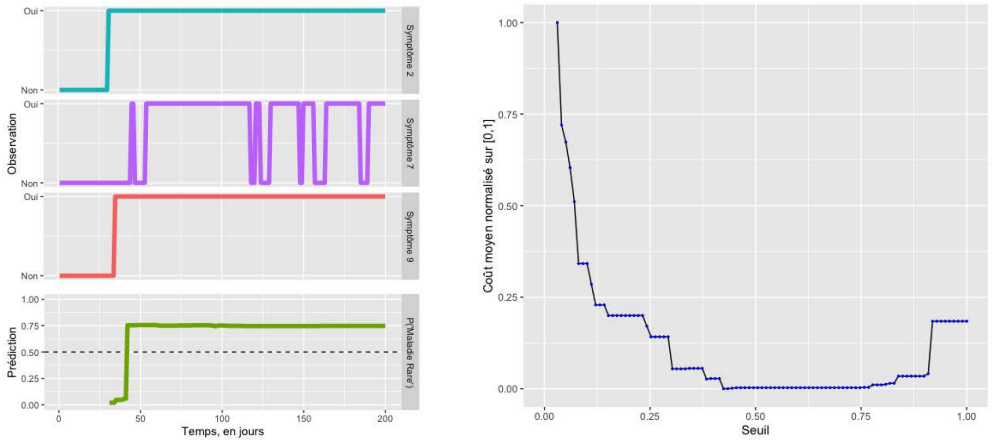


FIGURE 2 – *(Gauche)* Données simulées sur 200 jours, patient atteint du syndrome #1 (maladie rare pour rappel) qui donne souvent lieu aux symptômes 4 et 7, ici les symptômes 2 puis 9 et finalement 7 vont être observés (trois premières courbes), le symptôme 4 est présent mais non observé car il s'agit d'un symptôme latent. La prédiction du risque de maladie rare, opérée dès l'observation du premier symptôme, le symptôme 2, franchit le seuil de 0.5 dès l'observation du symptôme 7. *(Droite)*. En simulant les données comme représentées à gauche pour les différents syndromes et un grand nombre de fois, nous avons pu établir le coût moyen, comme défini dans la section 2.3 en fonction du seuil utilisé pour prendre la décision d'envoyer la personne en centre de maladie rare. Ce coût est ici représenté sur l'échelle [0 ;1]. A l'extrême gauche, nous avons le coût associé au fait d'envoyer tous les patients en centre spécialisé directement. A l'extrême droite, nous avons le coût associé au fait de n'envoyer aucun patient en centre spécialisé.

4 Conclusion et discussion

En France, un patient souffrant d'une maladie rare doit en moyenne attendre deux ans avant d'être diagnostiqué. A cette errance médicale sont associés des coûts économiques et humains élevés alors même que des centres experts maladies rares ont été déployés sur le territoire et donnent satisfaction lorsqu'ils sont sollicités.

Nous avons donc proposé ici une procédure pour la création d'un système d'alerte permettant de détecter dans des bases de données de parcours patients les patients atteints de maladie rare et devant être orientés vers des centres experts. Nous prenons en compte dans ces décisions les différents coûts engendrés par les diverses actions possibles. Ce système d'alerte est entraîné en générant des données issues d'un simulateur de parcours patient que nous construisons à partir de données expertes et dont les paramètres devront être calibrés avec des données cliniques. Les premiers résultats, obtenus sur données simulées, apparaissent prometteurs. Toutefois un important travail doit être encore être fourni sur les points suivants :

Calibrer le modèle Symptômes-Syndrome Ce modèle doit être, à termes, calibré à partir d'un mélange de données expertes et de parcours patients. Ces données peuvent être combinées au prix d'un travail significatif. Il s'agit de constituer un module associant à chaque évènement de santé d'un EHR un identifiant HPO. Une annotation experte doit également permettre de mieux modéliser l'apparition des symptômes au cours du temps : information sur l'âge typique d'apparition d'un symptôme, les durées typiques, la nature du symptôme.

Améliorer et valider la modélisation Dans notre modèle de parcours patient, le médecin n'était présent qu'une fois la décision de se rendre en CRMR prise ou la durée de simulation terminée. Dans la réalité, l'occurence d'un symptôme incitera probablement la personne à consulter une personne du corps médical. Ces agents doivent être impérativement intégrés au parcours patient. Des approches de validation par indicateurs comme celle développé dans [Prodel, 2017] seront utilisées.

Le problème de contrôle Nous avons considéré ici une stratégie particulière pour la réorientation vers les centres experts qui exécute un choix binaire si l'on dépasse un certain seuil de probabilité que le patient présente une maladie rare. Une future direction est de formuler le problème par un processus décisionnel de Markov où la stratégie pourra être apprise par des algorithmes d'apprentissage par renforcement se basant sur des données du simulateur et n'ayant plus nécessairement une forme aussi spécifique.

Remerciements : Ce travail a bénéficié d'une aide de l'État gérée par l'Agence Nationale de la Recherche au titre du programme d'Investissements d'avenir portant la référence ANR-19-P3IA-0001.

Références

[Colbaugh et al., 2018] Colbaugh, R., Glass, K., Tremblay, M., and Rudolf, C. (2018). Learning to identify rare disease patients from electronic health records. *AMIA Annu Symp Proc.*

[Hastie et al., 2009] Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning : data mining, inference, and prediction.* Springer Science & Business Media.

[Köhler and al, 2016] Köhler, S. and al (2016). The Human Phenotype Ontology in 2017. *Nucleic Acids Research*, 45.

[Nguengang and al, 2019] Nguengang, W. S. and al (2019). Estimating cumulative point prevalence of rare diseases : analysis of the orphanet database. *Eur J Hum Genet.*

[Prodel, 2017] Prodel, M. (2017). *Process discovery, analysis and simulation of clinical pathways using health-care data.* Theses, Université de Lyon.

[Tremblay et al., 2018] Tremblay, M., Colbaugh, R., Glass, K., and Rudolf, C. (2018). Robust ensemble learning to identify rare disease patients from electronic health records.

[Zhang et al., 2019] Zhang, J., Zhang, X., Sun, K., Yang, X., Dai, C., and Guo, Y. (2019). Unsupervised annotation of phenotypic abnormalities via semantic latent representations on electronic health records.

Étude de la dépendance des extrêmes en grande dimension

Nicolas MEYER [1] & Olivier WINTENBERGER [2]

[1] *Sorbonne Université, LPSM, France*
nicolas.meyer@upmc.fr
[2] *Sorbonne Université, LPSM, France*
olivier.wintenberger@upmc.fr

Résumé. Identifier les directions dans lesquelles des événements exceptionnels apparaissent est un des problèmes majeurs de la théorie multivariée des valeurs extrêmes. D'un point de vue théorique, la majeure partie de l'information concernant de tels événements est contenue dans la mesure spectrale, qui apparaît comme la limite de la composante angulaire de vecteurs aléatoires à variation régulière. Estimer cette mesure s'avère être un point délicat, notamment en grande dimension. Dans cette présentation, nous introduisons une méthode de réduction de la dimension basée sur la projection euclidienne sur le simplexe. Cette projection a été étudiée dans le cadre des valeurs extrêmes par Meyer & Wintenberger (2020+) qui ont établi plusieurs résultats théoriques. La présentation s'attachera à exposer une approche statistique qui s'appuie sur une sélection de modèles pour identifier des groupes de coordonnées susceptibles d'être extrêmes simultanément.

Mots-clés. Extrêmes multivariés, measure spectrale, projection sur le simplexe, réduction de la dimension, variation régulière, variation régulière parcimonieuse.

Abstract. Identifying directions where exceptional events occur is one of the major problems of multivariate extreme value theory. From a theoretical point of view most of the information concerning such events is contained in the spectral measure which appears as the limit of the angular component of regularly varying random vectors. Estimating this measure is a delicate point especially in large dimensions. In this presentation we introduce a dimension reduction method based on the Euclidean projection onto the simplex. This projection has been studied in the context of extreme values by Meyer & Wintenberger (2020+) who established several theoretical results. The presentation will focus on a statistical approach that uses model selection to identify groups of coordinates that are likely to be extreme simultaneously.

Keywords. Dimension reduction, multivariate extremes, projection onto the simplex, regular variation, sparse regular variation, spectral measure.

1 Valeurs extrêmes et variation régulière

1.1 Variation régulière

Étudier les valeurs extrêmes générées par un vecteur aléatoire $\mathbf{X} \in \mathbb{R}_+^d$, $d \geq 2$, revient à étudier le comportement de la queue de distribution de $\mathbf{X}$. Dans ce contexte, il est courant de supposer que le vecteur $\mathbf{X}$ est à variation régulière : il existe un vecteur aléatoire $\mathbf{\Theta}$ sur la sphère unité telle que

$$\mathbb{P}(|\mathbf{X}| > rt,\, \mathbf{X}/|\mathbf{X}| \in \cdot \mid |\mathbf{X}| > t) \xrightarrow{d} r^{-\alpha}\mathbb{P}(\mathbf{\Theta} \in \cdot)\,, \quad t \to \infty\,, \tag{1}$$

cf Resnick (2007). Dans ce cas, $|\cdot|$ désigne n'importe quelle norme sur $\mathbb{R}^d$. Le vecteur limite $\mathbf{\Theta}$ est alors appelé vecteur spectral tandis que sa loi est appelée mesure spectrale. La convergence (1) permet de séparer l'étude de la composante radiale des extrêmes $|\mathbf{X}|/t$ de celle de la composante angulaire $\mathbf{X}/|\mathbf{X}|$. Cette dernière concentre l'information sur la localisation et la dépendance des valeurs extrêmes. L'étude de la mesure spectrale est donc un point central de la théorie des extrêmes multivariés.

Il est souvent intéressant (voir par exemple Goix et al. (2017)) d'étudier le comportement du vecteur spectral sur les sous-ensembles C_β de la sphère définis par

$$C_\beta = \{\mathbf{x} \in \mathbb{R}_+^d,\, |\mathbf{x}| = 1,\, x_j > 0 \text{ pour } j \in \beta,\, x_j = 0 \text{ pour } j \notin \beta\}\,,$$

pour $\beta \subset \{1, \ldots, d\}$. En effet, la mesure spectrale met de la masse sur un tel ensemble si des événements extrêmes apparaissent conjointement dans la direction β. On est ainsi ramené à l'estimation des probabilités $\mathbb{P}(\mathbf{\Theta} \in C_\beta)$. Cependant, l'estimation de ces quantités se révèle délicate pour essentiellement deux raisons. Tout d'abord, le nombre de probabilités à estimer croît exponentiellement en la dimension. Par ailleurs, si $\beta \neq \{1, \ldots, d\}$ vérifie $\mathbb{P}(\mathbf{\Theta} \in C_\beta) > 0$, alors la mesure spectrale charge la frontière de C_β (qui est le sous-ensemble C_β lui-même) et donc la convergence (1) ne s'applique pas.

L'idée proposée par Meyer et Wintenberger (2020+) pour contourner ce problème est de remplacer le vecteur unitaire $\mathbf{X}/|\mathbf{X}|$ de (1) par un autre projeté qui permet de mieux tenir compte de la masse mise par la mesure spectrale sur les sous-ensembles C_β. Cette modification de la convergence (1) donne alors naissance à la notion de variation régulière parcimonieuse.

1.2 Variation régulière parcimonieuse

Introduite principalement par Duchi et al. (2008), la projection euclidienne sur le simplexe a connu un usage divers et varié, notamment en théorie de l'apprentissage.

Dans la suite, $|\cdot|$ désigne la norme ℓ^1 et $\mathbb{S}^{d-1}$ (respectivement $\mathbb{S}_+^{d-1}$) désigne la sphère associée (respectivement le simplexe). Pour $z > 0$ et $\mathbf{v} \in \mathbb{R}_+^d$ le vecteur projeté $\pi_z(\mathbf{v})$ est l'unique vecteur $\mathbf{w}$ de $\mathbb{S}_+^{d-1}(z) := \{\mathbf{x} \in \mathbb{R}_+^d,\, x_1 + \ldots + x_d = 1\}$ qui minimise la quantité

$|\mathbf{w}-\mathbf{v}|_2$, où $|\cdot|_2$ désigne la norme ℓ^2. On note alors π_z la projection euclidienne sur $\mathbb{S}_+^{d-1}(z)$ et plus généralement π la projection sur le simplexe $\mathbb{S}_+^{d-1}$. Cette manière de projeter permet de rendre les vecteurs parcimonieux, c'est-à-dire avec plusieurs coordonnées nulles. Elle permet également de mieux rendre compte du comportement des extrêmes sur les ensembles C_β.

Définition 1 (Variation régulière parcimonieuse)**.** *Un vecteur* $\mathbf{X}$ *à valeurs dans* $\mathbb{R}_+^d$ *est dit à variation régulière parcimonieuse s'il existe un vecteur aléatoire* $\mathbf{Z}$ *défini sur le simplexe et une variable aléatoire positive* Y *tels que*

$$\mathbb{P}\Big(\big(|\mathbf{X}|/t, \pi(\mathbf{X}/t)\big) \in \cdot \mid |\mathbf{X}| > t\Big) \xrightarrow{d} \mathbb{P}((Y, \mathbf{Z}) \in \cdot)\,, \quad t \to \infty\,. \tag{2}$$

Le vecteur limite $\mathbf{Z}$ doit être vu comme la limite angulaire obtenue après avoir remplacé $\mathbf{X}/|\mathbf{X}|$ par $\pi(\mathbf{X}/t)$ dans l'Equation (1). Par continuité de la projection, la notion de variation régulière standard (Equation (1)) implique celle de variation régulière parcimonieuse. Meyer et Wintenberger (2020+) ont prouvé que sous des hypothèses assez faibles, les deux notions sont en fait équivalentes.

L'intérêt principal de la Définition 1 est de pouvoir approcher le comportement des extrêmes de $\mathbf{X}$ sur les ensembles C_β. En effet, en reprenant les notations précédentes, on a la convergence suivante pour tout $\beta \subset \{1, \dots, d\}$:

$$\mathbb{P}(\pi(\mathbf{X}/t) \in C_\beta \mid |\mathbf{X}| > t) \to \mathbb{P}(\mathbf{Z} \in C_\beta)\,, \quad t \to \infty\,. \tag{3}$$

L'objectif est donc d'estimer le support de la distribution de $\mathbf{Z}$ via l'estimation des probabilités $\mathbb{P}(\mathbf{Z} \in C_\beta)$, pour $\beta \subset \{1, \dots, d\}$, le but étant de détecter lesquelles de ces probabilités sont positives. Autrement dit, il s'agit d'identifier l'ensemble

$$\mathcal{S}(\mathbf{Z}) := \{\beta \subset \{1, \dots, d\}, \mathbb{P}(\mathbf{Z} \in C_\beta) > 0\}\,.$$

Cet ensemble $\mathcal{S}(\mathbf{Z})$ rassemble toutes les directions β sur lesquelles le vecteur angulaire $\mathbf{Z}$ met de la masse. On note s^* son cardinal. L'objectif est alors de proposer une approche statistique pour décider quelles directions β appartiennent à $\mathcal{S}(\mathbf{Z})$.

2 Estimation

On considère désormais une suite de vecteurs aléatoires indépendants et identiquement distribués à variation régulière $\mathbf{X}, \mathbf{X}_1, \mathbf{X}_2, \dots$ de vecteur spectral $\mathbf{\Theta}$. On considère également une variable aléatoire Y de loi de Pareto de paramètre $\alpha > 0$, indépendante de $\mathbf{\Theta}$. Enfin, on pose $\mathbf{Z} = \pi(Y\mathbf{\Theta})$.

Le cadre classique en statistique des valeurs extrêmes est de considérer une suite positive $(u_n)_{n\in\mathbb{N}}$ telle que $u_n \to \infty$. Cette suite joue le rôle du seuil t dans les Equations (1) et (2). Cela signifie que pour $n \in \mathbb{N}$, la quantité u_n doit être vue comme le seuil

au-dessus duquel les données $\mathbf{X}_1, \cdots, \mathbf{X}_n$ sont considérées comme des valeurs extrêmes. Il est également usuel de définir un niveau $k = k_n = n\mathbb{P}(|\mathbf{X}| > u_n)$ et de supposer que $k_n \to \infty$ quand $n \to \infty$. Il est à noter que l'hypothèse $u_n \to \infty$ implique que $k_n/n = \mathbb{P}(|\mathbf{X}| > u_n) \to 0$. Ainsi, k_n tend vers l'infini à une vitesse plus lente que n. Un estimateur naturel non biaisé pour k_n est $\hat{k} = \hat{k}_n = \sum_{j=1}^n \mathbb{1}_{|\mathbf{X}_j|>u_n}$ qui correspond au nombre de dépassements au-dessus du seuil u_n, c'est-à-dire au nombre de valeurs extrêmes.

Notre objectif est d'estimer les probabilités $p(\beta) := \mathbb{P}(\mathbf{Z} \in C_\beta)$ pour $\beta \subset \{1, \ldots, d\}$. Ces probabilités apparaissent comme les limites des probabilités pré-asymptotiques $p_n(\beta) := \mathbb{P}(\pi(\mathbf{X}/u_n) \in C_\beta \mid |\mathbf{X}| > u_n)$ (voir Equation (3)). Le problème principal est alors de décider si $p(\beta)$ est positif ou nul. On définit pour cela l'estimateur

$$T_n(\beta) := \sum_{j=1}^{n} \mathbb{1}\{\pi(\mathbf{X}_j/u_n) \in C_\beta,\, |\mathbf{X}_j| > u_n\}\,,$$

pour $\beta \subset \{1, \ldots, d\}$. L'idée est de sélectionner parmi les valeurs extrêmes celles qui sont projetées dans l'ensemble C_β. Un calcul rapide montre que $\mathbb{E}[T_n(\beta)] = k_n p_n(\beta)$. D'autres résultats concernant ces estimateurs sont rassemblés dans le théorème suivant.

Théorème 1. *On reprend les notations précédentes.*

1. *(Consistance). Le vecteur $k_n^{-1}(T_n(\beta))_{\beta\subset\{1,\ldots,d\}}$ converge en probabilité vers $\mathbf{p} := (p(\beta))_{\beta\subset\{1,\ldots,d\}}$.*
2. *(Normalité asymptotique). On a la convergence en loi suivante :*

$$\sqrt{k_n}\,\mathrm{Diag}\left(\mathbf{p}(\mathcal{S}(\mathbf{Z}))\right)^{-1/2}\Big(\frac{\mathbf{T}_n(\mathcal{S}(\mathbf{Z}))}{k_n} - p_n(\mathcal{S}(\mathbf{Z}))\Big) \xrightarrow{d} \mathcal{N}(0, Id_{s^*})\,, \quad n \to \infty\,.$$

où $\mathbf{T}_n(\mathcal{S}(\mathbf{Z}))$ (resp. $\mathbf{p}_n(\mathcal{S}(\mathbf{Z}))$) correspond au vecteur de $\mathbb{R}^{s^*}$ dont les composantes sont les $T_n(\beta)$ (resp. $p_n(\beta)$) pour $\beta \in \mathcal{S}(\mathbf{Z})$ (on se restreint aux probabilités positives).

3 Sélection de modèle

La répartition des k_n données extrêmes sur les $2^d - 1$ sous-ensembles $(C_\beta)_{\beta\subset\{1,\ldots,d\}}$ suggère d'utiliser le modèle multinomial $\mathcal{M}(k_n, \tilde{\mathbf{p}})$ où le paramètre $\tilde{\mathbf{p}}$ est défini par

$$\tilde{\mathbf{p}} = (\overbrace{\tilde{p}_1, \ldots, \tilde{p}_s, \underbrace{\tilde{p}, \ldots, \tilde{p}}_{r-s}, 0, \ldots, 0}^{2^d-1 \text{ composantes}})\,,$$

avec $\tilde{p}_1 \geq \ldots \geq \tilde{p}_s, \tilde{p} \in (0, 1)$ satisfaisant la contrainte :

$$\tilde{p}_1 + \ldots + \tilde{p}_s + (r - s)\tilde{p} = 1\,.$$

L'idée est de séparer les sous-ensembles $(C_\beta)_{\beta \subset \{1,\dots,d\}}$ en trois catégories. La première correspond aux sous-ensembles sur lesquels les extrêmes apparaissent, leur probabilité d'apparition étant alors $\tilde{p}_j$. La deuxième catégorie correspond aux sous-ensembles sur lesquels peu de données extrêmes sont apparues. Ce phénomène survient notamment en raison d'un biais qui provient de l'aspect non-asymptotique de l'étude. Ce biais est modélisé par une probabilité d'occurence $\tilde{p}$ considérée comme proche de 0. Enfin, la dernière catégorie concerne les sous-ensembles C_β sur lesquels aucune donnée n'est apparue ; on estime alors que ces sous-ensembles ne concentrent pas de valeurs extrêmes (d'où une probabilité d'occurence nulle). L'objectif est alors d'ajuster au mieux le nombre s de faces significatives. Cet ajustement s'effectue via une sélection de modèle de type AIC.

La sélection de modèle doit aussi tenir compte du choix de k_n dans le modèle multinomial $\mathcal{M}(k_n, \tilde{\mathbf{p}})$, cette quantité correspondant au nombre de données considérées comme extrêmes. Il s'agit dès lors de comparer les modèles pour différents choix de k_n, ce qui modifie alors le nombre de données utilisées. L'approche classique de sélection de modèle doit alors être étendue pour pouvoir tenir compte de cette variation du nombre de données.

Bibliographie

Duchi, J. Shalev-Shwartz, S. Singer, Y. Chandra, T. (2008), *Efficient Projections onto the ℓ_1-Ball for Learning in High Dimensions*, ICML.

Goix, N. Sabourin, A. Clémençon, S. (2017), *Sparsity in Multivariate Extremes with Applications to Anomaly Detection*, Journal of Multivariate Analysis.

Meyer, N. et Wintenberger, O. (2020+), Detection of extremal directions via Euclidean projections, arXiv : 1907.00686.

Resnick, S.I. (2007), *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*, Springer.

Tests d'hypothèses sur les coefficients de Fourier dans un modèle de régression non paramétrique

Zaher Mohdeb[1] & Abdelkader Mokkadem[2]

[1]*Ecole Nationale Polytechnique de Constantine*
et Laboratoire de Mathématiques et Sciences de la Décision
Université frères Mentouri de Constantine, Algérie
E-mail: z.mohdeb@gmail.com
[2]*Département de Mathématiques,*
Université de Versailles Saint-Quentin En Yvelines,
45, Avenue des Etats-Unis, 78035 Versailles Cedex, France
E-mail: abdelkader.mokkadem@uvsq.fr

Résumé. On considère le modèle de régression non paramétrique de fonction de régression f. Une procédure de test d'hypothèse sur les coefficients de Fourier de f est proposée. On obtient le comportement asymptotique de la statistique de test proposée, on a donc ainsi le niveau et la puissance asymptotique du test. De tels tests peuvent, en particulier, être utilisés pour comparer deux signaux bruités dans une bande de fréquence. Un autre exemple est le test de l'hypothèse "f est un polynôme trigonométrique". Une étude par simulation est menée, pour des petites tailles d'échantillon, afin de montrer la performance du test proposé.

Mots-clés. Modèle de régression non linéaire, Coefficient de Fourier empirique, Test non paramétrique.

Abstract. We consider the nonparametric regression model with regression function f. A hypothesis testing procedure on the Fourier coefficients of f is proposed. We obtain the asymptotic weak behaviour of the proposed test, then we have the level and the asymptotic power of the test. Such tests can be used in particular to compare two noisy signals in a frequency band. Another example is the test of the hypothesis that "f is a trigonometric polynomial". A simulation study is conducted, for small sample size, to demonstrate the performance the proposed test.

Keywords. Nonlinear regression model, Empirical Fourier coefficient, Nonparametric test.

1 Introduction

On considère le modèle de régression non paramétrique

$$Y_{j,n} = f(t_{j,n}) + \varepsilon_{j,n} \quad j = 1, \ldots, n \tag{1}$$

où $t_{j,n} = j/n$, $f : [0,1] \longrightarrow \mathbb{R}$ est une fonction inconnue et $\varepsilon_{j,n}$, $j = 1, \ldots, n$, forment un tableau triangulaire de variables aléatoires centrées et de variance finie σ^2 et pour tout n les variables aléatoires $\varepsilon_{1,n}, \ldots, \varepsilon_{n,n}$ sont indépendantes.

L'objet de ce travail est de construire des tests d'hypothèses sur les coefficients de Fourier de f. Plus précisément, soit c_k, $k \in \mathbb{Z}$, les coefficients de Fourier de f et soit $\mathcal{I} = \{i_1, i_2, \ldots\}$ un sous-ensemble de $\mathbb{N}$ avec $i_1 < i_2 < \cdots$ et $Card(\mathcal{I}) = \infty$. On veut construire un test de l'hypothèse nulle

$$H_0 : \ c_k = 0 \ \ \forall k \in \mathcal{I} \ \text{ contre l'hypothèse } \quad H_1 : \ \exists k \in \mathcal{I} \ \ c_k \neq 0 \, . \tag{2}$$

Dans le test de l'hypothèse (2), il est clair que H_0 est vraie si et seulement si $\sum_{|k| \in \mathcal{I}} |c_k|^2 = 0$. Le test sera donc basé sur une estimation de cette quantité. On commence par estimer c_k par l'estimation empirique $\hat{c}_k = \frac{1}{n} \sum_{j=1}^{n} Y_{j,n} e^{-2i\pi kj/n}$. Ensuite on considère une suite d'entiers $p = p(n)$ telle que $\lim\limits_{n \to \infty} p(n) = \infty$ et on pose $\mathcal{I}_p = \{i_1, \ldots, i_p\}$. La statistique de test est alors $\hat{T}_{n,p} = \sum_{|k| \in \mathcal{I}_p} |\hat{c}_k|^2$. L'hypothèse H_0 est rejetée si $\hat{T}_{n,p} > t_\alpha$ où t_α est un nombre réel positif déterminé par le niveau α du test.

Nos résultats principaux énoncés ci-dessous donnent la loi asymptotique de $\hat{T}_{n,p}$ et permettent donc d'obtenir une valeur asymptotique de t_α.

La mise en oeuvre du test nécessite la connaissance de σ^2, ce qui, en pratique, n'est jamais le cas; on a donc besoin d'un estimateur. On peut utiliser l'estimateur de Gasser et al (1986); on propose également un estimateur adapté à l'hypothèse nulle et on donne les conditions qui permettent d'utiliser cet estimateur. Enfin, on peut comparer sur des simulations les effets de ces estimateurs sur la puissance empirique du test pour les petits échantillons ($n = 100$).

L'usage des coefficients de Fourier empiriques de f pour construire des tests d'hypothèses dans le modèle (1) est abordé par Eubank et Spiegelman (1990) et Mohdeb et Mokkadem (2001). Eubank et Spiegelman (1990) construisent un test de linéarité de f dans le cas d'un modèle normal en se ramenant à tester la nullité de la partie non linéaire. Mohdeb et Mokkadem (2001) construisent des tests d'hypothèses sur les coefficients de Fourier de f dans une bande de fréquences donnée. Cox et Koh (1989) donnent un test de l'hypothèse "f est un polynôme de degré inférieur à m". Jayasuriya (1996) généralise l'approche de Eubank et Spiegelman (1990) pour tester l'hypothèse: "f est un polynôme", dans le cas d'un modèle non normal.

2 Hypothèses et résultats

On s'intéresse au test (2) pour le modèle (1) et on suppose que

• (C1): f satisfait la condition de Hölder d'ordre δ, avec $\frac{1}{2} < \delta \leq 1$, i.e. il existe une constante positive M telle que $|f(s) - f(t)| \leq M|s - t|^\delta$ pour tout s, $t \in [0,1]$.

• (C2): Pour tout entier n, $\varepsilon_{j,n}$, $j = 1, \ldots, n$, sont des variables aléatoires réelles i.i.d. d'espérance nulle et de variance inconnue σ^2.

La convergence en loi est notée par $\xrightarrow{\mathcal{L}}$. On note aussi c_k, $k \in \mathbb{Z}$ les coefficients de Fourier de f et $\mathcal{I} = \{i_1, i_2, \ldots\}$ un sous-ensemble de $\mathbb{N}$ avec $i_1 < i_2 < \cdots$ et $Card(\mathcal{I}) = \infty$.

2.1 Résultats principaux

On considère $\mathcal{I}_p = \{i_1, \ldots, i_p\}$ où $p = p(n)$ est une suite croissante de limite infinie. On introduit les hypothèses suivantes:
• (A1): Pour tout entier n, $\varepsilon_{1,n} \sim \mathcal{N}(0, \sigma^2)$.
• (A2): $\lim_{n\to+\infty} \{n^{-2\delta+1} p(n)\} = 0$.
On a alors le théorème suivant.

Théorème 1 *Si les hypothèses* $(A1)$ *et* $(A2)$ *sont satisfaites, alors*

$$\frac{n \sum_{|k|\in\mathcal{I}_p} |\hat{c}_k - c_k|^2 - u_p\sigma^2}{\sigma^2\sqrt{2u_p}} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1)$$

où $u_p = 2p - 1$ *si* $0 \in \mathcal{I}$ *et* $u_p = 2p$ *si* $0 \notin \mathcal{I}$.

2.2 Construction du test

Le résultat du théorème 1 permet de construire le test quand la variance σ^2 est connue; cependant en pratique σ^2 est inconnue et il faut donc l'estimer. On peut utiliser l'estimateur de Gasser et al (1986). On peut aussi considérer un estimateur $\hat{\sigma}^2$ de σ^2 qui converge sous H_0. Plus précisément, soit

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{n} |Y_{j,n} - \hat{f}(j/n)|^2 \tag{3}$$

où $\hat{f}$ est défini de la manière suivante.
Notons J le complémentaire de $\mathcal{I}$ dans $\mathbb{N}$. Si $Card(J) = \infty$, on considère une suite croissante d'entiers $q = q(n)$ telle que $\lim_{n\to\infty} q(n) = +\infty$ et on pose

$$\hat{f}(t) = \begin{cases} \displaystyle\sum_{|k|\in J} \hat{c}_k e^{2i\pi kt} & \text{si} \quad Card(J) < \infty\,, \\ \displaystyle\sum_{|k|\le q, |k|\in J} \hat{c}_k e^{2i\pi kt} & \text{si} \quad Card(J) = \infty\,. \end{cases} \tag{4}$$

Quand J est fini, on a le résultat suivant.

Corollaire 1 *Si* $Card(J) < \infty$ *et si les hypothèses du théorème* 1 *sont satisfaites, alors sous* H_0*, on a:*

$$\frac{n \sum_{|k|\in\mathcal{I}_p} |\hat{c}_k|^2 - u_p\hat{\sigma}^2}{\hat{\sigma}^2\sqrt{2u_p}} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1)$$

où $u_p = 2p - 1$ *si* $0 \in \mathcal{I}$ *et* $u_p = 2p$ *si* $0 \notin \mathcal{I}$.

Quand J est infini, il nous faut introduire les hypothèses:
• (A3): $\lim_{n\to\infty}\{\sqrt{p(n)}\sum_{|k|>q(n)}|c_k|\}=0$.
• (A4): $\lim_{n\to\infty}\{n^{-1}p(n)q^2(n)\}=0$.
On a alors le corollaire suivant.

Corollaire 2 *Si $Card(J)=\infty$ et si les hypothèses $(A1)-(A4)$ sont satisfaites, alors sous H_0, on a:*

$$\frac{n\sum_{|k|\in\mathcal{I}_p}|\hat{c}_k|^2-u_p\hat{\sigma}^2}{\hat{\sigma}^2\sqrt{2u_p}}\xrightarrow[n\to+\infty]{\mathcal{L}}\mathcal{N}(0,1)$$

où $u_p=2p-1$ si $0\in\mathcal{I}$ et $u_p=2p$ si $0\notin\mathcal{I}$.

2.3 Simulations

On considère de le test de l'hypothèse $H_0: f\equiv 0$. Il s'agit de comparer, sur des simulations, les deux statistiques $T_n^{(i)}=\left(n\sum_{|k|\le p}|\hat{c}_k|^2-(2p+1)\hat{\sigma}_i^2\right)/\left(\hat{\sigma}_i^2\sqrt{2(2p+1)}\right)$, $i=1,2$, où $\hat{\sigma}_1^2=\hat{\sigma}^2$ donnée par (3) et $\hat{\sigma}_2^2=\frac{2}{3(n-2)}\sum_{j=2}^{n-1}(\frac{1}{2}Y_{j-1,n}+\frac{1}{2}Y_{j+1,n}-Y_{j,n})$ est l'estimateur de Gasser et al (1986). On simule un échantillon de taille $n=100$ du modèle (1) avec $\varepsilon_j\sim\mathcal{N}(0,\sigma^2)$ et $\sigma=0.05$, 0.10, 0.20, 0.50 et 1.00. Les valeurs de p et les niveaux considérés sont $p=5$, 10 et 17 et $\alpha=0.01$, 0.05 et 0.10. D'après les résultats obtenus de nos simulations, on constate comme Eubank et Spiegelman (1990), que les valeurs critiques empiriques pour $T_n^{(2)}$ sont plus grandes que les valeurs critiques théoriques. Par contre les valeurs critiques empiriques pour $T_n^{(1)}$ sont proches des valeurs théoriques.

Bibliographie

[1] Cox, D. and Koh, E. (1989), A smoothing Spline Based Test of Model Adequacy in Polynomial Regression. *Ann. Inst. Statist. Math.,* **41**, 2, 383-400.
[2] Eubank, R. L. and Spiegelman, C. H. (1990), Testing the Goodness-of-Fit of a Linear Model Via Nonparametric Regression Techniques. *J. Amer. Statist. Assoc.* **85**, 410, 387-392.
[3] Gasser, T., Sroka, L. and Jennen-Steinmetz, C. (1986), Residual Variance and Residual Pattern in Nonlinear Regression. *Biometrika,* **73**, 625-633.
[4] Jayasuriya, B. R. (1996), Testing for Polynomial Regression Using Nonparametric Regression Techniques. *J. Amer. Statis. Assoc.* **91**, 436, 1626-1631.
[5] Mohdeb, Z. and Mokkadem, A. (2001). Testing Hypotheses On Fourier Coefficients in Nonparametric Regression Model. *Journal of Nonparametric Statistics*, **13**, 605-629.
[6] Staniswalis, J. G. and Severini, T. A. (1991), Diagnostics for Assessing Regression Models. *J. Amer. Statis. Assoc.*, **86**, 415, 684-692.

Exploring the Hidden Parts of the Asteroid Belt with Data Imputation

Max Mahlke [1] & Rémi Flamary [1] & Pierre-Alexandre Mattei [2] & Benoit Carry [1]

[1] *Université Côte d'Azur, Observatoire de la Côte d'Azur, CNRS, Laboratoire Lagrange, France* *max.mahlke@oca.eu*

[2] *Université Côte d'Azur, Inria, Maasai project-team, Laboratoire J.A. Dieudonné, UMR CNRS 7351, France*

Résumé. L'étude des astéroïdes permet de contraindre les modèles de formation du système solaire. Nous cherchons donc à les classifier à travers leur composition chimique, à partir de leurs propriétés observables (astronomiquement). Cependant, les caractéristiques pertinentes ne sont pas toujours observées pour tous les astéroïdes du système solaire et il est nécessaire d'estimer ces valeurs avec des méthodes d'imputation statistique. Nous présentons dans ce document les données disponibles ainsi que de premiers résultats en imputation de données manquantes. Nous discutons également de questions ouvertes et des étapes suivantes en classification.

Mots-clés. Astronomie, planétologie, astéroïdes, imputation de données manquantes

Abstract. Asteroids provide tight constraints on current models of planetary system formation. We aim to classify them by their chemical compositions, using a clustering approach on different observational features. These features are not universally available for all asteroids, hence we explore how to fill gaps in the data using data imputation. We present the available data sets, the first preliminary results, highlight open questions, and outline the next steps of classification.

Keywords. astronomy, planetary science, asteroids, data imputation, clustering

1 Introduction

Asteroids are small bodies of the Solar System, mostly concentrated in the main asteroid belt between Mars and Jupiter. Their orbital distributions and chemical compositions provide tight constraints on models of the formation of the Solar System (DeMeo & Carry, 2014). Planetary bodies which formed on the outskirts of the Solar System have not been chemically altered by the solar irradiation and contain pristine materials, such as organics and water ice. Meanwhile, asteroids which accumulated in close vicinity to the Sun exhibit stony, metallic features (e.g. Chapman, 1971). These asteroids are referred to as C-types and S-types, respectively.

Observations of asteroids are based on their reflectance of the solar emission, which is brightest in the visible and near-infrared regime. Even though they are close sources in an astronomical context, they are challenging telescope targets as they tend to be small in

diameter (∼10 m-10 km) and display large apparent angular velocities in the sky. Hence, of the about 1 million known asteroids in the Solar System, only a few thousand are bright enough to acquire detailed reflectance spectra. For the remaining several hundred thousand, only sparse observational data is available.

We aim to fill in the gaps in the observational feature space of asteroids using data imputation. The final objective is to perform a clustering on the whole dataset using imputed data or even perform clustering simultaneously with imputation.

2 Observational data

We have gathered data on 280 000 asteroids, about a quarter of the known population. The data can be divided into four types of observational features, depicted in Figure 1. Below, we briefly outline these features and evaluate the content of information towards the chemical composition of the asteroid. We focus here on the three main compositions, the S- and C-type mentioned above, as well as the X-type.

Spectra Most of our knowledge on their chemical composition was derived by comparing the reflectance spectra of asteroids in the visible (VIS) and near-infrared (NIR) with those of meteorites. Four principal characteristics describe asteroids in the spectral space: the overall slope, the presence or absence of absorption bands at 0.7 μm, 1.0 μm, and 2.0 μm, introduced by different minerals. The spectra carry the most compositional information, however, they can only be acquired for the brightest asteroids. We have acquired about 1,000 VISNIR spectra, 1,000 NIR spectra, and 6,000 VIS spectra of asteroids (e.g. Bus and Binzel 2002). Figure 1 (a) shows the VISNIR spectra of the S-, C-, and X-type classes.

Colours The colours of an asteroid are observed integrating parts of the reflectance spectra. In essence, they are low-resolution spectra which depict its slope in different parts. In the data sample, we have gathered magnitudes (i.e. the brightness in certain wavelengths) of 280,000 asteroids acquired in the most observed wavelength ranges, denounced u, g, r, i, z (from the Sloan Digital Sky Survey, Ivezić et al. 2001), and Y, J, H, Ks (from ESO VISTA, Popescu et al. 2016). Figure 1 (c) depicts the magnitudes of the S, C, and X classes in these filters. "Colours" refers to the differences of these magnitudes.

Albedo The albedo of an asteroid describes the ratio of the incoming sunlight that is reflected by its surface. Asteroids with low albedos (close to zero) will appear dark, while some asteroids will be brighter with albedos around 0.5. Figure 1 (b) shows that the S-, C-, and X-types are degenerate to a certain degree in albedo space. We have albedo measurements of 80,000 asteroids.

Phase Function Just as the Moon, other celestial bodies also go through different phases of illumination, depending on the relative position between the Sun, the Earth, and the celestial body. The phase function of an asteroid describes how its apparent magnitude changes from a full phase (zero degree phase angle between the Sun, the asteroid, and Earth) and a larger phase. The phase function is empirically modeled and can be described by two parameters, G1 and G2. Figure 1 (d) displays how the asteroid classes are distributed in this parameter space. G1 and G2 are correlated with albedos, hence we see some degeneracy. We have calculated these parameters for 50,000 objects.

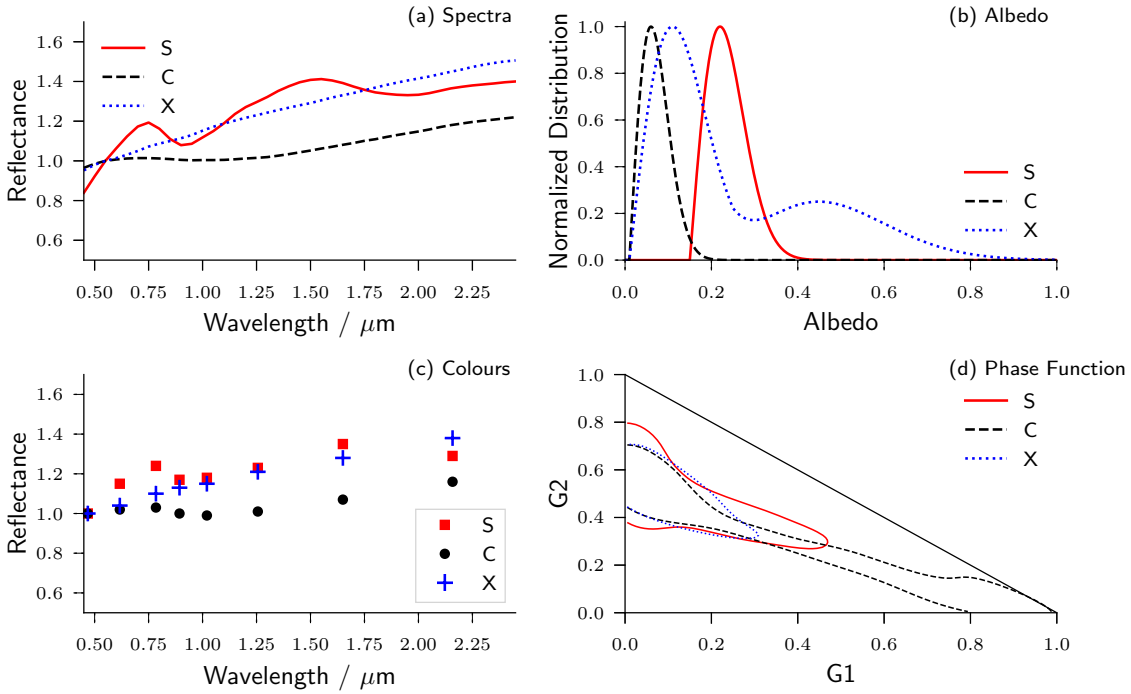


Figure 1: Observational features of three main types of asteroids

2.1 Data preparation

We build the input data by combining the four observational features outlined above. The input units are made up of observations of a single asteroid. As the spectra and colours far outweigh the albedo and phase function parameters in terms of chemical interpretability, we require for each input unit to contain at least a spectrum or a colour. In case an asteroid has been observed in the same spectral range or colour more than once, these observations are split into several rows. However, we do not repeat any datapoint, i.e. if an asteroid has multiple observations of a colour but only a single albedo measured, the albedo column is left partially filled. Nevertheless, we aim to create as many filled rows

as possible by merging the most complete feature sets for each asteroid.

Gaps in our data arise, e.g., when we have more colours of an asteroid than or albedo measurements. This is the general case, and in fact, most rows in the input dataframe are only partially filled. The missingness pattern is illustrated in Table 1.

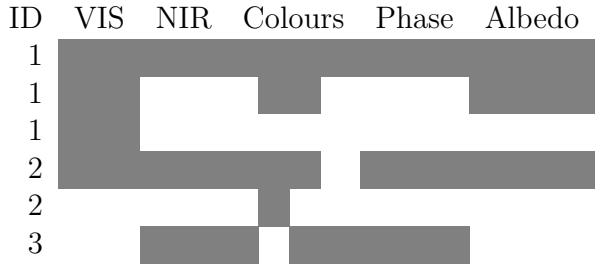


Table 1: Missingness pattern of the input data. Each ID refers to a unique asteroid.

While the availability of the features depends on the asteroid's size and distance to Earth, the actual value of the spectra, colours and phase function parameters are not biased by this dependence. We consider that these features are missing at random. For the albedo, we can immediately see that the availability of an observation is biased by the albedo value itself: Large albedos correspond to brighter asteroids, making an observation more likely. Values missing not a random rule out many imputation techniques, hence we apply a cut on the asteroids' brightness, removing the faintest ones from our sample. The remaining ones should be largely unbiased in the albedo.

An interesting question is how to provide the available measurement uncertainties to the imputation method. This point is particularly critical considering the wide range of precision affecting the available measurements.

3 Imputation

The dataset we collected contains a large fraction of the available observations on asteroids. Before attempting a clustering analysis, we combine the observations of single asteroids and fill in the gaps. For the latter, we currently use kNN imputation (Troyanskaya et al., 2001), a nonparametric technique able to capture nonlinear relationships between features.

3.1 Imputation strategy

The imputation is performed in two steps. Since asteroids do not produce their own emission but reflect the sunlight, we cannot compute their absolute reflectance values without additional information. The VIS, NIR, and VISNIR reflectance spectra are therefore relative values, typically normalized to unity at certain wavelengths. The input dataframe

contains VIS and NIR spectra which are not overlapping, hence we cannot use the same normalization wavelength for both.

As first step, we therefore train the imputer on the 2,000 rows containing all VISNIR and NIR spectra, and impute only the missing VIS part of the NIR spectra. Afterwards, we renormalize the VISNIR and NIR spectra to the same wavelength as the VIS spectra, before fitting the imputer on the whole 280,000 rows. Finally, we impute all remaining gaps using this imputer.

Currently, we are using 10 nearest-neighbours for imputation. This value will later be refined with a validation strategy.

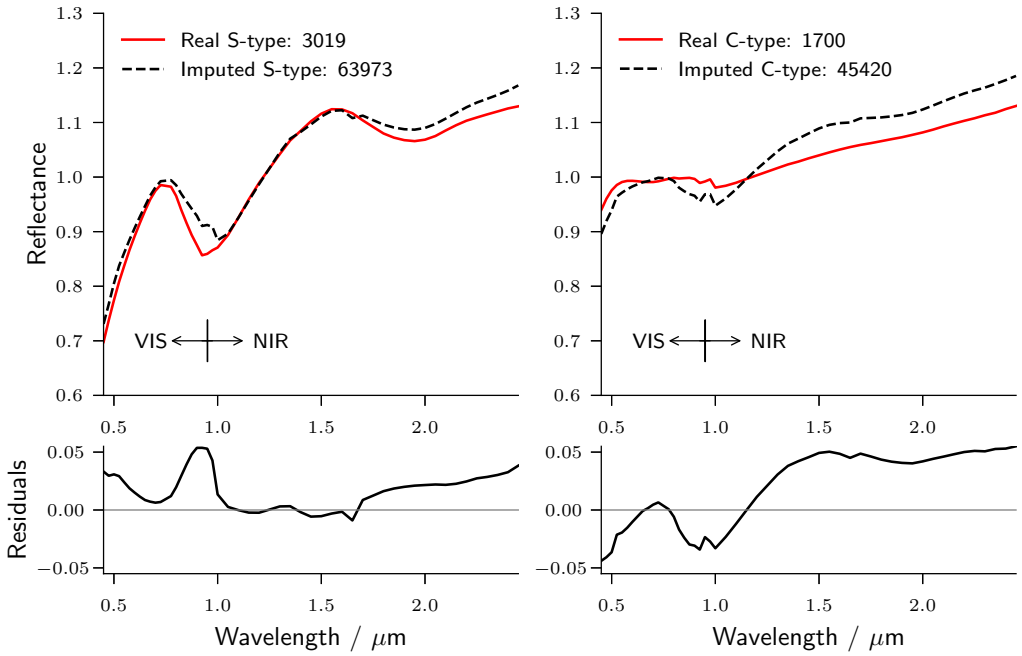


Figure 2: The mean reflectance of the real data (red) and the imputed data (black) at each wavelength for asteroids which have previously been classified as S-type (left) or C-type (right).

3.2 Results

As the reflectance spectra are pivotal for the classification, the imputed spectra are the key factor to estimate the quality of the imputation. Figure 2 depicts the mean reflectance at each wavelength of the real and the imputed data for asteroids which were previously classified as either S- or C-type. These two classes are the most represented in our dataset.

It is apparent from Figure 2 that the kNN-imputation works well in the VIS and the NIR parts of the spectra separately, however, the transition region around 1 μm is problematic. The band at 1 μm carries important information on the chemical composition, hence we need to improve in this area. For the C-type asteroids, we see that the imputed data resembles a flat S-type spectrum. The slope is close to the C-type asteroids, however, we observe spectral features at 1 μm and 2 μm which resemble the S-type. One factor contributing to this mixture is the heterogeneity of the input set of spectra in terms of wavelength coverage, resolution, and accuracy. Cleaning and homogenizing this data has been a major effort in the project and needs to be further improved.

We encounter similar results in the other dimensions shown above, which are not shown for brevity. In albedo-space, we find that the distribution of imputed values deviates from the underlying distribution, which resembles a Rayleigh-distribution. We are looking into providing a model to the imputer to improve this aspect.

4 Conclusion

Deriving the chemical compositions of asteroids requires an overview of the whole population in different feature spaces. Observational challenges cause gaps in the data, which we aim to impute using a kNN-approach. The first application on a large but sparsely-populated dataset shows that the imputation is overall successful but further data preparation is required to improve it on small scale errors.

Beyond imputation, clustering asteroids according to their chemical features remains our final goal. Therefore, an interesting avenue for future work would be to build a general model fit for *both imputation and clustering*. This could be possible using for example a mixture model.

Bibliography

Bus, S. J. and Binzel, R. P., (2002), Phase II of the Small Main-Belt Asteroid Spectroscopic Survey: The Observations, *Icarus*, 158, 106-145

Chapman, C., et al., (1971), A review of spectrophotometric studies of asteroids, *Physical Studies of Minor Planets*, 12, 51-65

DeMeo, F. E. and Carry, B., (2014), Solar System evolution from compositional mapping of the asteroid belt, *Nature*, 505, 629-634

Popescu, M. et al., (2016), Near-infrared colors of minor planets recovered from VISTA - VHS survey (MOVIS), *Astronomy & Astrophysics*, 591, A115

Ivezić, Z. et al., (2001), Solar System Objects Observed in the Sloan Digital Sky Survey Commissioning Data, *The Astronomical Journal*, 122, 2749-2784

Troyanskaya, O. et al., (2001), Missing value estimation methods for DNA microarrays, *Bioinformatics*, 17 (6), 520-525

Quelques tests de détection s'adaptant à la distribution du bruit de fond en astronomie

David Mary[1], Étienne Roquain[2], Marie Perrot-Dockes [2], Sophia Sulis[3] & Sébastien Bourguignon[4]

[1] *Université Côte d'Azur, Observatoire de la Côte d'Azur, CNRS, Laboratoire Lagrange, Boulevard de l'Observatoire, CS 34229, 06304, Nice cedex 4, France; david.mary@unice.fr*

[2] *Laboratoire de Probabilités, Statistique et Modélisation (LPSM, UMR 8001), Faculté des Sciences et Ingénierie, Tour 15/16, Étage 2, BC 158, 4, place Jussieu, 75252, PARIS cedex 05; etienne.roquain@upmc.fr, marie.perrocks@gmail.com*

[3] *Aix Marseille Université, CNRS, CNES, LAM, Marseille, France; Sophia.Sulis@lam.fr*

[4] *Laboratoire des Sciences du Numérique de Nantes (LS2N, UMR 6004), École Centrale de Nantes, 1 rue de la Noë, BP 92101, 44321 Nantes Cedex 3; Sebastien.Bourguignon@ec-nantes.fr*

Résumé. Dans cette note, nous considérons le problème de construction de tests de détection pour certains types de données issues de l'astrophysique. Une caractéristique commune des données considérées est que la distribution du bruit de fond est inconnue, ce qui invalide l'utilisation de nombreux tests classiques. Nous présentons des résultats récents proposant des solutions à ce problème pour deux applications spécifiques, la détection d'exoplanètes et celle de galaxies.

Mots-clés. détection, tests multiples, exoplanètes, galaxies.

Abstract.

In this note, we consider the problem of building detection tests for some types of astrophysical data. The distribution of the background noise is unknown in the considered cases, preventing from using most classical procedures. We present works that address this problem for two specific applications : the detection of exoplanets and of galaxies.

Keywords. detection, multiple testing, exoplanets, galaxies.

1 Problématique et contexte

La détection de sources est un enjeu majeur dans plusieurs domaines de l'astrophysique. Le contexte moderne d'instruments toujours plus complexes produisant des données toujours plus volumineuses conduit souvent à réaliser un nombre gigantesque de tests simultanément. La thématique des tests multiples permet de prendre en compte cette multiplicité de manière appropriée. Si ce domaine de recherche possède des origines très

anciennes en statistique, les deux dernières décennies y ont vu une explosion de travaux théoriques et appliqués : parmi les procédures qui ont vu le jour, on peut citer notamment la procédure de Benjamini-Hochberg [1] qui contrôle le *false discovery rate* (FDR), ainsi que les différentes versions du *higher criticism* (HC) [2] et des tests de Berk-Jones (BJ) [3] pour le cas de signaux rares et faibles [2].

En astrophysique, ce genre de procédures reste cependant assez peu utilisé. Une des raisons est que la distribution des données du bruit de fond (c'est-à-dire lorsque qu'il n'y a pas de signal astrophysique) est très souvent inconnue. Ceci pousse l'utilisateur à se tourner vers des procédures plus *ad hoc*, ce qui peut s'avérer dangereux. Un exemple récent est le cas de la détection d'une exoplanète près de l'étoile α Centauri Bb [4]. La détection de cette exoplanète, annoncée en 2012 sur la base d'une P-valeur évaluée à 0.02% [4], a été fortement remise en cause depuis [5, 6]. En effet, ces dernières analyses ont mis en lumière des effets mal pris en compte, en particulier le signal parasite émis par l'étoile elle-même (le "bruit stellaire"). Ces effets, très difficiles à contrôler, impactent fortement les taux d'erreur prédits par les tests qui les ignorent.

L'objet de cette note est de présenter des procédures de détection avec un risque correctement contrôlé même lorsque la distribution des statistiques de test sous la distribution nulle est mal connue. Nous considérons deux cas concrets : la détection d'exoplanètes par vélocimétrie radiale [7] (Section 2), et la détection de galaxies dans des images multi-longueurs d'onde de l'instrument MUSE [8, 9] (Section 3).

2 Détection d'exoplanètes par vitesses radiales

La présence d'une planète orbitant autour d'une étoile induit un mouvement de l'étoile autour du barycentre de masse du système étoile-planète. Ce mouvement module de façon quasi-périodique la vitesse radiale de l'étoile par rapport à un observateur terrestre et s'imprime par effet Doppler sur la lumière de l'étoile. En mesurant le décalage Doppler des raies du spectre stellaire en fonction du temps, on déduit ainsi la vitesse radiale de l'étoile. Les données se présentent donc sous la forme d'une série temporelle obtenue durant une fenêtre temporelle d'observation.

Le but est de tester, dans une telle série de vitesses radiales, H_0 : la moyenne est nulle (il n'y a pas d'exoplanète) contre H_1 : la moyenne est un signal quasi-périodique (il y a une ou plusieurs exoplanètes). La série temporelle est supposée stationnaire, mais sa matrice de covariance Σ est inconnue en raison du bruit stellaire. Pour s'adapter à l'alternative, une approche classique en échantillonnage régulier consiste à construire le périodogramme des données (module carré de la transformée de Fourier discrète de la série), puis à rechercher si ses composantes indiquent de façon significative la présence de signaux périodiques. Cependant, la distribution de ce périodogramme est inconnue sous H_0 car Σ est inconnue.

Pour contourner cette difficulté nous proposons ici une approche exogène, rendue pos-

sible par la disponibilité d'un certain nombre (L) de séries simulées sous H_0. En effet, il existe à l'heure actuelle des codes astrophysiques permettant de simuler de façon réaliste les vitesses radiales qui seraient observées pour un type d'étoile donné. Ceci permet, au moins sur une certaine plage de fréquences, de générer des séries exogènes sous H_0. Celles-ci sont cependant en nombre très limité (quelques unités à une dizaine) en raison du fort coût de calcul que ces simulations nécessitent. A l'aide de ces séries, nous proposons de construire un périodogramme de référence, utilisé pour normaliser le périodogramme des données.

Notons $Z_1, \ldots, Z_N$ un sous-ensemble de composantes du périodogramme ainsi standardisé. Si l'échantillonnage est régulier, alors les distributions considérées possèdent des expressions analytiques [10]. Ainsi, sous certaines conditions, on peut montrer que $Z_1, \ldots, Z_N$ sont asymptotiquement indépendants avec $Z_j \sim \mathcal{F}(2, 2L)$ (loi de Fisher de paramètres 2 et $2L$) sous l'hypothèse nulle et $Z_j \sim \mathcal{F}_{\lambda_j}(2, 2L)$ (loi de Fisher decentrée avec un certain de paramètre de décentrage λ_j et de paramètres 2 et $2L$) sous l'alternative. Soulignons que sous H_0, la distribution du périodogrammee standardisé est indépendante du paramètre de nuisance Σ. Cette propriété, complètée par celle de l'indépendance asymptotique, permet de construire des tests globaux de niveau α à partir des Z_j. Par exemple, une façon naturelle et explicite d'agréger ces tests est simplement le test du maximum. Dans le cas considéré, celui-ci rejette l'hypothèse nulle si $\max(Z_j, 1 \leq j \leq N)$ dépasse le seuil c_α, calibré de sorte que

$$1 - \Big(1 - \Big(\frac{L}{c_\alpha + L}\Big)^L\Big)^N = \alpha.$$

Des résultats similaires peuvent être obtenus pour de nombreux autres tests comme celui de Fisher [11], ses variantes [12, 13, 14] et ceci permet également d'utiliser des tests de détection de type HC et BJ.

Sur des simulations, on voit que l'approximation asymptotique est bonne quand la longueur de la série est suffisamment grande devant la durée caractéristique de corrélation du signal (typiquement un à deux ordres de grandeur supérieur), ce qui confirme que le test conduit est bien de niveau α. De plus, les modèles analytiques obtenus permettent d'étudier la puissance asymptotique de cette procédure de test. Là aussi, les simulations montrent que ces expressions sont valables pour des valeurs de N modérées . Ces résultats permettent de faire des études de détectabilité pour des planètes ou des instruments avec des caractéristiques données (cf Figure 1 pour un exemple avec le test du maximum).

Par ailleurs, dans le cas de l'échantillonnage irrégulier, l'hypothèse d'indépendance entre les Z_j ne peut être tenue et nous avons proposé dans [15] des techniques de *bootstrap* pour approcher la distribution de la statistique de test sous H_0.

3 Détection de galaxies dans les données MUSE

Le spectrographe intégral de champ MUSE installé sur un des télescopes de 8m au Very Large Telescope (Chili) permet d'obtenir des images multi-longueurs d'onde (typiquement 300×300 images dans 3600 canaux optiques). On cherche à détecter dans ce "cube" de données des galaxies très lointaines, qui se manifestent comme une surbrillance locale du flux (une raie en émission) dans une poignée de voxels. On connaît à peu près la forme de ces raies mais ni leurs hauteurs, ni leur nombre (quelques dizaines à quelques centaines typiquement) ni leurs positions dans le cube. Par ailleurs, la hauteur de certaines raies peut être beaucoup plus faible que le niveau du bruit de fond et que celui d'autres sources brillantes et étendues (dûes à d'autres étoiles et galaxies ou à des artefacts instrumentaux; on regroupe ces sources sous le terme de signaux de nuisance).

Dans ce cadre, nous avons considéré une approche de détection en deux temps : les signaux de nuisance sont d'abord supprimés et l'étape de détection des galaxies se fait ensuite, dans le résidu. Pour s'adapter à l'alternative, l'approche considère les maximas locaux du cube de données résiduel, après suppression des nuisances et divers filtrages. En raison de ces prétraitements, la distribution sous H_0 des maxima locaux n'est pas connue. Le problème considéré ici est plus complexe que le précédent puisqu'il y a plusieurs hypothèses nulles, chacune liée à un maximum local [16]. Si nous notons x, y, z la position d'un maximum local, on teste $H_{0,x,y,z}$: il n'y a pas de raie d'émission à la position (x, y, z), contre $H_{1,x,y,z}$: il y en a une à cette position. Le critère d'erreur considéré est celui du FDR, moyenne du taux de fausses découvertes parmi les positions déclarées comme correspondant à une galaxie.

L'approche que nous proposons pour ce problème est endogène. Nous montrons par simulations que sous certaines hypothèses, la distribution sous H_0 peut être estimée à partir des données elles-mêmes et que la procédure de détection globale (incluant la suppression des nuisances et les divers filtrages) contrôle le FDR. Nous appliquons cette approche à un cube de données issu de l'instrument MUSE et comparons avec les résultats du télescope spatial Hubble [9].

4 Perspectives

Dans la communauté statistique mathématique, la validation de procédures de test apprennant automatiquement la distribution nulle remonte aux procédures de test par permutation ou par randomisation. Dans le cadre de tests multiples et du FDR, certaines justifications ont été apportées récemment, voir par exemple [17, 18, 19]. Dans certaines des études considérées ici, la validation des approches est faite par des simulations numériques, qui nécessitent un choix particulier (et donc nécessairement un peu arbitraire) du jeu de paramètres. Une direction de recherche de ces travaux est d'obtenir des garanties théoriques en simplifiant éventuellement les procédures et/ou les modèles astrophysiques tout en s'inspirant des résultats théoriques récents.

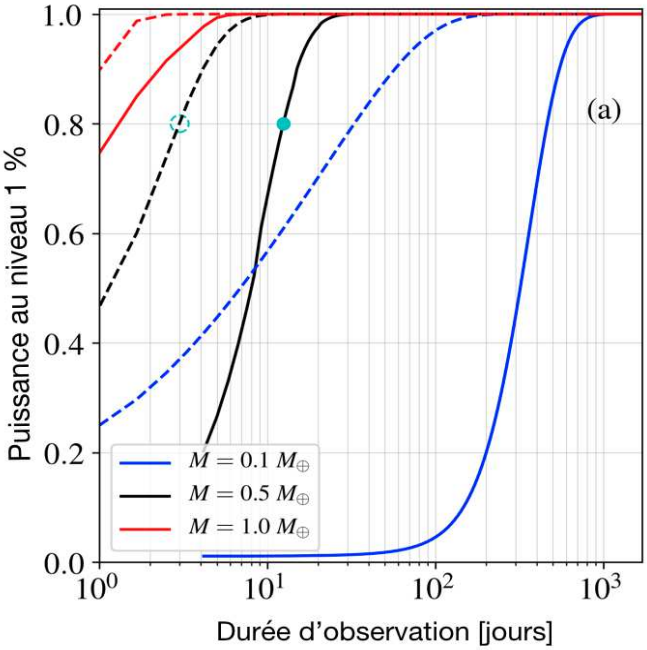


Figure 1: Puissance du test de détection au niveau 1% en fonction du temps pour une planète orbitant circulairement une étoile de type solaire en 17,5 h. La masse de l'exoplanète est de $0,1$ (bleu), 0,5 (noir) et 1 (rouge) masse terrestre. Le pas d'échantillonnage des données est de 2h et $L = 20$. Les courbes pleines correspondent au cas où le bruit généré par l'étoile est corrélé (cas réaliste) et les courbes en pointillés au cas où il est blanc. L'écart-type du bruit est de 49 cm.s^{-1} et l'amplitude du signal planétaire de quelques centimètres par seconde. La détection est plus difficile dans le cas d'un bruit coloré (courbes pleines) car la période de la planète correspond à une zone de fréquences où le bruit de fond créé par l'étoile est très énergétique. Dans cette configuration, la détection d'une exoplanète d'une demi masse terreste serait possible avec une probabilité de 80% en 12,4 jours; pour un bruit de même puissance mais blanc, cette durée tomberait à 3 jours (disque et cercle bleus).

References

[1] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B*, 57(1):289–300, 1995.

[2] D. Donoho and J. Jin. Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Stat.*, 2004.

[3] V. Gontscharuk et al. The intermediates take it all: Asymptotics of higher criticism statistics and a powerful alternative based on equal local levels. *Biom. J.*, 57(1):159–180, 2014.

[4] X. Dumusque et al. An Earth-mass planet orbiting α Centauri B. *Nature*, 491:207–211, 2012.

[5] A. Hatzes. The Radial Velocity Detection of Earth-mass Planets in the Presence of Activity Noise: The Case of α Centauri Bb. *ApJ*, 770:133, 2013.

[6] V. Rajpaul et al. Ghost in the time series: no planet for Alpha Cen B. *MNRAS*, 456:L6–L10, 2016.

[7] Sophia Sulis. *Statistical methods using hydrodynamic simulations of stellar atmospheres for detecting exoplanets in radial velocity data*. Theses, Université Côte d'Azur, October 2017.

[8] R. et al. Bacon. The muse hubble ultra deep field survey - i. survey description, data reduction, and source detection. *A&A*, 608:A1, 2017.

[9] D. Mary, R. Bacon, S. Conseil, L. Piqueras, and A. Schutz. Origin: Blind detection of faint emission line galaxies in muse datacubes. *Astronomy Astrophysics*, Jan 2020.

[10] S. Sulis, D. Mary, and L. Bigot. A study of periodograms standardized using training datasets and application to exoplanet detection. *IEEE Transactions on Signal Processing*, 65(8):2136–2150, April 2017.

[11] R.A. Fisher. Tests of Significance in Harmonic Analysis. *Proc. R. Soc. London, Ser. A*, 125:54–59, 1929.

[12] S.T. Chiu. Detecting periodic components in a white gaussian time series. *J. R. Stat. Soc. Series B*, 51(2):249–259, 1989.

[13] M. Shimshoni. On fisher's test of significance in harmonic analysis. *Geophys. J. R. Astronom. Soc.*, pages 373–377, 1971.

[14] E. Bölviken. New tests of significance in periodogram analysis. *Scandinavian J. Stat.*, 10(1):1–9, 1983.

[15] S. Sulis, D. Mary, and Lionel Bigot. A bootstrap method for sinusoid detection in colored noise and uneven sampling. application to exoplanet detection. In *EUSIPCO 2017, Kos, Greece, August 28 - September 2, 2017*, pages 1095–1099. IEEE, 2017.

[16] Dan Cheng and Armin Schwartzman. Multiple testing of local maxima for detection of peaks in random fields. *Ann. Statist.*, 45(2):529–556, 04 2017.

[17] Rina Foygel Barber and Emmanuel J. Candès. Controlling the false discovery rate via knockoffs. *Ann. Statist.*, 43(5):2055–2085, 2015.

[18] Ery Arias-Castro and Shiyun Chen. Distribution-free multiple testing. *Electron. J. Stat.*, 11(1):1983–2001, 2017.

[19] Etienne Roquain and Nicolas Verzelen. On using empirical null distributions in Benjamini-Hochberg procedure. *arXiv e-prints*, page arXiv:1912.03109, Dec 2019.

Classification de patterns de catégorisation chez l'humain par deux modèles d'apprentissage

Giulia Mezzadri [1] & Patricia Reynaud-Bouret [1] & Thomas Laloë [1] & Fabien Mathy [2]

[1] *Laboratoire J.A. Dieudonné, CNRS UMR 7351, Nice*
giulia.mezzadri@univ-cotedazur.fr
patricia.reynaud-bouret@univ-cotedazur.fr
thomas.laloe@univ-cotedazur.fr
[2] *Bases, Corpus, Langage, CNRS UMR 7320, Nice*
fabien.mathy@univ-cotedazur.fr

Résumé. La catégorisation est un processus cognitif qui permet de regrouper des objets ayant une ou plusieurs caractéristiques en commun. Un des problèmes majeurs dans ce domaine est la multitude des modèles cognitifs qui ont comme but de reproduire les performances de participants humains dans une tâche de catégorisation. Nous nous proposons comme objectif de comparer deux modèles influents en psychologie : ALCOVE et Component-Cue, tous deux fondés sur une structure de réseaux de neurones artificiels. Après une présentation de ces deux modèles d'apprentissage, nous détaillerons la méthode que nous avons utilisée pour déterminer lequel de ces modèles s'ajuste au mieux à un jeu de données expérimentales. Nous montrerons enfin que la moitié des participants suit une logique plus proche d'ALCOVE, tandis que l'autre moitié suit une logique plus proche de Component-Cue.

Mots-clés. Cognition, Neurosciences, apprentissage de catégories, ALCOVE, réseau de neurones artificiels, Component-Cue, estimation par maximum de vraisemblance, méthode des moindres carrée.

Abstract. Categorization is a cognitive process that allows the grouping of objects with one or more common features. One of the major problems in this field is the multitude of cognitive models which aim to reproduce the performances of human participants in a categorization task. We aim to compare two influential models in psychology : ALCOVE and Component-Cue, both based on a structure of artificial neural networks. After a presentation of these two learning models, we will present the method we used to determine which of these models best fits an experimental dataset. We will finally show that half of the participants follow a logic closer to ALCOVE, while the other half follows a logic closer to Component-Cue.

Keywords. Cognition, neuroscience, categorization learning, ALCOVE, Component-Cue, artificial neural network, maximum likelihood estimation (MLE), Sum of squared differences (SSD).

1 Introduction

En observant le ciel, nous étiquetons immédiatement le temps comme étant beau ou mauvais ; une musique peut-être classée comme un morceau de jazz, folk, rap, ou autres styles ; en lisant un texte nous l'identifions comme un article de journal, un roman, une poésie ou autre encore. La catégorisation s'intéresse à l'ensemble de ce type de processus cognitifs qui permettent à un apprenant, qu'il soit un humain, un animal ou un algorithme, de placer des objets similaires dans un même groupe.

Un des problèmes majeurs dans ce domaine est la multitude et la variété des modèles qui ont comme but de reproduire les performances de participants humains dans une tâche de catégorisation. Un objectif est donc celui de comparer les modèles de catégorisation les plus influents en psychologie cognitive. Avant de se plonger dans cette analyse, il est important de comprendre en quoi consiste une expérience de catégorisation.

Dans une expérience de catégorisation, l'expérimentateur choisi un ensemble d'objets qu'il place arbitrairement dans deux catégories. L'objectif du participant est de comprendre et d'assimiler la règle de catégorisation choisie à travers un apprentissage par essai-erreur. La tâche de catégorisation est divisée en deux phases : une première phase, appelée phase d'apprentissage, dans laquelle le participant est censé apprendre la règle de catégorisation choisie par l'expérimentateur ; et une deuxième phase, appelée phase de transfert, dans laquelle l'introduction de nouveaux objets permet de tester la généralisabilité de l'apprentissage du participant.

Étant donné la différence de nature des phases d'apprentissage et de transfert, il y a des modèles, dits de transfert, qui sont plus adaptés à la phase de transfert, et d'autres, dits d'apprentissage, qui sont plus adaptés à la phase d'apprentissage. Les modèles de transfert sont utiles pour reproduire les performances finales des participants (celles obtenues à la fin du processus d'apprentissage), tandis que les modèles d'apprentissage peuvent rendre compte à la fois de la dynamique d'apprentissage et de la phase de transfert. Ici, nous nous focalisons sur deux modèles d'apprentissage qui ont montré dans le passé leur qualité d'ajustement aux données humaines (voir [3] et [4]) : ALCOVE et Component-Cue. Ces modèles sont tous deux fondés sur des réseaux de neurones et utilisent des conditions de similarité.

Après une présentation des spécificités des deux modèles d'apprentissage considérés, nous présenterons la méthode utilisée pour sélectionner le modèle qui s'adapte au mieux aux données. Nous montrerons enfin les résultats obtenus sur données simulées et expérimentales.

2 ALCOVE et Component-Cue

ALCOVE et Component-Cue sont deux modèles d'apprentissage basés sur une structure de réseau de neurones artificiels (Figure 1). Cette structure, conceptualisée pour reproduire le fonctionnement de l'esprit humain dans une tâche de catégorisation, est

constituée de trois parties : un premier nœud qui reçoit le stimulus (input), des nœuds centraux qui élaborent l'information, et une dernière couche de nœuds (outputs) qui est reliée aux précédents à travers des poids (spécifiques à chaque catégorie) et qui produit une sortie pour chacune des catégories possibles. À chaque fois que le réseau reçoit un stimulus, les nœuds s'allument selon une logique propre au modèle et la somme pondérée de tous les nœuds produit les sorties des catégorie considérées. Les poids qui relient les

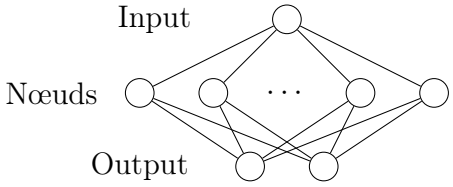


FIGURE 1 – Structure de réseau de neurones commune à ALCOVE et Component-Cue.

nœuds centraux à ceux de sortie sont mis à jour après chaque réception d'un stimulus. Pour mettre à jour les poids, les modèles utilisent un algorithme de descente de gradient qui cherche à minimiser l'écart entre la réponse du participant et les sorties du réseau. En particulier, cette étape rend le processus d'apprentissage dépendant de tout le passé, en augmentant ainsi la difficulté à analyser statistiquement ce type de modèles.

Les nœuds centraux d'ALCOVE et de Component-Cue élaborent l'information différemment. Dans ALCOVE chaque nœud code un objet de l'expérience et son activation dépend de la similarité entre l'objet associé au nœud et le stimulus d'input. Au contraire, dans Component-Cue chaque nœud code une des caractéristiques possibles de l'objet et son activation se base sur la présence ou l'absence de la caractéristique considérée dans le stimulus d'input. Si on dénote A et B les deux catégories dans lesquelles nous pouvons classifier les objets, alors pour les deux modèles la sortie associée à la catégorie A au temps t pour l'objet x est la suivante :

$$O_A^{(t)}(x) = \sum_{i=1}^{n} a_i(x) \cdot w_{iA}^{(t)},$$

où n représente le nombre de nœuds, $a_i(x)$ dénote l'activation du i-ème nœud suite à la réception de l'input x, et $w_{iA}^{(t)}$ représente le poids au temps t qui relie le i-ème nœud au nœud de sortie de la catégorie A. Au contraire, la probabilité de classifier un objet dans une des deux catégories revêt une forme différente dans les deux modèles. Pour ALCOVE, la probabilité de catégoriser l'objet x dans la catégorie A est donnée par :

$$\mathbb{P}(A \,|\, x, \mathcal{H}_{t-1}) = \frac{O_A^{(t)}(x) + b}{O_A^{(t)}(x) + O_B^{(t)}(x) + 2b},$$

tandis que pour Component-Cue, elle est donnée par :

$$\mathbb{P}(A \,|\, x, \mathcal{H}_{t-1}) = \frac{e^{\phi \cdot O_A^{(t)}(x)}}{e^{\phi \cdot O_A^{(t)}(x)} + e^{\phi \cdot O_B^{(t)}(x)}},$$

où b et ϕ sont deux paramètres fixes à estimer, et $\mathcal{H}_{t-1}$ représente le passé du processus d'apprentissage jusqu'au temps $t-1$.

Il est important de souligner que ALCOVE et Component-Cue peuvent aussi être utilisés sur des données provenant d'une phase de transfert. Pour ce faire, il suffit de fixer les poids du réseau de neurones. Pour citer un exemple, dans l'article [3], ALCOVE et Component-Cue ont été utilisés dans leur forme dynamique de réseau de neurones sur des données d'apprentissage, et ensuite dans leur forme statique sur des données de transfert.

3 Comment compare-t-on deux modèles d'apprentissage ?

Pour déterminer le modèle qui s'ajuste au mieux aux données, nous exploitons la division en deux phases, d'apprentissage et de transfert, caractéristique des expériences de catégorisation. La phase d'apprentissage est utilisée pour l'estimation des paramètres du modèle considéré, tandis que la phase de transfert est utilisée pour calculer l'erreur entre les réponses du participant et les prévisions du modèle. La phase d'apprentissage permet d'estimer les paramètres, tandis que la phase de transfert constitue notre test. L'estimation des paramètres du modèle est faite par maximum de vraisemblance (maximum likelihood estimation, MLE) :

$$\hat{\theta} \in \arg\min_{\theta} \mathcal{L}(\theta \,;\, \mathcal{D}_A),$$

où $\mathcal{L}$ représente la vraisemblance du modèle considéré et $\mathcal{D}_A$ représente les données en phase d'apprentissage. Au contraire, le calcul de l'erreur entre les prévisions et les réponses des participants est fait en utilisant la méthode des moindres carrées (sum of squared differences, SSD) :

$$\text{SSD} = \sum_{i \in \mathcal{D}_T} \left(p_i^{\hat{\theta}} - r_i \right)^2,$$

où, $p_i^{\hat{\theta}}$ dénote la probabilité donnée par le modèle de classifier le stimulus i dans la première catégorie (cette probabilité est obtenue en utilisant le vecteur des paramètres estimés $\hat{\theta}$), r_i dénote la réponse du participant au temps i et $\mathcal{D}_T$ dénote les données en phase de transfert. L'utilisation du calcul de la SSD comme critère pour déterminer le modèle qui s'ajuste au mieux aux données est motivé par son utilisation en psychologie (voir [3] et [4]).

4 Application aux données simulées

Avant d'appliquer la méthode décrite ci-dessus aux données expérimentales, nous devons nous assurer que, quand nous l'appliquons aux données simulées, nous sommes bien capables de retrouver le modèle avec lequel nous avons créé les données. Pour ce faire, nous simulons des données artificielles avec un modèle donné et nous appliquons la méthode illustrée dans le paragraphe précédent afin d'établir, parmi tous les modèles, celui qui s'ajuste au mieux aux données simulées. Si nous retrouvons bien le modèle avec lequel nous avons simulées les données artificielles, nous pouvons affirmer que la méthode permet de discriminer les modèles.

Pour simuler les données artificielles nous avons utilisé les paramètres trouvés à travers la procédure d'estimation (des paramètres) sur données réelles. Les résultats en Figure 2 (à gauche) montrent que, si ALCOVE a la plus petite SSD, alors dans 90% des cas les données ont été obtenues en utilisant ALCOVE. Au contraire, si Component-Cue a la plus petite SSD, alors dans 84.8% des cas les données proviennent de Component-Cue. Ces résultats nous permettent d'avoir un regard confiant vis-à-vis des conclusions que nous obtiendrons sur données réelles.

5 Application aux données réelles

Nous pouvons enfin appliquer cette méthode aux données expérimentales recueillies par Mathy [5] pour établir quel modèle, parmi ALCOVE et Component-Cue, se rapproche au mieux aux réponses de chaque individu. Le résultat (Figure 2 à droite), montre que la moitié des participants suit une logique plus proche d'ALCOVE, tandis que l'autre moitié suit une logique plus proche de Component-Cue. Étant donné la fiabilité de l'analyse sur données artificielles, nous pouvons affirmer qu'il y a deux types d'individus : une classe d'individus qui a un raisonnement plus proche d'ALCOVE et une autre qui utilise un raisonnement plus proche de Component-Cue.

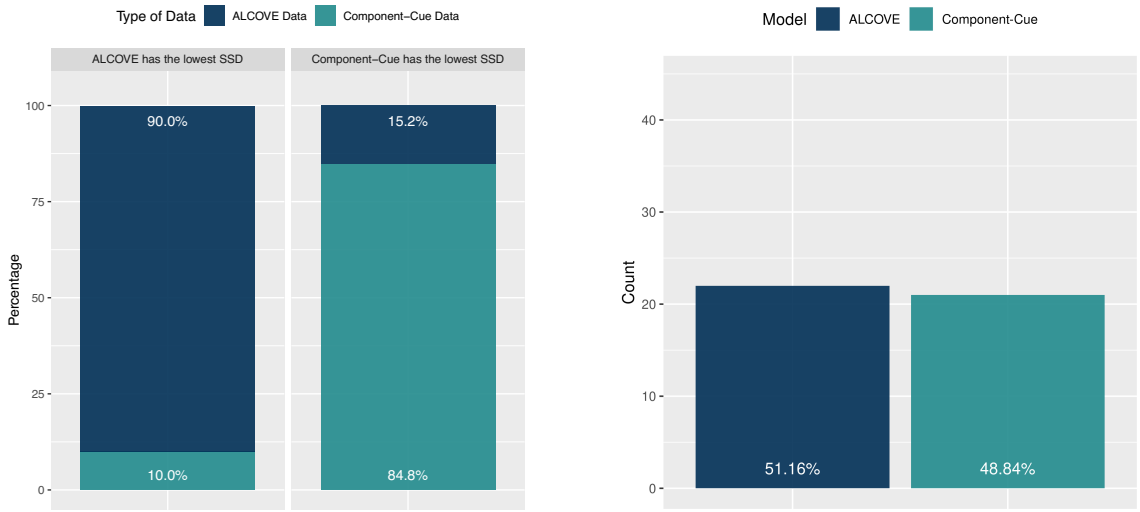


FIGURE 2 – Comparaison entre ALCOVE et Component-Cue. À gauche, sur données simulées, le pourcentage de données simulées provenant d'ALCOVE ou de Component-Cue, en fonction du modèle ayant la plus petite SSD. À droite, sur données réelles, le nombre et le pourcentage d'individus dont les données s'ajoute au mieux à ALCOVE ou à Component-Cue.

Références bibliographiques

[1] Kruschke, J. K. *ALCOVE : An exemplar-based connectionist model of category learning*. Psychological Review, 1992.

[2] Gluck M. A. & Bower G. H. *From conditioning to category learning : An adaptive network model*. Journal of Experimental Psychology : General, 1988.

[3] Nosofsky, R. M. & Kruschke J. K. & McKinley S. C. *Combining Exemplar-Based Category Representations and Connectionist Learning Rules*. Journal of Experimental Psychology : Learning, Memory and Cognition, 1992.

[4] Nosofsky, R. M. & Gluck, M. A. & Palmeri T. J. & McKinley S. C. & Glauthier P. *Comparing modes of rule-based classification learning : A replication and extension of Shepard, Hovland, and Jenkins (1961)*. Memory & Cognition, 1994.

[5] Mathy, F., & Feldman, J. *The influence of présentation order on category transfer*. Experimental Psychology, 2016.

Prise en compte d'un acteur manquant dans l'inférence de réseaux d'interactions d'espèces par mélange d'arbres à partir de données de comptages

Raphaëlle Momal[1]*, Stéphane Robin[1], Christophe Ambroise[2]

1: Université Paris-Saclay, AgroParisTech, INRAE, UMR MIA-Paris, Paris, France.
2: Laboratoire de Mathématiques et Modélisation d'Évry, 23 bvd de France, Évry, France

Résumé

L'inférence de réseau est utilisée dans de nombreux domaines tels que la génomique ou l'écologie pour déduire la structure d'indépendance conditionnelle entre les variables, à partir des mesures de l'expression des gènes ou de l'abondance des espèces par exemple. Dans de nombreux cas pratiques, il est probable que toutes les variables impliquées dans le réseau n'ont pas été réellement observées. Les échantillons observés sont donc tirés d'une distribution où certaines variables non observées ont été marginalisées.

Nous introduisons un modèle statistique générique pour l'inférence de réseau à partir de données d'abondance avec acteur manquant. Le modèle comprend des effets fixes pour la prise en compte des covariables environnementales et des efforts d'échantillonnage, et des effets aléatoires corrélés pour coder les interactions des espèces. La structure de corrélation est celle d'un modèle graphique gaussien marginalisé sur une ou plusieurs variables, correspondant aux acteurs manquants. Le réseau inféré est obtenu comme une moyenne sur tous les arbres couvrants, d'une manière efficace sur le plan des calculs.

Mots-clés : acteur manquant, algorithme EM variationnel, données d'abondance, inférence de réseaux, modèles graphiques, modèle Poisson log-Normal, théorème arbre-matrice

Abstract

Network inference is used in many areas such as genomics or ecology to infer the structure of conditional independence between covariates, based on the measures of gene expression or species abundance for example. In many experiments, it is likely that not all covariates involved in the network were actually observed. Then observed samples are drawn from a distribution where some unobserved covariates were marginalized.

We introduce a generic staitstical model for network inference from abundance data with missing actor. The model includes fixed effects to take account of environmental covariates and sampling efforts, as well as correlated random effects to encode species interactions. The correlation structure is that of a gaussian graphical model marginalized on one ore more covariates, corresponding to the missing actors. The inferred network is obtained by averaging on all spanning trees, in a computationnally efficient way.

Key-words: graphical models, network inference, missing actor, abundance data, Variational EM algorithm, matrix tree theorem, Poisson log-Normal model

*Electronic address: `raphaelle.momal@agroparistech.fr`; Corresponding author

Introduction

L'inférence de réseau est utilisée dans de nombreux domaines tels que la génomique ou l'écologie pour déduire la structure d'indépendance conditionnelle entre les variables, à partir des mesures de l'expression des gènes ou de l'abondance des espèces par exemple. Cette inférence repose sur la modélisation de la distribution jointe des différentes expressions ou abondances, pour laquelle les modèles graphiques fournissent un cadre naturel et bien étudié. Ce cadre permet notamment de distinguer les liens "directs" reliant des variables dépendantes conditionnellement à toutes les autres, des liens "indirects" entre des variables liées, par exemple, à une même troisième. La prise en compte de covariables décrivant les conditions environnementales ou expérimentales permet également d'éviter l'inférence de liens entre des variables répondant conjointement à des fluctuations de ces conditions.
Les modèles graphiques gaussiens (GGM) sont particulièrement populaires mais ne s'adaptent pas aux données de comptages comme les données d'abondances d'espèces ou les mesures d'expression de gènes obtenues par séquençage. Dans ce cas, une modélisation classique consiste à introduire une couche latente gaussienne, conditionnellement à laquelle les données observées sont distribuées selon une loi pertinente pour des comptages comme la loi de Poisson. Le modèle Poisson-logNormal (Aitchison and Ho, 1989) entre dans cette catégorie. Son inférence pose des difficultés liées à la loi conditionnelle des variables latentes sachant les variables observées, dont la complexité peut être contournée aux moyens d'approximations variationnelles (Blei et al., 2017). Le recours à une couche latente gaussienne ouvre l'accès à toute une série de méthodes disponibles, comme l'inférence de réseaux par GGM (Chiquet et al., 2019). Par ailleurs l'utilisation d'un mélange d'arbre couvrants comme structure de dépendance de la couche latente permet d'obtenir des probabilités pour chaque arête (Momal et al., 2020).
Cependant, dans de nombreux cas pratiques, il est probable que toutes les variables impliquées dans le réseau n'ont pas été réellement observées. Les échantillons observés sont donc tirés d'une distribution où certaines variables non observées, appelées ici 'acteurs', ont été marginalisées, comme l'illustre la Figure 1a) ci-dessous.
L'objectif de ce travail est d'introduire un cadre statistique permettant d'inférer l'existence et la position dans le réseau d'acteurs manquants. Des travaux analogues existent dans le cadre gaussien (Lauritzen and Meinshausen, 2012; Robin et al., 2019) mais l'extension au cas de données de comptages est, à notre connaissance, nouvelle.

Figure 1: *a*) Exemple de la marginalisation d'une variable x non observée dans un graphe à 4 variables. *b*) Modèle graphique reliant les données $\boldsymbol{Y}$, la couche latente de paramètres gaussiens $\boldsymbol{Z} = (\boldsymbol{Z}_O, \boldsymbol{Z}_H)$, ainsi que l'arbre latent T.

1 Modélisation

Modèle Poisson-logNormal. Nous commençons par rappeler le modèle PLN en prenant pour exemple le cas de données d'abondance. On considère p espèces observées sur n sites. Les abondances par site sont décrites dans la matrice $n \times p$ $\boldsymbol{Y}$ où Y_{ij} est l'abondance de l'espèce j dans le site i, et $\boldsymbol{Y}_i$ est le vecteur d'abondances collecté au site i (i ième ligne de $\boldsymbol{Y}$). Un vecteur de covariables $\boldsymbol{x}_i$ de taille d est mesuré dans chaque site i et toutes les covariables sont

rassemblées dans la matrice $n \times d$ $\boldsymbol{X}$. Le modèle Poisson-logNormal (PLN) prévoit qu'un vecteur gaussien (latent) $\boldsymbol{Z}_i$ de dimension p est associé à chaque site:

$$\{\boldsymbol{Z}_i\}_{1 \leq i \leq n} \text{ iid}, \qquad \boldsymbol{Z}_1 \sim \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Omega}^{-1}), \tag{1}$$

les sites étant donc supposés indépendants. On réunit l'ensemble des vecteur latents $\boldsymbol{Z}_i$ dans une matrice $n \times p$ notée $\boldsymbol{Z}$. Le modèle PLN suppose ensuite que les abondances des différentes espèces dans chaque site sont conditionnellement indépendantes et que leur distribution dépend de l'environnement et de la variable latente associée:

$$\{Y_{ij}\}_{1 \leq i \leq n, 1 \leq j \leq p} \mid \boldsymbol{Z} \text{ iid}, \qquad Y_{ij} \mid \boldsymbol{Z} \sim \mathcal{P}\left(\exp(o_{ij} + \boldsymbol{x}_i^\intercal \boldsymbol{\theta}_j + Z_{ij})\right), \tag{2}$$

où o_{ij} est un terme d'*offset* connu qui rend compte de l'effort d'échantillonnage. Le vecteur $d \times 1$ de coefficients de régression $\boldsymbol{\theta}_j$ décrit l'effet de l'environnement $\boldsymbol{x}_i$ sur l'espèce j. La dépendance entre les abondances des différentes espèces est entièrement contrôlée par la structure de dépendance latente encodée dans $\boldsymbol{\Omega}$.

Structure de dépendance parcimonieuse. L'inférence de réseau repose généralement sur l'hypothèse que peu de paires d'espèces sont directement dépendantes, ce qui signifie que le modèle graphique décrivant la dépendance entre leurs abondances est peu dense. On peut forcer cette parcimonie en imposant à la matrice de précision $\boldsymbol{\Omega}$ d'être fidèle à un arbre couvrant T (soit $\boldsymbol{Z}_1 \sim \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Omega}_T^{-1})$ où les termes non-nuls de $\boldsymbol{\Omega}_T$ correspondent aux arêtes de l'arbre T), mais cette hypothèse est très restrictive car elle n'autorise que $p-1$ liens parmi p espèces (Chow and Liu, 1968). Une approche plus flexible consiste à supposer que les vecteurs latents sont issus d'un mélange de lois (normales) chacune fidèle à un arbre T:

$$\boldsymbol{Z}_1 \sim \sum_{T \in \mathcal{T}_p} \pi_T \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Omega}_T^{-1}), \tag{3}$$

où $\mathcal{T}_p$ est l'ensemble des arbres couvrants à p noeuds. Si on suppose de plus que la loi sur les arbres donnée par $\{\pi_T\}_{T \in \mathcal{T}_p}$ est factorisable sur les arêtes:

$$\pi_T = \prod_{(j,k) \in T} \beta_{jk} \Big/ B, \qquad B = \sum_{T \in \mathcal{T}_p} \prod_{(j,k) \in T} \beta_{jk}, \tag{4}$$

alors l'inférence d'un tel modèle est accessible d'un point de vue calculatoire grâce au théorème arbre-matrice qui permet de calculer un terme de la forme de la constante de normalisation B en $O(p^3)$ (Chaiken and Kleitman, 1978; Meilă and Jaakkola, 2006; Kirshner, 2008). La probabilité d'un arbre est alors proportionnelle au produit des poids de ses arêtes $\{\beta_{jk}\}_{(j,k) \in T}$.

Introduction d'acteurs manquants. La prise en compte d'acteurs manquants dans le modèle PLN passe par l'hypothèse que le vecteur latent $\boldsymbol{Z}_i$ associé au site i est en fait de dimension $(p + r) \times 1$ et se décompose en $\boldsymbol{Z}_i^\intercal = [\boldsymbol{Z}_{Oi}^\intercal, \boldsymbol{Z}_{Hi}^\intercal]^\intercal$ où $\boldsymbol{Z}_{Oi}$ est de dimension p et correspond aux espèces observées et $\boldsymbol{Z}_{Hi}$ est de dimension r et correspond aux acteurs (espèces ou autres) manquants. Le modèle prévoit alors que

(*i*) les vecteurs d'abondances $\boldsymbol{Y}_i$ des p espèces observées sont distribuées selon (2), en remplaçant $\boldsymbol{Z}$ par $\boldsymbol{Z}_O$,

(*ii*) les $\boldsymbol{Z}_i$ sont distribués selon le mélange décrit par (3) et (4) mais avec des lois normales de dimension $(p + r)$, et des arbres pris dans $\mathcal{T}_{p+r}$.

La Figure 1b) donne le modèle graphique associé à ce modèle. Les données observées $\boldsymbol{Y}$ sont indépendantes de l'arbre T conditionnellement aux variables latentes $\boldsymbol{Z} = (\boldsymbol{Z}_O, \boldsymbol{Z}_H)$, et leur distribution conditionnelle ne dépend même que de $\boldsymbol{Z}_O$. De ce fait $\boldsymbol{Z}_O$ constitue la partie latente associée aux données observées, alors qu'aucune donnée observée n'est directement associée à la partie $\boldsymbol{Z}_H$. Finalement, le modèle PLN original concerne uniquement les variables $\boldsymbol{Y}$ et $\boldsymbol{Z}_O$. L'utilisation d'une structure de dépendance par mélange d'arbres permet une inférence parcimonieuse et efficace, et les acteurs manquants sont pris en compte dans la variable $\boldsymbol{Z}_H$.

2 Inférence par algorithme EM Variationnel

Vraisemblance La vraisemblance jointe des données observées et cachées se factorise comme suit

$$\begin{aligned} p(\boldsymbol{Y}, \boldsymbol{Z}, T) &= p_{\boldsymbol{\beta}}(T)\, p_{\boldsymbol{\Omega}_T}(\boldsymbol{Z} \mid T)\, p_{\boldsymbol{\theta}}(\boldsymbol{Y} \mid \boldsymbol{Z}) \\ &= p_{\boldsymbol{\beta}}(T)\, p_{\boldsymbol{\Omega}_T}(\boldsymbol{Z}_O \mid T)\, p_{\boldsymbol{\Omega}_T}(\boldsymbol{Z}_H \mid \boldsymbol{Z}_O, T)\, p_{\boldsymbol{\theta}}(\boldsymbol{Y} \mid \boldsymbol{Z}_O) \end{aligned}$$

en notant $\boldsymbol{\beta}$ la matrice contenant les poids des arêtes β_{jk}. Cette factorisation est suggérée par le modèle graphique donné en Figure 1b). L'obtention des estimateurs du maximum de vraisemblance demanderait à évaluer la loi conditionnellle de l'ensemble des variables manquantes sachant les données observées, à savoir $p(T, \boldsymbol{Z} \mid \boldsymbol{Y})$.

Approximation variationnelle. Dans le cas présent la loi $p(T, \boldsymbol{Z} \mid \boldsymbol{Y})$ n'a pas de forme simple et nous adoptons donc une approche variationnelle qui vise à maximiser une borne inférieure de la log-vraisemblance de données observées $\log p(\boldsymbol{Y})$, à savoir

$$\begin{aligned} \mathcal{J}(\boldsymbol{Y}; g, h) &= \log p(\boldsymbol{Y}) - KL\left[q(T, \boldsymbol{Z}) || \; p(T, \boldsymbol{Z} \mid \boldsymbol{Y})\right] \\ &= \mathbb{E}_q\big[\log p_{\boldsymbol{\theta}}(\boldsymbol{Y} \mid \boldsymbol{Z}_O) + \log p_{\boldsymbol{\Omega}_T}(\boldsymbol{Z}_O \mid T)\big] - \mathbb{E}_q[\log h(\boldsymbol{Z}_O)] + \mathbb{E}_q[\log p_{\boldsymbol{\Omega}_T}(\boldsymbol{Z}_H \mid \boldsymbol{Z}_O, T)] \\ &\quad + \mathbb{E}_q\big[\log p_{\boldsymbol{\beta}}(T) - \log g(T)\big] - \mathbb{E}_q[\log h(\boldsymbol{Z}_H)] \end{aligned}$$

où $KL\left[q(T, \boldsymbol{Z}) || \; p(T, \boldsymbol{Z} \mid \boldsymbol{Y})\right]$ est la divergence de Küllback-Leibler entre la loi approchée $q(T, \boldsymbol{Z})$ et leur vraie loi conditionnelle.

Lois approchées L'approximation que nous adoptons repose sur une factorisation de la loi q en le produit de deux lois h et g portant respectivement sur $\boldsymbol{Z}$ et sur T: $q(T, \boldsymbol{Z}) = g(T)h(\boldsymbol{Z})$. Du fait de cette factorisation supposée, la contribution de $\mathbb{E}_h[\log p(\boldsymbol{Z} \mid T)]$ ainsi que $\log p(T)$ – seuls termes qui font intervenir T dans la borne inférieure – prennent la forme de sommes sur les arêtes. La loi g minimisant la divergence de Küllback doit donc également prendre une forme factorisable sur les arêtes, à savoir : $g(T) = \left(\prod_{kl} \widetilde{\beta}_{kl}\right) / \widetilde{B}$.

De plus, il découle de l'indépendance des échantillons que h est une loi produit : $h(\boldsymbol{Z}) = \prod_i h_i(\boldsymbol{Z}_i)$. Notre approximation variationnelle consiste à supposer que chacune de ses composantes est une loi normale multivariée : $h(\boldsymbol{Z}_i) = \mathcal{N}(\boldsymbol{Z}_i; \boldsymbol{m}_i, \boldsymbol{S}_i)$, où $\boldsymbol{S}_i$ est diagonale.

2.1 Algorithme proposé.

Optimisation en $\boldsymbol{\theta}$ et en $h(\boldsymbol{Z}_O)$. Nous choisissons de tirer partie de l'algorithme VEM implémenté par Chiquet et al. (2018a,b) dans le package `PLNmodels` pour optimiser la borne inférieure $\mathcal{J}(\boldsymbol{Y}; g, h)$. Il est en effet possible de faire apparaître dans $\mathcal{J}(\boldsymbol{Y}; g, h)$ la borne inférieure d'un modèle PLN simple, notée $\mathcal{J}_{PLN}$. On a :

$$\begin{aligned} \mathcal{J}(\boldsymbol{Y}; g, h) \leq & \mathcal{J}_{PLN}(\boldsymbol{Y}; h(\boldsymbol{Z}_O), \boldsymbol{\theta}) \\ & + \mathbb{E}_{gh}[\log p(\boldsymbol{Z}_H \mid \boldsymbol{Z}_O, T)] + \mathbb{E}_g[\log p(T) - \log g(T)] - \mathbb{E}_h[\log h(\boldsymbol{Z}_H)]. \end{aligned}$$

L'étape VE de l'algorithme maximise $\mathcal{J}_{PLN}$ et fournit les moments d'ordre 1 et 2 de la partie "observée" de la loi approchée h: $\boldsymbol{m}_{Oi}$ et $\boldsymbol{S}_{Oi}$. Reprendre ces estimateurs pour la suite de l'optimisation permet de maximiser $\mathcal{J}(\boldsymbol{Y}; g, h)$ en $\boldsymbol{\theta}$ et en partie en $h(\boldsymbol{Z}_O)$. Il reste une dépendance en $\boldsymbol{Z}_O$ dans $\mathbb{E}_{gh}[\log p(\boldsymbol{Z}_H \mid \boldsymbol{Z}_O, T)]$, c'est donc une solution sous-optimale en $h(\boldsymbol{Z}_O)$.

Etape VE : optimisation en g et $h(\boldsymbol{Z}_H)$. L'étape VE minimise la distance entre les distributions conditionnelle et approchée :

$$\arg \min_{g, h(\boldsymbol{Z}_H)} KL\left(g(T)h(\boldsymbol{Z}) \,||\, p(\boldsymbol{Z}, T \mid \boldsymbol{Y})\right)$$

La mise à jour des poids $\widetilde{\beta}_{jk}$ des arêtes entre deux noeuds observés (resp. un noeud observé et un noeud caché) est obtenue en annulant les termes de la distance qui dépendent des arêtes entre deux noeuds observés (resp. un noeud observé et un noeud caché). La mise à jour de $h(\boldsymbol{Z}_H)$ est obtenue en dérivant l'expression et en reprenant les paramètres de la loi $h(\boldsymbol{Z}_O)$ optimisée à l'étape précédente.

Etape M : optimisation en β et Ω_T. L'étape M maximise la borne inférieure $\mathcal{J}(\boldsymbol{Y}; g, h)$ en les paramètres de $p(\boldsymbol{Z}, T \mid \boldsymbol{Y})$ qui n'ont pas encore été optimisés, soit en $\boldsymbol{\beta}$ et $\boldsymbol{\Omega}_T$:

$$\underset{\boldsymbol{\beta}, \boldsymbol{\Omega}_T}{\operatorname{argmax}} \mathcal{J}(\boldsymbol{Y}; g, h) = \underset{\boldsymbol{\beta}, \boldsymbol{\Omega}_T}{\operatorname{argmax}} \{\mathbb{E}_{gh}[\log(p_{\boldsymbol{\beta}}(T) p_{\boldsymbol{\Omega}_T}(\boldsymbol{Z} \mid T))]\}$$

La loi de l'arbre est mise à jour au travers de $\boldsymbol{\beta}$ grâce à la formule explicitée dans Momal et al. (2020). La mise à jour pour $\boldsymbol{\Omega}_T$ est la même que celle obtenue dans Robin et al. (2019) concernant les éléments hors de la diagonale, en revanche la diagonale est obtenue de manière explicite et ne nécessite pas d'optimisation numérique.

3 Discussion

Nous présentons une approche variationnelle pour l'inférence de réseau à partir de données de comptage avec acteur manquant. Des simulations préliminaires montrent que pour pouvoir être détecté, un acteur manquant doit être lié à beaucoup de variables et donc avoir un effet majeur sur le réseau. Un tel acteur peut être par exemple une espèce écologique centrale, ou une variable environnementale importante comme la température, ou la profondeur.
Ajouter un acteur manquant dans l'inférence de réseau permet d'obtenir certaines de ses caractéristiques, qui pourraient donner une idée de sa nature. L'algorithme VEM développé ici donne plusieurs informations, notamment les probabilités de liens avec les autres noeuds, ainsi que les moyennes et variances de l'acteur estimées sur chacun des sites. Cette présentation sera complétée par des simulations portant sur la capacité de l'algorithme à retrouver le réseau complet, et des exemples illustratifs d'acteurs manquants sur des jeux de données écologiques.

References

J. Aitchison and C. Ho, Biometrika **76**, 643 (1989).

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, Journal of the American Statistical Association **112**, 859 (2017).

J. Chiquet, M. Mariadassou, and S. Robin, in *International Conference on Machine Learning* (2019).

R. Momal, S. Robin, and C. Ambroise, Methods in Ecology and Evolution (2020), URL `https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13380`.

S. Lauritzen and N. Meinshausen, The Annals of Statistics (2012).

G. Robin, C. Ambroise, and S. Robin, Statistical Modelling **19**, 545 (2019).

C. Chow and C. Liu, IEEE Transactions on Information Theory **14**, 462 (1968).

S. Chaiken and D. J. Kleitman, Journal of combinatorial theory, Series A **24**, 377 (1978).

M. Meilă and T. Jaakkola, Statistics and Computing **16**, 77 (2006).

S. Kirshner, in *Advances in Neural Information Processing Systems* (2008), pp. 761–768.

J. Chiquet, M. Mariadassou, and S. Robin, The Annals of Applied Statistics **12**, 2674 (2018a).

J. Chiquet, M. Mariadassou, and S. Robin, arXiv preprint arXiv:1806.03120 (2018b).

Cartes spatiales dans le cerveau des mammifères : modélisation et analyse d'enregistrements expérimentaux

Rémi Monasson & Simona Cocco

Laboratoire de Physique de l'Ecole Normale Supérieure, PSL Research & CNRS UMR8023, 24 rue Lhomond, 75005 Paris

Résumé. Les mammifères sont capables de créer et mémoriser des représentations de leur environnement, appelées cartes spatiales. Quels modèles théoriques permettent de comprendre comment ces cartes sont mémorisées et codent pour la position de l'animal? Je montre ici le point de vue d'un physicien statisticien pour répondre à cette question et comment ces modèles peuvent être confrontés quantitativement aux données expérimentales (enregistrements multi-électrodes dans l'hippocampe) chez le rat.

Mots-clés. cartes spatiales, hippocampe, physique et inférence statistiques, décodage

Abstract. Mammals are able to create and store neural representations of their environment, called spatial maps. What are the theoretical models allowing us to understand how those maps are memorized and code for the animal position? I report here a statistical physicist's viewpoint on this question and how these models can be quantitatively confronted with experimental data (multi-electrode recordings in the hippocampus) in behaving rats.

Keywords. spatial maps, hippocampus, statistical physics and inference, decoding

1 Introduction : cartes spatiales dans le cerveau des mammifères.

Il est crucial pour les animaux de pouvoir établir et mémoriser des représentations mentales de leur environnement. Ces représentations spatiales, ou cartes, sont nécessaires pour permettre la navigation, en particulier, la planification de trajets permettant d'atteindre une destination (source de nourriture, abri, etc ...). Au début de la seconde moitié du XXe siècle, il est apparu, notamment grâce à l'étude du patient HM, que l'hippocampe, une région du cerveau présente chez tous les mammifères, jouait un rôle crucial dans l'acquisition et la mémorisation de ces cartes spatiales. Dans les années 70, J. O'Keefe et ses collaborateurs enregistrèrent l'activité électrique des neurones dans l'hippocampe du rat et découvrit l'existence des cellules de lieu (place cells en anglais), découverte qui fut récompensée par l'attribution du prix Nobel de physiologie et médecine en 2014 [O'Keefe 1978; Moser 2008]. Ces neurones ont la propriété remarquable d'être actifs lorsque l'animal

se trouve dans des régions spécifiques de l'environnement, appelée champs de lieu, ou place fields, et qui dessinent approximativement des disques, et d'être silencieux lorsque l'animal se situe en dehors de cette région. Ainsi, la séquence d'activation des neurones dépend de la trajectoire de l'animal dans l'espace et de la carte qui est utilisée par l'hippocampe (Figure 1, carte spatiale A ou B). Une fois que ces neurones acquièrent leurs champs de lieu (ce qui prend environ une dizaine de minutes dans un nouvel environnement), ceux-ci sont stables : le même neurone sera actif au même endroit lorsque l'animal est replacé quelques jours ou semaines plus tard dans son environnement, et les champs de lieu subsistent dans l'obscurité : les stimuli visuels sont important pour l'acquisition des cartes spatiales, mais pas pour leur réactivation dans un environnement donné. Notons que les neurones de lieu existent aussi en 3D, comme l'ont montré des expériences sur les chauve-souris : les champs de lieu sont approximativement sphériques.

2 Une seule carte : le point de vue de la physique statistique

Comment comprendre l'existence de neurones de lieu? Un modèle relativement simple, introduit par Lebowitz et Penrose à la fin des années 60 [Lebowitz 1966] pour analyser la transition liquide-gaz, permet d'imaginer un mécanisme. Imaginons un gaz sur réseau, par exemple, une grille régulière en dimension $D = 2$ ou 3, et appelons $\vec{r}_i$ les positions des noeuds i dans l'espace. Chacun des N noeuds i du réseau peut être occupé ou pas par une particule du gaz, et on appelle $\sigma_i = 0, 1$ la variable d'occupation correspondante. Ces particules interagissent de manière attractive à courte portée (si elles occupent des sites proches l'un de l'autre) et n'interagissent pas à grande distance.
Dans l'ensemble canonique à température T, la probabilité d'une configuration des nombres d'occupations, $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, ..., \sigma_N)$, est donnée par la distribution de Boltzmann associée à l'énergie

$$E[\boldsymbol{\sigma}] = -\sum_{i<j} J_{ij}\ \sigma_i\ \sigma_j \quad \text{où} \quad J_{ij} = J\big(|\vec{r}_i - \vec{r}_j|\big)\ , \tag{1}$$

avec la contrainte que le nombre de particules est fixé: $\sum_{i=1}^{N} \sigma_i = \rho\ N$, où ρ est la densité. Ici, le couplage effectif $J(d)$ est positif à courte distance d et s'annule lorsque d augmente, par exemple $J(d) \sim e^{-d}$. A température suffisamment basse, ce modèle prédit l'existence d'une phase ferromagnétique, où la majeure partie des $\rho\, N$ particules s'accumulent autour d'une position sur le réseau, formant une goutte liquide de haute densité, tandis que le reste du réseau est envahit par une vapeur saturante de basse densité. La position de la goutte liquide est arbitraire du fait que les interactions J_{ij} entre sites ne dépendent que de leur distance relative $|\vec{r}_i - \vec{r}_j|$: la fonction énergie (1) est invariant par translation. Ainsi,

ce modèle définit une population de variables en interaction, dont les activité 'codent' pour une variable collective, $\vec{r} = \sum_i \sigma_i \vec{r}_i$, qui n'est autre que le centre de la goutte. Lorsque $\vec{r}$ varie, c'est-à-dire que la goutte se déplace, chacune des variables σ_i se comporte comme un 'neurone' (binaire) de lieu : elle sera en moyenne peu active (car faisant partie de la phase vapeur) si $\vec{r}$ est loin de $\vec{r}_i$ et très active (car étant dans la goutte) si $\vec{r} \simeq \vec{r}_i$: on peut donc interpréter $\vec{r}_i$ comme le centre du champ de lieu de la variable i, dont l'extension coïncide avec celle de la goutte (fixée par ρ et la portée des interactions).

Ce modèle de particules peut être traduit directement dans le langage des neurosciences computationnelles (où il a été redécouvert indépendamment par Amari [Amari 1977]) en utilisant la table de correspondance ci-dessous :

particules sur réseau	cellules de lieu
site (occupé ou vide)	neurone (actif ou silencieux)
densité de particules	activité neuronale moyenne
attraction à courte portée	interactions excitatrices [1]
position $\vec{r}_i$ du site i	centre du champ de lieu du neurone i
goutte de liquide	'goutte' d'activité neuronale
position $\vec{r}$ de la goutte	variable collective codée par la population de neurones
densité de particules en $\vec{r}$	activité moyenne des neurones codant pour $\vec{r}$
température T	bruit neuronal

Une conséquence remarquable de l'invariance par translation et de cette correspondance est, que en l'absence de force ou stimulus extérieur déplaçant la goutte neuronale, celle-ci devrait diffuser librement et donc correspondre à une marche aléatoire (virtuelle) dans l'environnement. C'est précisément ce que montre l'analyse de l'activité neuronale d'un rat en train de dormir [Stella 2019].

3 Plusieurs cartes

Se pose alors le problème suivant : alors qu'un animal a un seul hippocampe (dans chaque hémisphère), il est capable d'acquérir et mémoriser un très nombre de cartes spatiales. Comment toutes ces cartes peuvent-elles coexister dans un même réseau neuronal sans (trop) interférer, et comment étendre le modèle (1) ci-dessus à plusieurs cartes? La réponse standard à ces questions [Samsonovitch 1997], consiste à remplacer les couplages

[1] Les interactions excitatrices entre neurones pourraient être une conséquence du phénomène de plasticité synaptique : l'hypothèse actuelle est que, lorsque l'animal parcourt un nouvel environnement, les stimuli sensoriels (visuels, olfactifs, proprioceptifs, ...) activent de manière plus ou moins al'eatoires les neurones dans l'hippocampe ce qui, par apprentissage hebbien renforcerait les connexions entre neurones activés simultanément.

$J_{ij} = J\big(|\vec{r}_i - \vec{r}_j|\big)$ dans l'énergie (1) par

$$J_{ij} = \sum_{\ell=1}^{L} J\big(|\vec{r}_i^{(\ell)} - \vec{r}_j^{(\ell)}|\big) \tag{2}$$

où la somme porte sur les L différents environnements à mémoriser. Dans la formule précédente, $\vec{r}_i^{(\ell)}$ représente la position du centre du champ de lieu de la cellule i dans l'a carte ℓ. Il est en effet connu expérimentalement que la même cellule peut avoir des champs de lieu en des positions apparemment sans aucune relation les unes avec les autres, un phénomène appelé remapping en anglais (Figure 1, comparer les cartes A et B). Le modèle (2) souffre d'un défaut rédhibitoire lorsque le nombre d'environnements, L, est grand (plus précisément, lorsque N et L tendent à l'infini à rapport $\alpha = L/N$ fixé). Si une goutte se forme dans une carte ℓ (ce qui veut dire que les neurones/nombres d'occupations actifs correspondent à des positions proches les unes des autres dans cette environnement, mais pas dans les autres), les couplages codant pour les cartes $\ell' \neq \ell$ brisent l'invariance par translation. De grandes barrières énergétiques, dont on peut calculer les distributions et les corrélations spatiales [Monasson 2014] s'opposent au déplacement de la goutte et la piègent dans quelques positions particulières. L'activité neuronale ne peut plus représenter l'ensemble des positions de l'animal dans l'environnement et la carte devient inutilisable. Si le rapport α devient trop grand (dépasse une valeur critique dépendant de ρ et de la fonction $J(d)$), ces quelques positions n'existent même plus : il devient impossible de former une goutte dans une carte quelconque [Monasson 2013]. Très récemment, nous avons proposé une règle d'apprentissage des couplages J_{ij}, bien plus efficace que (2), qui permet de concilier grande capacité de mémorisation et haute résolution spatiale [Battista 2020].

En dépit de ces défauts, la prescription simple (2) est intéressante pour mémoriser un nombre L petit de cartes (fini, alors que N est envoyé à l'infini). Ce cas limite a une réalisation expérimentale approchée : dans un travail publié en 2011, K. Jezek et ses collaborateurs ont entraîné un rat à explorer deux pièces (appelée A et B) en tous points identiques, sauf pour leur faible éclairage (sur le plafond ou sur le coté selon la pièce) [Jezek 2011]. Ils observent que les cellules de lieu (une trentaine sont enregistrées de manière simultanée) ont des champs de lieu différent dans les deux pièces, par exemple une même cellule sera active dans le coin en haut a gauche de la première pièce et au centre de la deuxième, en accord avec la propriété de remapping (Figure 1). Ensuite, le rat est placé dans une troisième pièce, identique aux deux précédentes mais munie d'un interrupteur permettant de changer abruptement les conditions d'éclairage, c'est-à-dire, en quelque sorte, de téléporter l'animal de la pièce A à B et vice versa! Nous nous sommes intéressés à la modélisation et l'analyse de cette expérience :

D'abord, nous avons montré que, sous l'effet du bruit thermique/neuronal, la goutte d'activité peut spontanément passer d'une carte à l'autre. Deux scénarios sont possibles

[Monasson 2015] : soit la goutte s'évapore de la carte A puis se condense dans la carte B, soit elle transite de A vers B à travers un état mixte (semi-goutte dans les deux cartes en même temps), ce qui n'est possible qu'en des positions très particulières où les topologies des deux cartes se ressemblent localement (les champs de lieu d'un même sous-groupe de cellules se recouvrent dans les deux cartes autour de positions qui forment des sortes de 'trous de vers' neuronaux permettant le passage d'une carte à l'autre). Ces transitions spontanées sont observées dans l'expérience de K. Jezek et ses collaborateurs, où l'activité hippocampale peut changer brutalement de celle de la carte A à celle de la carte B sans que l'interrupteur ne soit manipulé[2].

Ensuite, L. Posani, S. Cocco et moi avons ré–analysé les enregistrements expérimentaux en construisant un décodeur bayesien dual (faisant appel à des modèles graphiques), de la carte et de la position dans la carte, en fonction du vecteur d'activité instantané (Figure 1) [Posani 2018]. Nous avons pu montrer que, bien que la carte chargée par l'hippocampe puisse changer très rapidement (de $A \to B$ ou $B \to A$) sur un temps caractéristique ~ 100 msec, notamment dans les quelques secondes qui suivent un basculement de l'interrupteur, la position codée par la goutte neuronale reste très précisément celle de l'animal. En d'autres mots, la population neuronale exprime à tout moment une information positionnelle précise dans des 'langages' (cartes) qui varient rapidement au cours du temps. Nous avons proposé un modèle, fondé sur un conflit entre les différentes entrées sensorielles, qui permet d'expliquer ce résultat.

Bibliographie

Amari SI. Dynamics of pattern formation in lateral-inhibition type neural fields. Biological cybernetics 27, 77-87 (1977)

Batista A, Monasson R. Capacity-resolution trade-off in the optimal learning of multiple low-dimensional manifolds by attractor neural networks. Phys. Rev. Lett. 124, 048302 (2020)

Jezek K, Henriksen EJ, Treves A, Moser EI, Moser MB. Theta-paced flickering between place-cell maps in the hippocampus. Nature 478, 246-249 (2011)

Lebowitz J, Penrose O. Rigorous treatment of the Van der Waals-Maxwell theory of the liquid-vapor transition. Journal of Mathematical Physics 7, 98-113 (1966)

Monasson R, Rosay S. Crosstalk and transitions between multiple spatial maps in an attractor neural network model of the hippocampus: Phase diagram. Phys. Rev. E 87, 062813 (2013)

Monasson R, Rosay S. Crosstalk and transitions between multiple spatial maps in an attractor neural network model of the hippocampus: Collective motion of the activity.

[2]Il ne faut pas oublier que les rats de laboratoires ont des expériences sensorielles extrêmement pauvres et ne connaissent qu'un nombre très limité de situations environnementales différentes par rapport aux rats vivant en milieu naturel.

Phys. Rev. E 89, 032803 (2014)
Monasson R, Rosay S. R. Transitions between spatial attractors in place-cell models. Phys. Rev. Lett. 115, 098101 (2015)
Moser EI, Kropff E, Moser MB. Place Cells, Grid Cells, and the Brain's Spatial Representation System. Annual Review of Neuroscience 31, 69-89 (2008)
O'Keefe J, Nadel L, *The Hippocampus as a Cognitive Map*, Oxford University Press (1978)
Posani L, Cocco S, Monasson R. Integration and multiplexing of positional and contextual information by the hippocampal network. PLoS Computational Biology 14, e1006320 (2018)
Samsonovich A, McNaughton BL. Path Integration and Cognitive Mapping in a Continuous Attractor Neural Network Model. J. Neurosci. 17, 5900 (1997).
Stella F, Baracskay P, O' Neill J, Csicsvari J. Hippocampal Reactivation of Random Trajectories Resembling Brownian Diffusion. Neuron 102, 450-461 (2019)

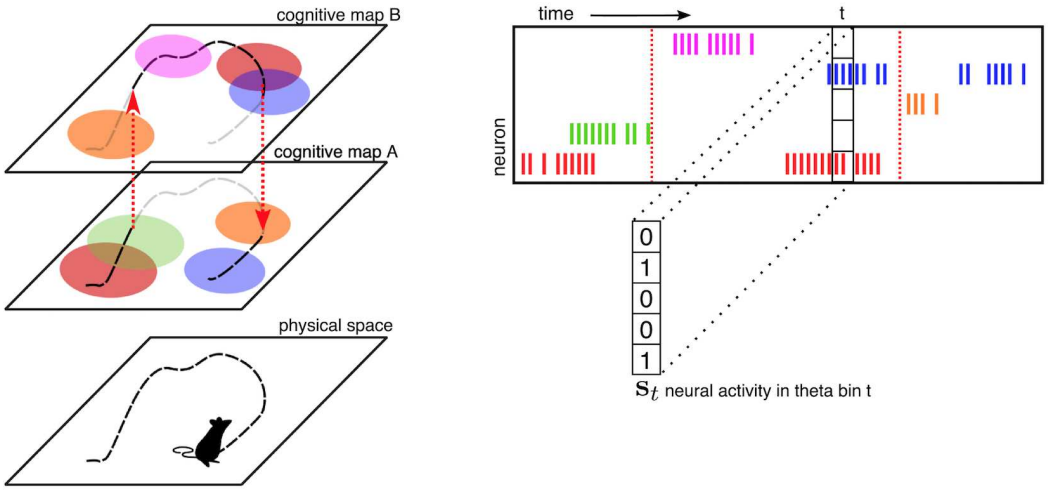


Figure 1: Cellules de lieu, cartes spatiales et décodage de la carte et de la position [Posani 2018]. Gauche : lorsqu'un rongeur se déplace dans un environnement, il traverse les champs de lieu de neurones de l'hippocampe, qui s'activent séquentiellement. Droite : activité des cellules, que l'on peut résumer dans un vecteur binaire (0: cellule silencieuse, 1: active) à chaque temps t. Chaque couleur correspond à une cellule de lieu, qui peut avoir un champ de lieu dans une ou plusieurs cartes spatiales (gauche). Dans l'expérience de Jezek et al., à tout moment, l'hippocampe utilise une des deux cartes spatiales apprises par le rat; les changements de cartes sont indiqués par des flèches verticales en pointillés.

Approximation stochastique de vecteurs et valeurs propres. Application à l'ACG en ligne.

Jean-Marie Monnez [1,2,*]

[1] *Université de Lorraine, CNRS, Inria*, IECL**, F-54000 Nancy, France*
**Inria, Project-Team BIGS, F-54600 Villers-lès-Nancy*
***IECL, Institut Elie Cartan de Lorraine, F-54506 Vandœuvre-lès-Nancy*
[2] *INSERM U1116, Centre d'Investigation Clinique Plurithématique 1433, Université de Lorraine, Nancy, France*
∗jean-marie.monnez@univ-lorraine.fr
Financement : Programme Investissement d'Avenir ANR-15-RHU-0004

Résumé. Nous avons étendu le domaine d'application du processus d'approximation stochastique de vecteurs propres de Oja en en démontrant la convergence presque sûre sous des hypothèses plus générales. Nous étudions l'application à l'analyse canonique généralisée (ACG) d'un vecteur aléatoire Z dans le cas de données massives ou en flux. Les composantes générales de l'ACG sont les composantes principales de l'ACP de Z avec une métrique particulière M. Nous définissons des processus d'approximation stochastique où l'on peut utiliser à chaque étape toutes les observations de Z effectuées jusqu'à cette étape sans avoir à les stocker au lieu uniquement des nouvelles observations à ce pas, pour estimer simultanément la métrique M, les composantes générales de l'ACG et les valeurs propres associées.

Mots-clés. Analyse canonique généralisée, Approximation stochastique, Données massives, Estimation en ligne, Flux de données, Valeurs propres, Vecteurs propres.

Abstract. We widened the scope of the 0ja's eigenvector stochastic approximation process proving its almost sure convergence under more general assumptions. We study the application to generalized canonical correlation analysis (gCCA) of a random vector Z in the case of big or streaming data. The general components of gCCA are principal components of PCA of Z with a particular metric M. We define stochastic approximation processses using at each step all observations up to this step without storing them instead of the new observations at this step only, to estimate simultaneously the metric M, the general components of gCCA and the corresponding eigenvalues.

Keywords. Big data, Data stream, Eigenvalues, Eigenvectors, Generalized canonical correlation analysis, Online estimation, Stochastic approximation.

1 Analyse de données massives ou en flux

Dans le contexte d'un flux de données, on peut utiliser des algorithmes récursifs pour estimer en ligne des paramètres d'intérêt, par exemple :

- les paramètres d'une fonction de régression linéaire (Duarte et al, 2018) ou logistique (Lalloué et al, 2019) ;

- les centres de classe en classification non supervisée (Cardot et al, 2012) ;

- des composantes principales en ACP (Monnez et Skiredj, 2019).

Le principe en est que chaque vecteur de données entrant est utilisé pour actualiser l'estimation courante des paramètres d'intérêt. Dans le cas d'un tableau de données massives, ce type de méthode peut aussi être utilisé en effectuant dans le temps une succession de tirages au hasard de lignes du tableau.

Les avantages de ces méthodes séquentielles sont multiples :

- on n'a pas besoin de stocker les données ;

- on peut prendre en compte beaucoup plus de données qu'avec les méthodes non séquentielles durant la même durée de temps ;

- elles utilisent moins de place que les méthodes non séquentielles.

Nous avons montré que l'on pouvait utiliser à chaque étape dans certaines méthodes, au lieu d'un lot de nouvelles données, toutes les données jusqu'à l'étape courante sans avoir à les stocker, donc prendre en compte toute l'information contenue dans les données précédentes et améliorer ainsi en général la vitesse de convergence des processus, comme cela a été vérifié empiriquement dans le cas de la régression linéaire en ligne (Duarte et al, 2018) et l'ACP en ligne (Monnez et Skiredj, 2019).

2 L'algorithme de Oja

Soit B une matrice (p,p) symétrique de vecteurs propres normés $V_1, ..., V_p$ associés aux valeurs propres $\lambda_1 > ... > \lambda_p$. Soit (a_n) une suite de nombres réels positifs. $\|x\|$ désigne la norme euclidienne usuelle d'un vecteur x de $\mathbb{R}^p$, la norme matricielle est la norme spectrale.

Supposons que B soit inconnue et qu'il existe une suite $(B_1, ..., B_n, ...)$ de matrices aléatoires symétriques mutuellement indépendantes, bornées presque sûrement et d'espérance mathématique B. Soit le processus stochastique normé $(X_n, n \geqslant 1)$ défini par Oja et Karhunen (1985) tel que :

$$X_{n+1} = \frac{(I + a_n B_n)\, X_n}{\|(I + a_n B_n)\, X_n\|}. \tag{1}$$

La convergence presque sûre de ce processus vers un vecteur propre normé associé à la plus grande valeur propre de B a été établie par Oja et Karhunen (1985). Sa rapidité de convergence est étudiée dans Balsubramani et al (2013). Remarquons que, T_n étant la tribu engendrée par $X_1, B_1, ..., B_{n-1}$, on a $E\left[B_n \mid T_n\right] = B$.

Dans le cas de l'ACP d'un vecteur aléatoire Z, $\mathbb{R}^p$ étant muni d'une métrique M qui peut dépendre de caractéristiques de Z, pour déterminer les composantes principales on recherche les premiers vecteurs propres de la matrice M^{-1}symétrique $B =$

$ME\left[(Z-E[Z])(Z-E[Z])'\right]$. Dans le cas d'un flux d'observations de Z, $E[Z]$ et M ne sont pas connues a priori, mais peuvent être estimées en ligne. L'hypothèse d'indépendance des B_n n'est alors pas vérifiée.

De façon générale, soit une matrice Q-symétrique B, pour $n>1$, Q_n une métrique connue à l'étape $n-1$ convergeant presque sûrement vers Q, $\langle .,. \rangle_n$ et $\|.\|_n$ le produit scalaire et la norme induits par la métrique Q_n. Nous établissons dans (Monnez, 2020) un théorème de convergence presque sûre de processus $(X_n^i), i=1,2,...,r$, obtenus par une orthonormalisation de Gram-Schmidt à l'étape n par rapport à Q_{n+1}, vers $\pm V_i$ et $(\Lambda_n^i), i=1,2,...,r$, vers λ_i, tels que :

$$\widetilde{Y}_{n+1}^i = (I+a_nB_n)\widetilde{X}_n^i \tag{2}$$

$$\widetilde{X}_{n+1}^i = \widetilde{Y}_{n+1}^i - \sum_{j<i}\left\langle \widetilde{Y}_{n+1}^j, X_{n+1}^j\right\rangle_{n+1} X_{n+1}^j, \quad X_{n+1}^i = \frac{\widetilde{X}_{n+1}^i}{\left\|\widetilde{X}_{n+1}^i\right\|_{n+1}} \tag{3}$$

$$\Lambda_{n+1}^i = (1-a_n)\Lambda_n^i + a_n\left\langle B_nX_n^i, X_n^i\right\rangle_n. \tag{4}$$

Ce théorème peut être utilisé dans des cas où $E[B_n \mid T_n]$ converge p.s. vers B ou B_n converge p.s. vers B.
Il étend ou complète ceux donnés par Benzécri (1969), Oja et Karhunen (1985), Duflo (1997) et Brandière (1998) pour l'ACP, Monnez et Skiredj (2019). Nous en énonçons un corollaire dans le cas où $B_n - B = B_n^1 + B_n^2$ en supposant :

H1 (a) B est Q-symétrique. (b) Les valeurs propres de B sont simples.

H2 (a) $\sum_1^\infty a_n\|B_n^1\| < \infty$ p.s., $E[B_n^2 \mid T_n] = 0$ p.s., $E\left[\sup_n \|B_n^2\|^2\right] < \infty$.(b) Pour tout n, $I+a_nB_n$ est inversible.

H3 (a) $a_n > 0, \sum_1^\infty a_n = \infty, \sum_1^\infty a_n^2 < \infty$.

H3 (b) $a_n = \frac{a}{n^\alpha}$ avec $a>0, \frac{2}{3} < \alpha \leq 1$ et $a > \frac{1}{2}$ pour $\alpha = 1$ (une hypothèse plus générale peut être formulée).

H4 $Q_n \longrightarrow Q, \sum_1^\infty a_n\|Q_n - Q\| < \infty$ p.s.

H5 Pour $i=1,...,r, X_1^i$ est une variable aléatoire absolument continue indépendante de $B_1,...,B_n,....$

Théorème 1 *Sous les hypothèses H1a,b,2a,b,3b,4,5, on a presque sûrement pour* $i=1,...,r$ *:* $X_n^i \longrightarrow V_i$ ou $-V_i$, $\Lambda_n^i \longrightarrow \lambda_i$, $\sum_1^\infty a_n|\langle BX_n^i, X_n^i\rangle_n - \lambda_i| < \infty$. *Dans le cas où* $B_n^2 = 0$, *on a les mêmes conclusions en remplaçant H3b par H3a ; en outre* $\sum_1^\infty a_n|\langle B_nX_n^i, X_n^i\rangle_n - \lambda_i| < \infty$ *et* $\sum_1^\infty a_n|\Lambda_n^i - \lambda_i| < \infty$ *presque sûrement.*

3 L'analyse canonique généralisée

3.1 Formulation probabiliste

Supposons que l'ensemble des composantes d'un vecteur aléatoire Z dans $\mathbb{R}^p$ soit partitionné en q sous-ensembles de variables aléatoires réelles $\left\{Z^{k1},...,Z^{kr_k}\right\}$, $k=1,...,q$. Soit Z^k le vecteur aléatoire dans $\mathbb{R}^{r_k}$ dont les composantes sont $Z^{k1},...Z^{kr_k}$. Soit C^k la matrice de covariance de Z^k, C celle de Z. Supposons qu'il n'y ait pas de relation affine entre les composantes de Z, ainsi $C^k, k=1,...,q$ et C sont inversibles.

Soit le problème suivant : pour $l=1,...,r\leq p$, déterminer au pas l une combinaison linéaire de toutes les composantes centrées de Z, $U_l=\theta_l'\left(Z-E\left[Z\right]\right)$, appelée $l^{ième}$ composante générale, de variance 1 et non corrélée à $U_1,...,U_{l-1}$, et, pour $k=1,...,q$, une combinaison linéaire de variance 1 des composantes centrées de Z^k, $V_l^k=\left(\eta_l^k\right)'\left(Z^k-E\left[Z^k\right]\right)$, appelée $l^{ème}$ composante canonique du $k^{ième}$ sous-ensemble de variables, qui maximisent $\sum_{k=1}^q\rho^2\left(U_l,V_l^k\right)$, ρ désignant le coefficient de corrélation linéaire.

Soit M la matrice inconnue d'ordre p diagonale par blocs dont le $k^{ième}$ bloc diagonal est $M^k=\left(C^k\right)^{-1}$. Soit $\theta_l=\left(\left(\theta_l^1\right)'...\left(\theta_l^q\right)'\right)', \theta_l^k\in\mathbb{R}^{m_k}, k=1,...,q$. On montre que θ_l est un vecteur propre C-normé de la matrice M^{-1}-symétrique $B=MC$ correspondant à sa $l^{ième}$ plus grande valeur propre λ_l et que pour $k=1,...,q$, il existe $\alpha_l^k\in\mathbb{R}$ tel que $\eta_l^k=\alpha_l^k\theta_l^k$. Remarquons que $\sqrt{\lambda_l}\left(\theta_l\right)'\left(Z-E\left[Z\right]\right)$ est la $l^{ième}$ composante principale de l'ACP de Z avec la métrique M. L'objectif est donc de réaliser l'ACP en ligne de Z en utilisant à l'étape n un estimateur convergent M_n de M.

Dans le cas où pour tout k, $m_k=1$, cette analyse est équivalente à l'ACP normée.

Dans le cas $q=2$, cette analyse est équivalente à l'analyse canonique de deux ensembles de variables, qui a pour cas particuliers l'analyse factorielle discriminante et l'analyse factorielle des correspondances. Pour $k=1,2$, $\left(\theta_l^k\right)'\left(Z^k-E\left[Z^k\right]\right)$ est colinéaire à la $l^{ième}$ composante canonique du $k^{ième}$ ensemble de variables.

3.2 Approximation stochastique dans le cas d'un flux

Soit $(Z_{11},...,Z_{1m_1},...,Z_{n1},...,Z_{nm_n},...)$ un échantillon i.i.d. de Z avec $Z_{ij}=\left(Z_{ij}^1,...,Z_{ij}^q\right)$; $Z_{n1},...,Z_{nm_n}$ sont observés à l'étape n du processus. On note T_n la tribu du passé à l'étape n, par rapport à laquelle $Z_{11,}...,Z_{n-1,m_{n-1}}$ sont mesurables. On note $\overline{Z}_n$ la moyenne des Z_i observés jusqu'à l'étape n, et $\overline{Z}_n^k$ celle des Z_i^k pour $k=1,...,q$. On note C_n la matrice de covariance des Z_i observés jusqu'à l'étape n, et C_n^k celle des Z_i^k pour $k=1,...,q$, qui peuvent être calculées de façon récursive :

$$C_n^k=\frac{1}{\mu_n}\sum_{i=1}^n\sum_{j=1}^{m_i}Z_{ij}^kZ_{ij}^{k\prime}-\overline{Z}_n^k\overline{Z}_n^{k'},\mu_n=\sum_{i=1}^n m_i. \tag{5}$$

$M^k = \left(C^k\right)^{-1}$ est solution de l'équation en X : $C^k X - I = 0$ ou

$$E\left[\left(Z^k Z^{k\prime} - E\left[Z^k\right] E\left[Z^k\right]'\right) X - I\right] = 0. \tag{6}$$

Pour estimer $M^k = \left(C^k\right)^{-1}$, on définit le processus d'approximation stochastique $\left(M_n^k, n \geqslant 1\right)$ tel que :

$$M_n^k = M_{n-1}^k - a_n \left(C_n^k M_{n-1}^k - I\right). \tag{7}$$

On fait les hypothèses :

H6 (a) Il n'y a pas de relation affine entre les composantes de Z ; (b) Z admet des moments d'ordre 4.

H3 (c) $\sum_1^\infty \frac{a_n}{\sqrt{n}} < \infty$.

Théorème 2 *Sous les hypothèses H6a,b,3a,c, on a presque sûrement :* $M_n^k \longrightarrow \left(C^k\right)^{-1}$, $\sum_{n=1}^\infty a_n \left\|M_n^k - \left(C^k\right)^{-1}\right\| < \infty$.

Soit M_n et N_n les matrices diagonales par bloc dont les $k^{ièmes}$ blocs diagonaux sont respectivement M_n^k et $C_n^k, k = 1, ...q$. N_n est symétrique positive de plein rang à partir d'un certain n que l'on suppose égal à 1. Soit :

$$B_n = M_n\left(\omega_{1n} C_n + \omega_{2n}\left(\frac{1}{m_n}\sum_{j=1}^{m_n} Z_{nj} Z_{nj}' - \overline{Z}_n \overline{Z}_n'\right)\right), \omega_{1n} \geqslant 0, \omega_{2n} \geqslant 0, \omega_{1n} + \omega_{2n} = 1. \tag{8}$$

B_n est la somme de deux termes, le premier correspondant à l'utilisation de toutes les observations jusqu'à l'étape n et le deuxième à celle d'un lot de m_n observations entrées à cette étape, pondérés par un poids compris entre 0 et 1 et dépendant de n. On peut aussi définir B_n de telle manière qu'à partir d'une étape N, on ne tienne plus compte des observations effectuées avant cette étape. On définit les processus $(X_n^i, n \geqslant 1)$ et $(\Lambda_n^i, n \geqslant 1)$ comme précédemment en faisant l'orthonormalisation de Gram-Schmidt à l'étape n par rapport à N_n (on a $Q_{n+1} = N_n$). On fait l'hypothèse :

H6 (c) Z est p.s. bornée.

Théorème 3 *Sous les hypothèses H1b,3b,c,5,6a,c, on a presque sûrement pour* $i = 1, ..., r$ *:* $X_n^i \longrightarrow V_i$ ou $-V_i$, $\Lambda_n^i \longrightarrow \lambda_i$, $\sum_1^\infty a_n \left|\langle B X_n^i, X_n^i\rangle_n - \lambda_i\right| < \infty$. *Dans le cas où* $\omega_{2n} = 0$, *on a les mêmes conclusions en remplaçant H3b par H3a et H6c par H6b ; on a en outre* $\sum_1^\infty a_n \left|\langle B_n X_n^i, X_n^i\rangle_n - \lambda_i\right| < \infty$ *et* $\sum_1^\infty a_n \left|\Lambda_n^i - \lambda_i\right| < \infty$ *presque sûrement.*

4 Conclusion

Nous avons présenté un corollaire et une application du théorème général de convergence presque sûre du processus de Oja établi dans (Monnez, 2020) à l'analyse canonique généralisée en ligne d'un vecteur aléatoire en estimant par un processus d'approximation

stochastique la métrique inconnue et en utilisant à chaque étape des processus définis toutes les observations du vecteur aléatoire effectuées jusqu'à cette étape. On peut aussi utiliser à chaque étape seulement les nouvelles observations entrantes à cette étape ($\omega_{1n} = 0$). Les expériences effectuées dans le cas de l'ACP en ligne dans les cas $\omega_{1n} = 0$ (utilisation d'un lot de nouvelles observations à l'étape courante) ou $\omega_{2n} = 0$ (utilisation de toutes les observations jusqu'à l'étape courante) ont montré en général une plus grande rapidité de convergence des processus avec $\omega_{2n} = 0$ (Monnez et Skiredj, 2020).

5 Bibliographie

Balsubramani, A., Dasgupta S. et Freund, Y. (2013), The fast convergence of incremental PCA, *NIPS*, pp. 3174-3182.

Benzécri, J.P. (1969), Approximation stochastique dans une algèbre normé non commutative, *Bull. Soc. Math. France,* 97, pp. 225-241.

Brandière, O. (1998), Some pathological traps for stochastic approximation, *Siam J. Control Optim.,* 36, No. 4, pp. 1293-1314.

Cardot, H., Cénac, P. et Monnez, J.M. (2012), A fast and recursive algorithm for clustering large datasets with k-medians, *Computational Statistics and Data Analysis*, 56, pp. 1434-1449.

Duarte, K., Monnez, J.M. et Albuisson, E. (2018), Sequential linear regression with online standardized data, *PLoS ONE*, 13 (1) : e0191186.

Duflo, M. (1997), *Random Iterative Models*, Applications in Mathematics, 34, Springer-Verlag, Berlin.

Lalloué, B., Monnez, J.M. et Albuisson, E. (2019), Streaming constrained binary logistic regression with online standardized data, *Soumis.*

Monnez, J.M. (2020), Stochastic approximation of eigenvectors and eigenvalues of the Q-symmetric expectation of a random matrix, *Pré-publication.*

Monnez, J.M. et Skiredj, A. (2019), Convergence of a normed eigenvector stochastic approximation process and application to online principal component analysis of a data stream, *hal-01844419.*

Monnez, J.M. et Skiredj, A. (2020), Widening the scope of an eigenvector stochastic approximation process and application to streaming PCA and related methods, *Soumis.*

Oja, E. et Karhunen, J. (1985), On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix, *Journal of Mathematical Analysis and Applications,* 106, pp. 69-84.

Sous-produits de la théorie des modèles dépendant du temps

Guy Mélard [1] & Rajae Azrak [2]

[1] *Université libre de Bruxelles, Solvay Brussels School of Economics and Management, ECARES, Bruxelles, Belgique, gmelard@ulb.ac.be*
[2] *Université Mohammed V - Rabat, Faculté des Sciences juridiques, économiques et sociales, Salé, Maroc, rajae.azrak@gmail.com*

Résumé. Cette contribution constitue une continuation de plusieurs articles dans lesquels nous avons étudié des modèles ARMA et VARMA à coefficients dépendant du temps, respectivement tdARMA et tdVARMA. Les coefficients des modèles sont supposés être des fonctions déterministes du temps t, et d'un petit nombre de paramètres. Nous montrons ici trois applications de la théorie qui a été exposée dans un article récent noté AALM : (i) une démonstration des propriétés asymptotiques des modèles VARMA classiques, (ii) une étude de modèles à seuils, un type de modèles avec plusieurs régimes, (iii) une étude de modèles pour données de comptage. Ces trois applications sont possibles parce que la théorie asymptotique présentée dans AALM ne requiert pas la stationarité, ni l'ergodicité et n'emploie pas d'arguments d'analyse spectrale.

Mots-clés. Série temporelle, modèle VARMA, modèle à seuils, modèle pour données de comptage

Abstract. This contribution is a follow-up to several papers where we have studied ARMA and VARMA models with time-dependent coefficients, or tdARMA and tdVARMA models. The model coefficients are supposed to be deterministic functions of time t, and of a small number of parameters. Here we show three applications of the theory that was exposed in a recent paper denoted AALM: (i) a proof of the asymptotic properties of standard VARMA models, (ii) a study of threshold ARMA models, a kind of models with different regimes, (iii) a study of count data models. These three applications are possible because the asymptotic theory in AALM does not require stationarity, nor ergodicity and does not make use of spectral arguments.

Keywords. Time series, time-dependent model, VARMA model, threshold model, count data model

1 Introduction

Cette contribution constitue une continuation de plusieurs articles et communications (notamment aux journées de statistique de 2015, 2017 et 2019) dans lesquels nous avons

étudié des modèles ARMA et VARMA à coefficients dépendant du temps, respectivement tdARMA et tdVARMA. Les coefficients des modèles sont supposés être des fonctions déterministes du temps t, et d'un petit nombre de paramètres. Nous montrons ici trois applications de la théorie qui a été exposée dans un article récent Alj *et al.* (2017) noté AALM : (i) une démonstration des propriétés asymptotiques des modèles VARMA classiques, (ii) une étude de modèles à seuils, un type de modèles avec plusieurs régimes, (iii) une étude de modèles pour données de comptage. Ces trois applications sont possibles parce que la théorie asymptotique présentée dans AALM ne requiert pas la stationarité, ni l'ergodicité et n'emploie pas d'arguments d'analyse spectrale. En conséquence, d'abord la démonstration des propriétés asymptotiques des modèles VARMA classiques esquissée dans le paragraphe 2 diffère d'autres démonstrations. En second lieu, l'analyse de modèles non standard des paragraphes 3 et 4 diffère d'autres approches qui exigent d'exhiber une solution de l'équation du modèle et de montrer qu'elle est strictement stationaire et ergodique. Au lieu de cela, on montre que la théorie de AALM peut être appliquée. Les résultats des paragraphes 2 et 3 sont basés sur une représentation d'un modèle tdARMA ou tdVARMA sous la forme de modèle tdVAR d'ordre 1. Dès lors, les hypothèses énoncées dans AALM peuvent être vérifiées. Le résultat du paragraphe 4 est basé sur un modèle log-linéaire généralisé légèrement différent des modèles usuels. On emploie aussi une quasi-vraisemblance gaussienne de manière à pouvoir utiliser la théorie de AALM.

2 Propriétés asymptotiques de modèles VARMA classiques

2.1 Modèles VARMA à coefficients dépendant du temps

Soit $\{x_t : t \in \mathbf{N}\}$ un processus stochastique à valeurs dans $\mathbf{R}^r$. Notons $\theta = (\theta_1, ..., \theta_m)^T$ le vecteur des paramètres d'intérêt, de vraie valeur θ^0. Un modèle vectoriel autorégressif-moyenne mobile (VARMA) à coefficients dépendant du temps d'ordre (p, q), de moyenne 0, est défini par l'équation

$$x_t = \sum_{i=1}^{p} A_{ti}(\theta) x_{t-i} + e_t(\theta) + \sum_{j=1}^{q} B_{tj}(\theta) e_{t-j}(\theta), \tag{1}$$

où les $e_t(\theta)$ sont les résidus tels que $e_t(\theta^0) = \epsilon_t$ et $\{\epsilon_t : t \in \mathbf{N}\}$ est un processus bruit blanc constitué de vecteurs aléatoires indépendants, de moyenne 0, de matrice de covariance Σ $r \times r$ inversible dont les éléments sont des paramètres de nuisance. Les coefficients $A_{ti}(\theta)$, $i = 1, ..., p$, et $B_{tj}(\theta)$, $j = 1, ..., q$, sont des matrices $r \times r$, fonctions déterministes du temps t et de θ. En abrégé on écrira tdVARMA(p, q). Supposons une série d'observations $\{x_t : t = 1, \ldots, n\}$. L'estimation de θ s'effectue en minimisant la log-vraisemblance gaussienne qui est proportionnelle à la somme pour $t = 1, \ldots, n$ de

$$\alpha_t(\theta) = \log\left(\det\left(\Sigma_t(\theta)\right)\right) + e_t^T(\theta)\Sigma_t^{-1}(\theta)e_t(\theta), \tag{2}$$

où $\Sigma_t(\theta) = E(e_t(\theta)e_t^T(\theta))$. La théorie développée dans AALM permet, sous des hypothèses très générales esquissées dans l'appendice, d'établir que l'estimateur $\hat{\theta}_n$ de θ converge presque sûrement vers θ^0 quand $n \longrightarrow \infty$, et qu'il existe une matrice inversible V telle que $n^{1/2}(\hat{\theta}_n - \theta^0) \longrightarrow N(0, V^{-1})$ en loi, lorsque $n \longrightarrow \infty$.

2.2 Préparatifs

On montre d'abord que le modèle tdVARMA(p, q) (1) peut être mis sous la forme (3) d'un modèle tdVAR(1). On définit les vecteurs $X_t(\theta) = (x_t, \ldots, x_{t-p+1}, e_t(\theta), \ldots, e_{t-q+1}(\theta))^T$ et $E_t(\theta) = (e_t(\theta), 0, \ldots, 0, e_t(\theta), 0, \ldots, 0)^T$, de dimension $(p+q)r$, et une matrice $\phi_t(\theta)$ tels que

$$X_t(\theta) = \phi_t(\theta)X_{t-1}(\theta) + E_t(\theta), \tag{3}$$

où le bloc supérieur gauche de $\phi_t(\theta)$ est la matrice compagnon du polynôme réciproque $A_t^*(z, \theta) = z^p - \sum_{i=1}^p A_{ti}(\theta)z^{p-i}$. Soit $\tilde{X}_t = (x_t, 0, \ldots, 0, x_t, 0, \ldots, 0)^T$. Il existe une matrice K telle que $\tilde{X}_t = KX_t(\theta^0)$. Soit J une matrice identité d'ordre $(p+q)r$ où on remplace les blocs $r \times r$ $(1,1)$ et $(p+1, p+1)$ par une matrice 0_r et le bloc $(p+1, 1)$ par $-I_r$, où I_r est la matrice identité d'ordre r. Alors le bloc inférieur droit de $J\phi_t(\theta)$ est la matrice compagnon du polynôme réciproque $B_t^*(z, \theta) = z^q + \sum_{j=1}^q B_{tj}(\theta)z^{q-j}$. Notons $E_t^0 = E_t(\theta^0)$ et $\phi_t^0 = \phi_t(\theta^0)$. Après plusieurs manipulations on obtient

$$X_t(\theta) = \sum_{k=0}^{t-1} \Psi_{tk}(\theta)E_{t-k}^0, \quad \Psi_{tk}(\theta) = \sum_{s=0}^{k} \left\{ \left(\prod_{\ell=0}^{k-s-1} J\phi_{t-\ell}(\theta) \right) K \left(\prod_{j=0}^{s-1} \phi_{t-k+s-j}^0 \right) \right\}, \tag{4}$$

et donc, en considérant le bloc de r lignes en position $p+1$ de (3), $e_t(\theta) = \sum_{k=0}^{t-1} \psi_{tk}(\theta)\epsilon_{t-k}$, $\psi_{tk}(\theta) = U_{p+1}^T \Psi_{tk}(\theta)U_1$, avec les matrices $(p+q)r \times r$ $U_{p+1} = (0 \ \ldots \ 0 \ I_r \ 0 \ \ldots \ 0)^T$, où I_r est en position $(p+1)$, $U_1 = (I_r \ 0 \ \ldots \ 0)^T$, où I_r est en position 1. Dès lors, en notant $\psi_{tik}(\theta) = \partial\psi_{tk}(\theta)/\partial\theta_i$, pour $i = 1, \ldots, m$,

$$\frac{\partial e_t(\theta)}{\partial \theta_i} = \sum_{k=1}^{t-1} \psi_{tik}(\theta)\epsilon_{t-k}, \quad \Psi_{tik}(\theta) = \frac{\partial \Psi_{tk}(\theta)}{\partial \theta_i}, \quad \psi_{tik}(\theta) = U_{p+1}^T \Psi_{tik}(\theta)U_1, \tag{5}$$

généralisant Francq et Gautier (2004). On peut procéder de même pour les dérivées secondes et troisièmes des $e_t(\theta)$.

2.3 Modèles VARMA classiques

Dans un modèle VARMA classique, les coefficients ne dépendent pas de t, donc $A_{ti}(\theta) = A_i(\theta)$, $i = 1, ..., p$, et $B_{tj}(\theta) = B_j(\theta)$, $j = 1, ..., q$. Les paramètres sont les éléments des matrices de coefficients, de vraies valeurs respectives A_i^0 et B_j^0, le temps t va de $-\infty$ à $+\infty$ et on suppose la stationnarité, l'inversibilité et l'indentifiabilité du processus. On

suppose aussi l'existence de moments d'ordre $4+\delta$, $\delta > 0$, pour les ϵ_t. Pour satisfaire les deux premières conditions, on suppose que les polynômes $A^0(z) = 1 - \sum_{i=1}^{p} A_i^0 z^i$ et $B^0(z) = 1 + \sum_{i=1}^{q} B_j^0 z^j$ sont tels que les racines des équations $\det(A^0(z)) = 0$ et $\det(B^0(z)) = 0$ sont supérieures à 1 en module. Pour une démonstration classique des propriétés asymptotique des estimateurs, voir Yao et Brockwell (2006) pour les modèles ARMA ou Hannan et Deistler (1998) pour les modèles VARMA. Pour un modèle VARMA, l'application de la technique du paragraphe 2.2, avec $\psi_{tik}(\theta)$ dans (5) remplacé par $\psi_{ik}(\theta)$, conduit pour les deux produits dans (4) à des puissances de matrices. Sous les conditions sur les deux polynômes matriciels, la norme de Frobenius de ces produits peut être majorée par des nombres inférieurs à 1. Plus précisément $||\psi_{ik}(\theta^0)||_F < \Phi^k$ pour un $\Phi < 1$. On peut alors vérifier en quelques pages les hypothèses de AALM esquissées dans l'appendice, voir l'article complet de Mélard (2020). On y discute également une condition suffisante d'identifiabilité. Un théorème dû à Ruiz (1996) est employé.

3 Analyse d'un modèle à seuils

Nous allons traiter ici le modèle le plus simple de type TAR(1) défini, voir Tong (1990), par

$$x_t = \begin{cases} \theta_1 x_{t-1} + \epsilon_t, & \text{si } x_{t-1} \leq 0, \\ \theta_2 x_{t-1} + \epsilon_t, & \text{si } x_{t-1} > 0, \end{cases} , \tag{6}$$

où θ_1 et θ_2 sont des paramètres et les ϵ_t i.i.d. $(0, \sigma^2)$. Soit $I_{t-1} = 1$, si $x_{t-1} \leq 0$ et $I_{t-1} = 0$, autrement, on peut écrire (6) comme

$$x_t = \{\theta_1 I_{t-1} + \theta_2(1 - I_{t-1})\} x_{t-1} + \epsilon_t. \tag{7}$$

L'équation (7) ressemble à celle d'un modèle tdAR(1). Les conditions principales sont pour les coefficients $\psi_{tik}(\theta^0)$. Une condition suffisante est que les $|\psi_{tik}(\theta^0)| < \Phi^k$ pour un $\Phi < 1$ $\psi_{tik}(\theta) = \theta_1^{N_{t,k-1}} \theta_2^{k-1-N_{t,k-1}}$, où $N_{t,k-1} = \sum_{\ell=1}^{k-1} I_{t-1-\ell}$ est le numbre d'occurrences du premier régime entre $t-1$ and $t-k+1$. Notons $\Phi(\theta) = \max(|\theta_1|, |\theta_2|)$ et $\Phi = \Phi(\theta^0)$. Dès lors une condition suffisante mais non neécessaire est que $\max(|\theta_1^0|, |\theta_2^0|) < 1$. Notons que cette condition est plus restrictive que celle de Petruccelli et Woolford (1984) qui est $\theta_1^0 < 1$, $\theta_2^0 < 1$, et $\theta_1^0 \theta_2^0 < 1$. Mais l'existence de la matrice de covariance asymptotique V doit être supposée. Il est possible de traiter des modèles TARMA(p, q) mais on doit employer la paragraphe 2.2 et on doit tenir compte que les ψ_{tk} et les ψ_{tik} sont aléatoires donc l'application de AALM n'est pas directe. Voir Azrak et Mélard (2020) pour plus de détails.

4 Analyse d'un modèle pour données de comptage

On suppose que les $x_t \in \mathbf{N}$ sont i.i.d. avec une loi de Poisson(λ_t), où les λ_t sont elles-mêmes des variables aléatoires satisfaisant l'équation

$$\log(\lambda_t + 1) = \theta_1 \log(\lambda_{t-1} + 1) + \epsilon_t + \theta_2 \epsilon_{t-1}, \tag{8}$$

(θ_1 et θ_2 sont des paramètres et les ϵ_t's sont comme dans le paragraphe 3). Le modèle log-linéaire pour données de comptage de Liboschik *et al.* (2017) est illustré par une équation similaire mais différente

$$\log(\lambda_t) = \theta_0 + \theta_1 \log(x_{t-1} + 1) + \theta_2 \log(\lambda_{t-1}). \tag{9}$$

Une première différence de (9) par rapport à (8) est que le logarithme de λ_t est employé, pas celui de $\lambda_t + 1$. Dans notre cas nous employons les observations x_{t-i} au lieu de λ_{t-i}, $i = 0, 1$ dans (8), de sorte que les paramètres sont estimés en employant le modèle suivant :

$$\log(x_t + 1) = \theta_1 \log(x_{t-1} + 1) + \epsilon_t + \theta_2 \epsilon_{t-1}.$$

Une seconde différence est qu'une quasi-vraisemblance gaussienne est employée ici au lieu d'une quasi-vraisemblance de Poisson. Ainsi, pour n grand, la fonction objectif apparaît comme une somme de carrés de résidus. On peut donc employer la théorie de AALM et utiliser les propriétés asymptotiques d'un modèle ARMA ou VARMA standard. On en déduit les résultats asymptotiques pour $\hat{\theta}_n$ pourvu que les vraies valeurs de θ satisfont $|\theta_1^0| < 1$ et $|\theta_2^0| < 1$. Contrairement au modèle TAR(1), l'existence de V est évidente ici. Contrairement au cas d'une quasi-vraisemblance de Poisson, c'est fait à nouveau, comme pour le modèle TAR(1), sans avoir besoin de trouver une solution et sans montrer qu'elle est stationnaire et sans prouver l'ergodicité. Ceci a bien été réalisé pour le modèle log-linéaire (9) mais pas pour des modèles d'ordre supérieur. Au contraire, notre approche avec une quasi-vraisemblance gaussienne ne pose pas de problème pour des modèles d'ordre supérieur sous des hypothèses très générales, essentiellement sur les coefficients des polynômes. Voir Azrak et Mélard (2020) pour plus de détails. Des extensions peuvent aussi être traitées sans problème.

Bibliographie

Alj, A., Azrak, R., Ley, C. and Mélard, G. (2017). Asymptotic properties of QML estimators for VARMA models with time-dependent coefficients, *Scandinavian Journal of Statistics* 44, pp. 617–635.

Azrak, R. and Mélard, G. (2020). Asymptotic properties of conditional least-squares estimators for array time series, ECARES Working paper 2020-12, Université libre de Bruxelles.

Francq, C. and Gautier, A. (2004). Estimation of time-varying ARMA models with Markovian changes in regime, *Statistics and Probability Letters* 70, pp. 243–251.
Hannan, E. J. and Deistler, M. (1988). *The Statistical Theory of Linear Systems*, John Wiley and Sons, New York.
Liboschik, T., Fokianos, K., and Fried, R. (2017). tscount : An R package for analysis of count time series following generalized linear models, *Journal of Statistical Software* 82 (5).
Mélard, G. (2020). An indirect proof for the asymptotic properties of VARMA model estimators, ECARES Working paper 2020-10, Université libre de Bruxelles.
Petruccelli, J.D. and Woolford, S.W. (1984). A threshold AR(1) model, *Journal of Applied Probability* 21, pp. 270–286.
Ruiz, S. M. (1996). An algebraic identity leading to Wilson's theorem, *The Mathematical Gazette* 80 (489), pp. 579–582.
Tong, H. (1990). *Nonlinear Time Series : A Dynamical Systems Approach*, Oxford University Press, Oxford.
Yao, Q. and Brockwell, P. J. (2006). Gaussian maximum likelihood estimation for ARMA models. I. Time series, *Journal of Time Series Analysis* 27, pp. 857–875.

Appendice : les principales hypothèses de AALM

Nous utilisons les coefficients $A_{ti}(\theta)$ et $B_{tj}(\theta)$, les résidus $e_t(\theta)$ et les vraies erreurs ϵ_t du modèle (1), et les notations introduites dans le paragraphe 2 : $\psi_{tik} = \psi_{tik}(\theta^0)$ dans (5) et $\alpha_t(\theta)$ dans (2).

On suppose (i) que les $A_{ti}(\theta)$ et $B_{tj}(\theta)$ sont des fonctions trois fois continûment différentiables de θ, (ii) que les ϵ_t ont des moments d'ordre $4+\delta$, $\delta > 0$ et (iii) l'existence de bornes comme les suivantes $\sum_{k=\nu}^{t-1} \|\psi_{tik}\|_F^2 < N_1 P(\nu)\Phi^{\nu-1}$, $\sum_{k=\nu}^{t-1} \|\psi_{tik}\|_F^4 < N_2 P(\nu)\Phi^{\nu-1}$, $i = 1, ..., m$, $\nu = 1, ..., t-1$, où N_1 et N_2 sont des constantes positives, $P(\nu)$ est un polynôme en ν, et $0 < \Phi < 1$. Ensuite on suppose (iv) l'existence de la matrice strictement définie positive V, telle que $V = \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} V_t^{(n)}$, où les éléments de $V_t^{(n)}$ sont donnés par $E_{\theta^0}((\partial e_t^T(\theta)/\partial\theta_i)\Sigma_t^{-1}(\theta)(\partial e_t(\theta)/\partial\theta_j))$, $i, j = 1, ..., m$. Enfin, on suppose (v) que deux triples sommes sont $O(1/n)$ dont la première est $\frac{1}{n^2} \sum_{d=1}^{n-1} \sum_{t=1}^{n-d} \sum_{k=1}^{t-1} \|\psi_{tik}\|_F \|\psi_{t+d,i,k+d}\|_F$.

Notons que dans AALM, on remplace $e_t(\theta)$ par $g_t(\theta)e_t(\theta)$ où $g_t(\theta)$ est une matrice avec des éléments qui sont des fonctions déterministes du temps et de θ. Dans ce cas, outre des bornes sur $\Sigma_t(\theta)$, $\Sigma_t^{-1}(\theta)$ et leurs dérivées en θ^0, on doit supposer l'existence de la matrice définie positive W telle que $W = \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} W_t^{(n)}$, où les éléments de $W_t^{(n)}$ sont donnés par $\frac{1}{4} E_{\theta^0}((\partial\alpha_t(\theta)/\partial\theta_i)(\partial\alpha_t(\theta)/\partial\theta_j))$, pour $i, j = 1, ..., m$. On doit alors remplacer la matrice de covariance asymptotique V^{-1} par $V^{-1}WV^{-1}$ mais dans les conditions du présent article on a $W = V$.

Optimal transport-based machine learning sheds light on Huntington's disease

Thi Thanh Yen Nguyen [1,2] & Warith Harchaoui [1,2,3] & Olivier Bouaziz [1,2] & Christian Néri [4] & Antoine Chambaz [1,2]

[1] *MAP5 (UMR CNRS 8145), Université de Paris*
[2] *Fédération Parisienne de Modélisation Mathématique (FR CNRS 2036)*
[3] `Oscaro.com`
[4] *BrainC-lab, IBPS (UMR CNRS 8256), Biology of Adaptation & Aging, Sorbonne Université*

Email: thi-thanh-yen.nguyen@parisdescartes.fr

Résumé. Nous présentons un nouvel algorithme de co-clustering dans le but d'apprendre le type de correspondance qui relie deux jeux de données avec la contrainte d'associer entre eux les éléments qui expriment une relation symétrique. L'objectif final est de décrire les interactions qui existent entre l'ARN messager et le micro ARN de souris atteintes de la maladie de Hutington.

L'algorithme fonctionne en deux étapes. Lors de la premiére étape, un plan de transport optimal P est appris. L'algorithme utilisé repose sur un contraste de type Sinkhorn et une descente de gradient stochastique. Lors de la deuxiéme étape, un algorithme de co-clustering spectral est appliqué (deux fois) à l'estimateur de la matrice de couplage optimale P et à une sous-matrice de P, afin d'en déduire les co-clusters finaux.

Une simple étude de simulation basée sur des données réelles semble indiquer que l'algorithme fonctionne bien. L'application aux données réelles est en cours.

Mots-clés. Algorithme de Sinkhorn; contraste de Sinkhorn; co-clustering spectral; génomique; maladie de Huntington; transport optimal.

Abstract.

We present a novel co-clustering algorithm for learning a pattern of correspondence between two data sets in situations where it is desirable to match elements that exhibit a mirroring relationship. The ultimate objective is to shed light on the interaction between mRNAs and miRNAs in Huntington's Disease based on mouse data.

The algorithm unfolds in two stages. During the first stage, an optimal transport plan P is learned. The corresponding algorithm relies on the Sinkhorn loss and a stochastic gradient descent. During the second stage, a spectral co-clustering algorithm is applied (twice) to the estimator of the optimal coupling matrix P and to a submatrix of P, in order to derive the final co-clusters.

A simple simulation study based on real data seems to indicate that the algorithm works well. The application to real data is ongoing.

Keywords. Genomics; Huntington's Disease; optimal transport; Sinkhorn algorithm; Sinkhorn loss; spectral co-clustering.

1 Introduction

The discriminative and biologically-precise analysis of high dimensional genomic data in biology and disease represents a challenging task in systems modeling and systems biology. Messenger RNAs (mRNAs) and micro RNAs (miRNAs) have been studied in model organisms and in cases of Huntington's Disease (HD) [6, 7], extensively but separately. Little is known about their interplay. The main objective of our study is to shed light on the interaction between mRNAs and miRNAs based on HD mouse data.

The data set was obtained by using deep mRNA sequencing (RNA-seq). The striatum of 2-month, 6-month and 10-month old mice that express one wildtype endogenous Htt allele and a second Htt allele with knock-in of human mHTT exon 1 carrying one of six different CAG lengths (Q20, Q80, Q92, Q111, Q140, Q175) were profiled. This resulted in $M = 19,051$ mRNA profiles, $X = \{x_1, \ldots, x_M\} \subset \mathbb{R}^{18}$, and in $N = 1,507$ miRNA profiles, $Y = \{y_1, \ldots, y_N\} \subset \mathbb{R}^{18}$.

It is known that the miRNAs and their target-mRNAs exhibit a many-to-many mirroring relationship. More specifically, it is expected that if a miRNA targets a mRNA, inducing the degradation of that mRNA, then the profile y_n of the former should be similar to minus the profile x_m of the latter. We will clarify what we mean by "similar", and will rely on this property. Figures 1 and 2 exhibit two profiles x_m and y_n that showcase a mirrored similarity. The corresponding miRNA and mRNA are believed to interact.

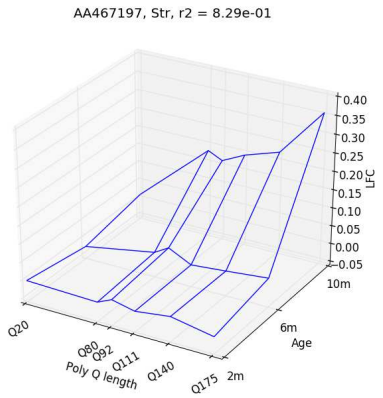


Figure 1: Profile x_m of a mRNA. It is believed that the mRNA is targeted by the miRNA from Figure 2.

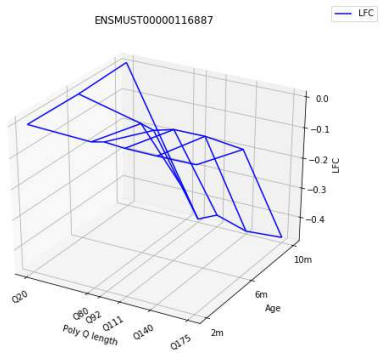


Figure 2: Profile y_n of a miRNA. It is believed that the miRNA targets the mRNA from Figure 1.

We look for mutually disjoint $I_1, \ldots, I_R$ subsets of $[\![M]\!] = \{1, \ldots, M\}$ and mutually disjoint $J_1, \ldots, J_R$ subsets of $[\![N]\!]$ such that, for all $r \in [\![R]\!]$, each x_m with $m \in I_r$ interacts with every y_n with $n \in J_r$. To reach this goal, we develop a co-clustering algorithm based on optimal transport [9] and spectral co-clustering (SCC) [3].

SCC is one way among others to carry out co-clustering, an unsupervised learning task to cluster simultaneously the rows and columns of a matrix in order to obtain homogeneous blocks. There are many efficient approaches to solving the problem, often characterized as model-based or metric-based methods [10]. The co-clustered matrix is derived from the data by optimal transport. We give some details in the next section.

2 Elements of optimal transport

Let $\mu_X = M^{-1} \sum_{m=1}^{M} \delta_{x_m}$ and $\nu_Y = N^{-1} \sum_{n=1}^{N} \delta_{y_n}$ be the empirical measures attached to X and Y. The celebrated Monge-Kantorovich problem [9, Chapter 2] consists in finding a joint law P over $X \times Y$ with marginals μ_X and ν_Y that minimizes the expected cost of transport with respect to some cost function $c : X \times Y \to \mathbb{R}_+$. We focus on c given by $c(x, y) = \|x-y\|_2^2$ (the squared Euclidean norm in $\mathbb{R}^{18}$). Specifically, denoting $\Pi(\mu_X, \nu_Y) = \{P \in \mathbb{R}_+^{M \times N} | P\mathbf{1}_N = \mu_X, P^\top \mathbf{1}_M = \nu_Y\}$ and $C \in \mathbb{R}^{M \times N}$ such that $C_{mn} = c(x_m, y_n)$ for each $(m, n) \in [\![M]\!] \times [\![N]\!]$, then the problem consists in solving $\min_{P \in \Pi(\mu_X, \nu_Y)} \langle C, P \rangle_F$ where $\langle C, P \rangle_F = \sum_{(m,n) \in [\![M]\!] \times [\![N]\!]} C_{mn} P_{mn}$. It is well known that it is very rewarding from a computational viewpoint to consider a regularized version of the above problem [9, Chapter 4]. The penalty term is proportional to the discretized entropy of P, that is, to $E(P) = -\sum_{(m,n) \in [\![M]\!] \times [\![N]\!]} P_{mn}(\log P_{mn} - 1)$, and the new problem consists, for any $\gamma > 0$, in finding P_γ that solves

$$\mathcal{W}_\gamma (\mu_X, \nu_Y) = \min_{P \in \Pi(\mu_X, \nu_Y)} \{\langle C, P \rangle_F - \gamma E(P)\} . \tag{1}$$

One of the advantages of entropic regularization is that one can solve (1) efficiently using the Sinkhorn-Knopp matrix scaling algorithm.

Finally, following [4], we use $\mathcal{W}_\gamma$ to define the so called Sinkhorn loss between μ_X and ν_Y as

$$\bar{\mathcal{W}}_\gamma (\mu_X, \nu_Y) = 2\mathcal{W}_\gamma (\mu_X, \nu_Y) - \mathcal{W}_\gamma (\mu_X, \mu_X) - \mathcal{W}_\gamma (\nu_Y, \nu_Y) .$$

This loss interpolates between $\mathcal{W}_0 (\mu_X, \nu_Y)$ and the maximum mean discrepancy of μ_X relative to ν_Y [4, Theorem 1].

3 Optimal transport-based machine learning

We introduce a parametric model Θ consisting of linear mappings $\theta : \mathbb{R}^{18} \to \mathbb{R}^{18}$ of the form $x \mapsto \theta(x) = \theta_1 x + \theta_2$, where $\theta_1 \in \mathbb{R}^{18 \times 18}$ and $\theta_2 \in \mathbb{R}^{18}$. Each $\theta \in \Theta$ is a candidate

to formalize the aforementioned mirroring relationship. The set Θ imposes constraints on the matrices θ_1, in particular that their diagonals are made of negative values. The parametrization is identifiable, in the sense that $\theta = \theta'$ implies $(\theta_1, \theta_2) = (\theta'_1, \theta'_2)$. The first program that we introduce is

$$\min_{\theta \in \Theta} \bar{\mathcal{W}}_\gamma \left(\mu_{\theta(X)}, \nu_Y \right), \tag{2}$$

where $\mu_{\theta(X)}$ is the empirical measure attached to $\theta(X) = \{\theta(x_1), \ldots, \theta(x_M)\}$ and we are interested in the minimizer $\hat{\theta}$ and the optimal joint matrix $\hat{P} \in \Pi(\mu_{\hat{\theta}(X)}, \nu_Y)$ that solves Equation (1) with the measures $\mu_{\hat{\theta}(X)}$ and ν_Y.In words, we look for an optimal mirroring function $\hat{\theta}$ and its optimal transport plan $\hat{P}$.

To acknowledge the fact that we do not expect to associate a y_n to every x_m eventually at the co-clustering stage, we build upon (2) to define our main program. The key addition is merely the introduction of a weighted version of the empirical measures $\mu_{\theta(X)}$ ($\theta \in \Theta$). Let $\Omega = \{\omega \in \mathbb{R}_+^M | \sum_{m \in [\![M]\!]} \omega_m = 1\}$ be the $(M-1)$-dimensional simplex. Moreover, for each $\omega \in \Omega$, let $\mu^\omega_{\theta(X)}$ be the ω-weighted empirical measure attached to $\theta(X)$, that is, $\mu^\omega_{\theta(X)} = \sum_{m \in [\![M]\!]} \omega_m \delta_{\theta(x_m)}$. Our main program is

$$\min_{\omega \in \Omega} \min_{\theta \in \Theta} \bar{\mathcal{W}}_\gamma \left(\mu^\omega_{\theta(X)}, \nu_Y \right). \tag{3}$$

Here, we are interested in the minimizer $(\hat{\omega}, \hat{\theta})$ and the optimal matrix $\hat{P} \in \Pi(\mu^{\hat{\omega}}_{\hat{\theta}(X)}, \nu_Y)$ solves Equation (1) with the measures $\mu^{\hat{\omega}}_{\hat{\theta}(X)}$ and ν_Y.

We propose to solve (3) iteratively by updating ω and then (θ, P). Given ω, we exploit the Sinkhorn algorithm. Given (θ, P), the updated ω is proportional to the vector in $\mathbb{R}_+^M$ of which the mth component equals $h^{-1} \sum_{n \in [\![N]\!]} \varphi((y_n - \theta(x_m))/h)$ where φ is the standard normal density and h is the arithmetic mean of the $c(y_n, y_{n'})$ for all $n \neq n' \in [\![N]\!]$.

Let $\tilde{P}$ be the approximation of $\hat{P}$ derived above. To co-cluster the mRNAs and miRNAs, we co-cluster $\tilde{P}$. We apply SCC a first time. Then we remove the rows and columns that correspond to diagonal blocks with a variance larger than the overall variance of $\tilde{P}$, because we identify the corresponding mRNAs and miRNAs as irrelevant. Then we apply SCC a second time. The resulting co-clusters should reveal the interplay between the remaining mRNAs and miRNAs in HD.

Our code is written in `python`. We adapt the Sinkhorn algorithm implemented by Aude Genevay and available here. The stochastic gradient descents relies on the machine learning framework `pytorch`. We use the implementation of SCC available in the `sklearn` module. The function requires that the number of co-clusters be given as an input parameter. We use the modularity measure [1, Sections 2 and 4], available in the `coclust` module. In the future, we will also look into the Integrated Completed Likelihood criterion [2].

Our algorithm bears a similarity to the one developed in [5]. The main differences are *(i)* our use of the parametric model Θ and weights ω, *(ii)* the fact that we apply

SCC to the approximation of the optimal transport matrix $\tilde{P}$. Our algorithm also bears a similarity to [11], a fast and certifiable point cloud registration algorithm. In a near future, we will study the similarities and differences closely.

4 Simulation study

To assess the performances of the algorithm sketched in Section 3, we conduct a simple simulation study. The synthetic law uses real observations, an explicit (linear) mirroring relationship (see θ in Section 3), a random number of elements in each co-cluster, a source of noise that is centered Gaussian (with different variances). Some of the synthetic x_m and y_n are expected to be left out of all co-clusters eventually. To validate the obtained results, we used the co-clustering error [8]. The results are satisfactory.

5 Conclusion

We develop an optimal transport-based machine learning co-clustering algorithm to match data points in a vector space (here, $\mathbb{R}^{18}$). The algorithm takes into account a geometric property that is known to be helpful to identify matching data points and can be formalized as a linear function. The application to real data is currently ongoing, where the points represent mRNA and miRNA activity in models of mice for HD, and the objective is to learn which collections of miRNAs interact with which collections of mRNAs.

References

[1] Melissa Ailem, François Role, and Mohamed Nadif. Graph modularity maximization as an effective method for co-clustering text data. *Knowledge-Based Systems*, 109: 160–173, 2016.

[2] Vincent Brault, Christine Keribin, Gilles Celeux, and Gerard Govaert. Estimation and selection for the latent block model on categorical data. *Statistics and Computing*, 25:1201–1216, 2014.

[3] Inderjit S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '01, page 269–274, New York, NY, USA, 2001. Association for Computing Machinery.

[4] Aude Genevay, Gabriel Peyre, and Marco Cuturi. Learning generative models with sinkhorn divergences. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and*

Statistics, volume 84 of *Proceedings of Machine Learning Research*, pages 1608–1617, Playa Blanca, Lanzarote, Canary Islands, 09–11 Apr 2018. PMLR.

[5] Charlotte Laclau, Ievgen Redko, Basarab Matei, Younès Bennani, and Vincent Brault. Co-clustering through Optimal Transport. In *34th International Conference on Machine Learning*, volume 70, pages 1955–1964, Sydney, Australia, August 2017.

[6] Peter Langfelder et al. Integrated genomics and proteomics define Huntingtin CAGlength–dependent networks in mice. *Nature Neuroscience*, 19, 02 2016.

[7] Peter Langfelder et al. MicroRNA signatures of endogenous Huntingtin CAG repeat expansion in mice. *PloS one*, 13(1), 2018.

[8] Anne Patrikainen and Marina Meila. Comparing subspace clusterings. *Knowledge and Data Engineering, IEEE Transactions on*, 18:902–916, 2006.

[9] G. Peyre and M. Cuturi. *Computational Optimal Transport: With Applications to Data Science*. Foundations and Trends in Machine Learning Series. Now Publishers, 2019.

[10] Beatriz Pontes, Raúl Giráldez, and Jesús S. Aguilar-Ruiz. Biclustering on expression data: A review. *Journal of Biomedical Informatics*, 57:163–180, 2015.

[11] Heng Yang, Jingnan Shi, and Luca Carlone. Teaser: Fast and certifiable point cloud registration, 2020. arXiv:2001.07715.

Bornes post hoc dans un modèle de Markov caché

Pierre Neuvial[1] & Marie Perrot-Dockès [2] & Etienne Roquain [3]

[1] *Institut de Mathématiques de Toulouse, UMR 5219, Université de Toulouse, CNRS. UPS IMT, F-31062 Toulouse Cedex 9, France. E-mail : pierre.neuvial@math.univ-toulouse.fr.*

[2] *Université Paul Sabatier, Institut Mathématiques de Toulouse & Sorbonne Université, Laboratoire de Probabilités, Statistique et Modélisation, LPSM, 4, Place Jussieu, 75252 Paris cedex 05, France. E-mail : marie.perrot-dockes@upmc.fr.*

[3] *Sorbonne Université, Laboratoire de Probabilités, Statistique et Modélisation, LPSM, 4, Place Jussieu, 75252 Paris cedex 05, France. E-mail : etienne.roquain@upmc.fr.*

Résumé. Dans les données contemporaines, de nombreuses questions ou hypothèses émergent après avoir « regardé » les données. Ceci pousse naturellement le praticien à sélectionner des variables d'intérêt, selon différents critères, souvent en utilisant les données plusieurs fois, puis à lancer des procédures d'analyse statistique classiques sur les variables sélectionnées. Cependant, en raison du biais de sélection, les garanties statistiques usuelles ne sont plus valables. Les bornes « post hoc » permettent de répondre à ce problème, en fournissant une borne de confiance sur le nombre de faux positifs quelle que soit la méthode de sélection employée par l'utilisateur (Genovese et Wasserman, 2006; Goeman et Solari, 2011). Ces bornes sont, par nature, assez conservatrices et un enjeu majeur est de développer de nouvelles bornes plus performantes, en ajoutant par exemple des contraintes de structure sur le modèle. Cette note présente de nouvelles bornes post hoc adaptées au cas où les données ont une structure linéaire, dans laquelle le signal est localisé à des positions contiguës. À cette fin, le processus de présence/absence du signal est modélisé par une chaîne de Markov cachée. Dans ce modèle, nous établissons la validité théorique de nouvelles bornes post hoc et nous montrons qu'elles améliorent l'état de l'art à l'aide d'expériences numériques.

Mots-clés. Inférence post hoc, modèle de Markov caché, faux positifs, tests multiples.

Abstract. Selective inference aims at providing guarantees for statistical procedures performed after a selection step. In particular, a task of primary interest is to build post hoc bounds, that is, confidence bounds for the number of false positives, with a coverage probability being valid uniformly over all the selected sets. Since such bounds are conservative by construction, a crucial aim is to build new post hoc bounds with better performances. In this note, we do so by incorporating a hidden Markov model structure into the bounds. We prove the theoretical validity of the new derived bounds and show that the improvement can be significant via numerical experiments.

Keywords. Post hoc inference, hidden Markov model, false positives, multiple testing.

1 Motivation

Nous observons une série de m mesures $X_1, \ldots, X_m \in \mathbb{R}$ dont la loi dépend d'une série de configurations inconnues $\theta_1, \ldots, \theta_m \in \{0, 1\}$, dans laquelle $\theta_i = 0/1$ code pour l'absence/la présence de « signal » à la position i. L'objectif de ce travail est de fournir, à partir de X et pour chaque ensemble de sélection $S \subseteq \mathbb{N}_m := \{1, \ldots, m\}$, une borne supérieure de confiance de niveau $1 - \alpha$, notée $V_\alpha(X, S)$, pour la quantité $\sum_{i \in S}(1 - \theta_i)$, qui correspond au nombre de faux positifs dans l'ensemble de sélection S. On recherche de plus une borne « post hoc », c'est-à-dire que la probabilité de couverture doit être assurée uniformément par rapport à l'ensemble de sélection S.

Cette situation recouvre plusieurs exemples pratiques. Ici, notre motivation principale est l'analyse de données de nombre de copies d'ADN en cancérologie. Dans ce cas, X_i peut être une mesure différentielle de la quantité d'ADN présente à la position i entre deux types de cancer. On notera ainsi $\theta_i = 0$ (resp. 1) si le nombre de copies d'ADN est identique (resp. différent) pour les deux types de cancer à la position i. On distingue les deux objectifs pratiques suivants :

— Trouver les positions où les deux types de cancer présentent une différence ;
— Pour un ensemble S de positions candidates, majorer le nombre d'erreurs dans S, c'est-à-dire le nombre de positions dans S où le nombre de copies est le même.

Nous poursuivons ici le second objectif. Notons que ce dernier est également lié au premier objectif : étant donnée une certaine borne $V_\alpha(X, S)$, une procédure de sélection peut consister à choisir un ensemble S de positions candidates pour lesquelles il y a une quantité d'erreurs acceptable, par exemple $V_\alpha(X, S)/|S| \leq 5\%$.

Dans cet exemple, les $\theta_1, \ldots, \theta_m$ présentent une forte structure locale : si le nombre de copies d'ADN est différent à la position i (i.e., $\theta_i = 1$), alors il y a davantage de chance qu'il le soit également à la position $i + 1$ (i.e., $\theta_{i+1} = 1$). Ainsi, il est naturel de supposer la suite $\theta_1, \ldots, \theta_m$ constante par morceaux, avec un petit nombre de « morceaux ».

2 Modèle

Les modèles de Markov cachés sont très utilisés pour modéliser une structure de dépendance uni-dimensionelle, voir par exemple Rabiner (1989); Robin *et al.* (2005), et notamment pour modéliser les données de nombre de copies d'ADN, voir par exemple Gassiat *et al.* (2013); Luong *et al.* (2013). Notre cadre de travail s'inspire largement de celui de Sun et Cai (2009), qui fut utilisé pour un objectif différent (tests multiples). Nous supposerons ainsi que $\theta = (\theta_1, \ldots, \theta_m)$ sont les états d'une chaîne de Markov de matrice de transition

$$A = \begin{pmatrix} a_{0,0} & a_{0,1} \\ a_{1,0} & a_{1,1} \end{pmatrix},$$

avec des probabilités de transition $a_{q,\ell} \in (0, 1)$. La matrice de transition possède une unique distribution stationnaire sur $\{0, 1\}$, de probabilités d'émission notés (π_0, π_1) et nous

supposerons qu'il s'agit de la probabilité initiale de la chaîne, c'est-à-dire, $\mathbb{P}(\theta_1 = q) = \pi_q$ pour $q \in \{0, 1\}$. Conditionnellement à θ, nous supposerons que les variables $X_1, \ldots, X_m \in \mathbb{R}$ sont mutuellement indépendantes, avec des probabilités marginales données par

$$\mathcal{L}(X_i \,|\, \theta) = P_{\theta_i}, \quad 1 \leq i \leq n,$$

où $\{P_0, P_1\}$ sont deux lois différentes sur $\mathbb{R}$. Nous considérons une approche asymétrique dans l'esprit des tests, dans laquelle la distribution P_0 est supposée connue alors que la distribution P_1 est inconnue. Les paramètres du modèle sont donc donnés par $\vartheta = (A, P_1)$. La structure de dépendance du modèle est illustrée par la Figure 2.

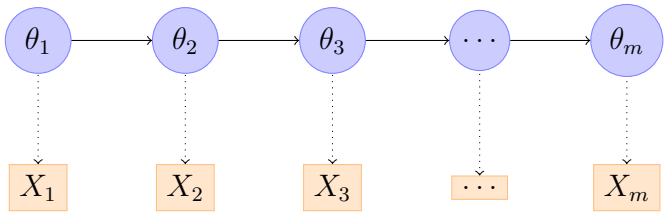


FIGURE 1 – Les dépendances au sein du modèle de Markov caché.

3 Nouvelles bornes post hoc

Formellement, l'objectif est de trouver une fonction $V_\alpha(X, \cdot) : S \subseteq \mathbb{N}_m \mapsto V_\alpha(X, S) \in \mathbb{N}$ telle que

$$\forall \vartheta, \qquad \mathbb{P}_\vartheta\Big(\forall S \subseteq \mathbb{N}_m, \sum_{i \in S}(1 - \theta_i) \leq V_\alpha(X, S)\Big) \geq 1 - \alpha, \tag{1}$$

Notons que lorsque (1) est satisfaite, nous avons pour n'importe quelle procédure de sélection $S(X)$, potentiellement dépendant des données, que l'inégalité $\sum_{i \in S(X)}(1 - \theta_i) \leq V_\alpha(X, S(X))$ est valable avec probabilité au moins $(1 - \alpha)$. Ainsi, comme annoncé en introduction, (1) fournira une borne de confiance pour le nombre de faux positifs de n'importe quelle procédure de sélection, calibrée en utilisant les mêmes données, autant de fois que l'utilisateur le souhaite.

3.1 Famille de référence

Pour établir (1) nous suivons la démarche proposée dans Blanchard *et al.* (2020); Durand *et al.* (2020) basée sur une famille de référence $(R_k)_{1 \leq k \leq K}$ et des bornes $(\zeta_k)_{1 \leq k \leq K}$ (pour $K \geq 1$), vérifiant le contrôle joint

$$\forall \vartheta, \qquad \mathbb{P}_\vartheta\Big(\forall k \in \{1, \ldots, K\}\,, \sum_{i \in R_k}(1 - \theta_i) \leq \zeta_k\Big) \geq 1 - \alpha, \tag{2}$$

c'est-à-dire, satisfaisant une contrainte de type (1), mais en restriction aux ensembles de référence $(R_k)_{1\leq k\leq K}$. Au vu de la structure supposée sur le signal, nous choisissons pour la famille de référence une partition régulière :

$$R_k = \{1 + (k-1)m/K, \ldots, km/K\}, \quad 1 \leq k \leq K. \tag{3}$$

(nous avons supposé que K divise m pour simplifier les notations.) Par suite, une simple étape d'interpolation permet de déduire (1) à partir de (2), en posant

$$V_\alpha(X, S(X)) = \sum_{k=1}^{K} |S \cap R_k| \wedge \zeta_k, \quad S \subseteq \mathbb{N}_m. \tag{4}$$

En effet, on a pour tout S, $\sum_{i\in S}(1-\theta_i) = \sum_{k=1}^{K} |S \cap R_k| \wedge \sum_{i\in S\cap R_k}(1-\theta_i)$ ce qui est plus petit ou égal à la borne annoncée avec probabilité au moins $(1-\alpha)$ d'après (2).

3.2 Choix des ζ_k

Pour obtenir une borne post hoc, il reste ainsi à calibrer les bornes $(\zeta_k)_{1\leq k\leq K}$ de sorte que (2) soit vérifiée. Pour cela, nous allons simplement utiliser une borne d'union et la règle de Bayes :

$$\mathbb{P}_\vartheta\Big(\exists k \in \{1, \ldots, K\}\,,\ \sum_{i\in R_k}(1-\theta_i) > \zeta_k\Big) \leq \sum_{k=1}^{K} \mathbb{E}_\vartheta\Big[\mathbb{P}_\vartheta\Big(\sum_{i\in R_k}(1-\theta_i) > \zeta_k \mid X\Big)\Big],$$

Ce qui conduit au choix

$$\zeta_k = \zeta_k(X;\vartheta) = \min\Big\{n \in \{1, \ldots, |R_k|\} \ :\ \mathbb{P}_\vartheta\Big(\sum_{i\in R_k}(1-\theta_i) > n \,\Big|\, X\Big) \leq \alpha/K\Big\}. \tag{5}$$

Classiquement, la loi de $(\theta_i)_{i\in\mathbb{R}_k}$ sachant X est celle d'une chaîne de Markov hétérogène (Rabiner, 1989). De plus les matrices de transition et la distribution initiale peuvent être déterminées de manière explicites en fonction des paramètres $\vartheta = (P_1, A)$ à l'aide d'un algorithme de type « forward-backward ». Par suite, calculer $\zeta_k = \zeta_k(X;\vartheta)$ revient à déterminer la loi du nombre d'occurrences d'un état dans une chaîne de Markov hétérogène. En adaptant les outils développés dans Robin *et al.* (2005) et notamment Roquain (2007) Section 7.2, nous pouvons calculer $\zeta_k(X;\vartheta)$ à l'aide d'une relation de récurrence.

3.3 Nouvelles bornes

En combinant (4) et (5), nous obtenons une nouvelle borne post hoc, notée V^*_{HMM}, qui satisfait (1). Cependant, elle utilise les paramètres $\vartheta = (A, P_1)$ qui sont généralement

inconnus. La borne post hoc V^*_{HMM} est ici appelée nouvelle borne « oracle ». Les paramètres A et P_1 peuvent être estimés à l'aide d'un algorithme EM, voir par exemple Sun et Cai (2009). Ici, nous combinons l'algorithme EM avec une estimation non-paramétrique de P_1 par une méthode à noyau (voir par exemple Robin *et al.*, 2007). La borne obtenue en remplaçant A et P_1 dans V^*_{HMM} par leur estimateurs respectifs est notre borne post hoc finale, que l'on note V_{HMM}.

Notons que les fluctuations induites par cette étape d'estimation ne sont pas prises en compte ici. Ainsi, la borne V_{HMM} peut ne plus satisfaire (1). Cependant, nous pensons que cette garantie est maintenue, au moins asymptotiquement en m, si les estimateurs utilisés sont consistants. Par ailleurs, les expériences numériques de la prochaine section semblent indiquées que la couverture reste valable pour des m « modérés ».

4 Expériences numériques

Nous comparons ici nos nouvelles bornes V^*_{HMM} et V_{HMM} à la borne classique de Simes (Goeman et Solari, 2011) et celle, plus récente, basée sur l'inégalité de DKW (Durand *et al.*, 2020). Cette dernière utilise la même famille de référence que V^*_{HMM} et V_{HMM}, mais une calibration des ζ_k différente, qui n'utilise pas la structure markovienne. Notons que dans notre cadre, les deux bornes Simes et DKW satisfont (1).

Pour des raisons de place, nous ne reportons ici qu'un seul résultat de simulation dans le cas simple où les bornes sont toutes évaluées en $S = \mathbb{N}_m$. En d'autres termes, nous cherchons simplement une borne de confiance sur $\sum_{i=1}^{m}(1-\theta_i)$, le nombre de « non-signal » dans les données totales. Ici, les bornes V^*_{HMM}, V_{HMM} et DKW, sont construites à partir d'une famille de référence avec un seul élément $R_1 = \mathbb{N}_m$ (i.e., $K = 1$). Les résultats sont reportés Figure 4. Cette expérience montre que la borne V^*_{HMM} est en général moins conservatrice que les bornes Simes et DKW. La borne et V_{HMM} est un peu trop conservatrice mais est en moyenne moins éloignée du nombre de faux positifs réels. Nous espérons obtenir de meilleurs résultats en améliorant l'estimation de P_1. De plus, lorsque les estimateurs $\widehat{A}$ et $\widehat{P}_1$ sont correctement choisis, la borne V_{HMM} semble maintenir le contrôle (1). Nous compléterons le panorama des expériences numériques dans la présentation orale.

Références

Blanchard, G., Neuvial, P. et Roquain, E. (2020). Post hoc confidence bounds on false positives using reference families. *Annals of Statistics*. Preprint : hal-01483585.

Durand, G., Blanchard, G., Neuvial, P. et Roquain, E. (2020). Post hoc false positive control for structured hypotheses. *Scandinavian Journal of Statistics*. Preprint : hal-01829037.

Gassiat, E., Cleynen, A. et Robin, S. (2013). Finite state space non parametric hidden markov models are in general identifiable. Preprint : `https://arxiv.org/abs/1306.4657`.

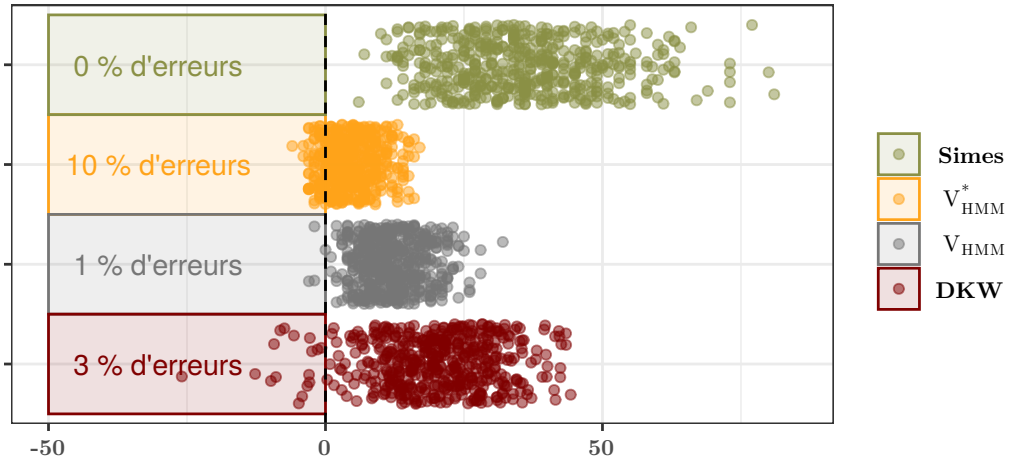


FIGURE 2 – Nuages de points représentant $V_\alpha(X, \mathbb{N}_m) - \sum_{i=1}^{m}(1-\theta_i)$ pour 500 réalisations de notre modèle pour quatre bornes post hoc. Les paramètres sont $m = 200$, $a_{0,0} = 0.95$, $a_{1,1} = 0.8$, $P_0 = \mathcal{N}(0, 1)$ et $P_1 = \mathcal{N}(2, 1)$.

GENOVESE, C. R. et WASSERMAN, L. (2006). Exceedance control of the false discovery proportion. *Journal of the American Statistical Association*, 101(476):1408–1417.

GOEMAN, J. J. et SOLARI, A. (2011). Multiple testing for exploratory research. *Statistical Science*, pages 584–597.

LUONG, T. M., ROZENHOLC, Y. et NUEL, G. (2013). Fast estimation of posterior probabilities in change-point analysis through a constrained hidden markov model. *Computational Statistics & Data Analysis*, 68:129–140.

RABINER, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

ROBIN, S., BAR-HEN, A., DAUDIN, J.-J. et PIERRE, L. (2007). A semi-parametric approach for mixture models : Application to local false discovery rate estimation. *Computational statistics & data analysis*, 51(12):5483–5493.

ROBIN, S., RODOLPHE, F. et SCHBATH, S. (2005). *DNA, words and models : statistics of exceptional words*. Cambridge University Press.

ROQUAIN, E. (2007). *Exceptional motifs in heterogeneous sequences. Contributions to theory and methodology of multiple testing*. Thèse de doctorat, Université Paris XI.

SUN, W. et CAI, T. T. (2009). Large-scale multiple testing under dependence. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 71(2):393–424.

Géométrie de la variété statistique gamma généralisée : application à la classification en neuroimagerie médicale et en transport aérien

Florence Nicol [1] & Stéphane Puechmorel [1] & Sana Rebbah [1 2]

[1] *Université de Toulouse, ENAC*
prenom.nom@enac.fr
[2] *Université de Toulouse,INSERM ToNIC*
sana.rebbah@inserm.fr

Résumé. En géométrie de l'information, les densités de probabilité peuvent être représentées comme des points d'une variété statistique munie d'une structure de variété riemannienne, la métrique étant généralement donnée par l'information de Fisher. Ce travail explore plus particulièrement de nouveaux résultats sur la variété des distributions de probabilité de type gamma généralisée, qui est particulièrement pertinente dans le contexte de la maladie d'Alzheimer et l'étude des retards de vol en transport aérien. Dans un premier temps, certaines propriétés géométriques de la variété gamma généralisée sont étudiées. Les résultats obtenus sur la métrique d'information de Fisher sont ensuite exprimés dans le contexte plus général des groupes de Lie. Une procédure de classification non supervisée a ensuite été adaptée en utilisant une distance géodésique dont une approximation est calculée numériquement avec un algorithme en deux étapes. Des résultats sont présentés pour des données réelles issues d'une population de patients atteints de la maladie d'Alzheimer et de patients sains, ainsi que pour pour des données de retards de vol en transport aérien.

Mots-clés. densités de probabilité, variété gamma généralisée, géométrie de l'information, distance géodésique, classification, maladie d'Alzheimer, retard de vol.

Abstract. Probability density functions may be represented as points of a statistical manifold using Information Geometry. Within this frame, densities are endowed with a Riemannian manifold structure, the metric being generally given by the Fisher information. The purpose is to present some new results about the generalized gamma manifold and how information geometry improved the performance of the classification of Alzheimer's disease population and flight delay distributions. This work investigates some information geometric properties of the generalized gamma family, that is particularly relevant in the Alzheimer's disease context and the air transport management. The Fisher information and results in the case of the generalized gamma manifold will be first detailed. Next, a clustering technique has been successfully extended by using a geodesic distance of which an approximation is numerically computed with a two steps algorithm.

Keywords. probability density functions, generalized gamma manifold, information geometry, geodesic distance, clustering, Alzheimer's disease, flight delay.

1 Introduction

Dans le domaine médical, des techniques d'imagerie médicale comme l'imagerie à résonace magnétique (IRM), présentent un intérêt majeur dans l'étude de maladies neurodégénératives comme la maladie d'Alzheimer [Cuingnet et al., 2011, Lama et al., 2017]. La maladie d'Alzheimer est la cause la plus commune de démence chez les personnes âgées. Elle affecte le fonctionnement du système nerveux central par des dysfonctionnements génétiques ou métaboliques conduisant notamment à la mort de neurones. Cette dégénérescence du cortex cérébral s'aggrave au cours du temps entraînant une altération progressive des capacités cognitives au cours de laquelle l'individu passe de sujet sain (vieillissement normal) vers la forme démentielle de la maladie, en passant par une phase de transition incluant des troubles cognitifs légers. Cette maladie entraînant une perte neuronale, l'atrophie cérébrale induite est donc un marqueur potentiel d'évolution vers cette maladie que la neuro-imagerie anatomique par IRM est capable de déceler indirectement en mesurant l'épaisseur corticale sur l'ensemble du cerveau [Busovaca et al., 2016].

Dans le but de détecter des modifications de l'activité cérébrale, des indicateurs statistiques de tels biomarqueurs peuvent être implémentés dans des algorithmes de classification. Usuellement, des indicateurs statistiques de mesure de tendance centrale, comme la moyenne ou le mode, sont utilisés. Cependant, ces indicateurs ne permettent pas d'exploiter toute l'information contenue dans la distribution des données. Dans des études plus récentes [Cercignani et al., 2001], des histogrammes ont été utilisés comme approximation des densités de probabilité. Mais seul un nombre limité de caractéristiques est extrait des histogrammes puis implémenté dans des analyses statistiques. Par ailleurs, cette approximation peut être de mauvaise qualité et peu robuste au choix des classes des histogrammes.

Comme alternative, nous proposons d'utiliser les densités de probabilité elles-mêmes en tant que biomarqueurs de l'activité globale du cerveau. Dans ce contexte, la géométrie de l'information fournit un cadre de travail particulièrement intéressant, dans lequel les distributions de probabilité sont considérées comme des points dans une variété riemannienne. Dans ce nouvel espace de représentation, la métrique d'information de Fisher, une métrique riemannienne donnée par l'information de Fisher, peut être implémentée dans un algorithme de classification. Dans ce travail, nous proposons d'étudier la géométrie de la variété de la famille des distributions de type gamma généralisée. Ce choix est particulièrement pertinent dans les deux applications étudiées : l'une en transport aérien pour l'étude de données de retards de vol et l'autre en neuroimagerie médicale pour l'étude d'une population de patients atteints de la maladie d'Alzheimer et de patients sains issue de la base de données ADNI, Alzheimer's Disease Neuroimaging Initiative [Mueller et al., 2005].

2 Géométrie de la variété des distributions de type gamma généralisée

2.1 Introduction à la géométrie de l'information

Dans beaucoup d'applications, les données d'intérêt n'appartiennent pas à un espace vectoriel comme c'est le cas par exemple des distributions de probabilité d'une famille de modèles statistiques. La géométrie de l'information fournit alors un cadre de travail original et pertinent pour étudier de telles données, qui sont vues comme des points sur une variété différentielle de modèles statistiques [Amari and Nagaoka, 2007]. La métrique d'information de Fisher, donnée par l'information de Fisher, munit une telle variété d'une structure riemanienne [Amari, 2016] et plus particulièrement d'une structure hessienne dans le cas de modèles statistiques appartenant à la famille exponentielle naturelle [Duistermaat, 2001]. La géométrie de l'information a donc pour objet d'étude les variétés différentielles de modèles statistiques. Bien que lié par le même cadre de travail fourni par la géométrie riemannienne, il convient de distinguer ce domaine de la statistique géométrique qui traite de l'analyse statistique d'objets géométriques vivant dans des variétés différentielles.

La métrique riemanienne est un produit scalaire défini localement sur chaque espace tangent de la variété, qui permet de calculer des angles, des longueurs et des distances. Les géodésiques sont les chemins de longueur minimale (localement) entre deux points de la variété, ce qui permet de définir une distance, appelée distance géodésique, entre ces points. En géométrie de l'information, les distances géodésiques calculées entre deux points d'une variété statistique permettent donc de calculer les distances entre les distributions de probabilité de cette variété. Par exemple, la géométrie de la variété statistique des lois normales univariées est celle de la géométrie hyperbolique. Les géodésiques partant d'un point de la variété sont les demi-droites verticales et les demi-cercles perpendiculaires à l'axe des abscisses dans le demi plan supérieur. Suivre les géodésiques sur cette variété statistique permet donc de réaliser des interpolations optimales entre les distributions de la loi normale univariée.

Considérons une famille de distributions de probabilité $\mathcal{P} = \{P_\theta = p_\theta \mu | \theta \in \Theta\}$, absolument continue par rapport à une mesure fixée μ et paramétrisée par $\theta \in \Theta$, où Θ désigne l'espace des paramètres et p_θ, la fonction de densité pour un paramètre θ donné. L'espace des paramètres Θ est une variété différentiable, souvent un ouvert de $\mathbb{R}^d$, qui peut être munie d'une métrique riemannienne basée sur l'information de Fisher $\mathsf{I}(\theta)$. Cette métrique, appelée métrique d'information de Fisher ou métrique de Fisher-Rao, est donnée en chaque point θ de Θ par

$$g_\theta(u, v) = u^T \mathsf{I}(\theta) v, \quad u,\, v \in T_\theta \Theta \simeq \mathbb{R}^d, \tag{1}$$

où $\mathsf{I}(\theta) = \mathbf{E}\left(\partial_\theta \log p_\theta \, \partial_\theta \log p_\theta^T \right)$ représente la métrique en coordonnées locales sous sa forme matricielle. Cette métrique permet de définir une distance géodésique sur la va-

riété Θ, appelée distance d'information de Fisher. Notons que l'information de Fisher présente deux propriétés qui rend son choix naturel pour construire une telle distance. Tout d'abord, la structure géométrique d'une variété statistique est préservée par transformation de variable [Amari and Nagaoka, 2007, Calin and Udrişte, 2014], correspondant à la conservation de l'information de Fisher par des statistiques suffisantes. Le théorème de Chentsov et ses nombreuses généralisations montre que la métrique de Fisher est la seule à vérifier cette propriété d'invariance sous certaines hypothèses. Par ailleurs, elle est compatible avec les changements de paramètres par des difféomorphismes $\eta \mapsto \theta(\eta)$. Cette propriété est très utile pour obtenir une forme diagonale de la métrique, plus facile d'utilisation pour les calculs. La distance géodésique étant invariante par changement difféomorphe de paramètres, la structure géométrique ne dépend donc pas du choix des paramètres. Cette propriété permet de considérer la métrique de Fisher comme une métrique riemanienne sur une variété dont un système de coordonnées locales est donné par les paramètres θ de la famille de densités p_θ.

2.2 Géométrie de la variété statistique gamma généralisée

La géométrie de l'information de la variété gamma associée à la famille des distributions gamma a largement été étudiée dans [Arwini et al., 2008]. Cependant, le choix de cette variété statistique ne s'est pas révélé pertinent pour étudier les applications visées en imagerie médicale et en transport aérien. En effet, les distributions de probabilité étudiées dans ces deux applications se sont avérées plutôt provenir de la famille des distributions de type gamma généralisée, qui ajoute un paramètre de forme supplémentaire. La distribution gamma généralisée a été introduite par [Stacy, 1962] comme une généralisation de la distribution gamma et peut être considérée comme un cas particulier de la distribution d'Amoroso [Amoroso, 1925] dans laquelle le paramètre de décalage est supprimé. Outre la distribution gamma, elle généralise également la distribution de Weibull et est couramment utilisée dans les modèles de survie. Seuls quelques résultats sont connus pour la variété des distributions de type gamma généralisée. Dans ce travail, nous avons donc particulièrement exploré la géométrie de cette variété statistique dont les détails peuvent être trouvés dans [Rebbah et al., 2019].

La densité de probabilité d'une loi gamma généralisée est définie sur $\mathbb{R}^+\backslash\{0\}$ et peut s'exprimer comme une généralisation d'une loi gamma exprimée sous sa forme exponentielle naturelle. La densité devient alors

$$p(x;\eta,\lambda,\beta) = \frac{\beta\,\eta^{\lambda}\,x^{\beta\lambda-1}\,e^{-\eta x^{\beta}}}{\Gamma(\lambda)}, \quad x>0, \tag{2}$$

with $\eta > 0$, $\lambda > 0$ and $\beta > 0$. Notons qu'à cause du paramètre puissance β, la distribution gamma généralisée n'est pas une famille exponentielle naturelle. Cependant, en utilisant le difféomorphisme $\Phi_\beta \colon x \mapsto x^\beta$, la propriété d'invariance de la métrique de Fisher permet de montrer que la métrique induite sur les sous-variétés $\beta =$ constante est indépendante

de β, et est exactement celle d'une variété gamma :

$$G(\eta,\lambda) = \begin{pmatrix} \frac{\lambda}{\eta^2} & -\frac{1}{\eta} \\ -\frac{1}{\eta} & \psi'(\lambda) \end{pmatrix}. \tag{3}$$

En coordonnées locales, la métrique d'information de Fisher d'une variété gamma généralisée est donnée par sa forme matricielle :

$$g_{\eta\eta} = \frac{\lambda}{\eta^2}, \tag{4}$$

$$g_{\eta\lambda} = -\frac{1}{\eta}, \tag{5}$$

$$g_{\lambda\lambda} = \psi'(\lambda), \tag{6}$$

$$g_{\eta\beta} = \frac{\lambda}{\eta\beta}\left(\psi(\lambda+1) - \log\eta\right), \tag{7}$$

$$g_{\lambda\beta} = \frac{1}{\beta}\left(\log\eta - \psi(\lambda)\right), \tag{8}$$

$$g_{\beta\beta} = \frac{1}{\beta^2}\left[1 + \lambda\log^2\eta - 2\lambda\psi(\lambda+1)\log\eta + \lambda\psi^2(\lambda+1) + \lambda\psi'(\lambda+1)\right]. \tag{9}$$

On remarque que le bloc (η,λ) correspond bien à la métrique de Fisher de la variété gamma, comme conséquence de la propriété d'invariance par statistique suffisante.

Du fait de la propriété de groupe de la famille des difféomorphismes Φ_β, tous les calculs peuvent être effectués dans le cadre de travail plus général des groupes de Lie. Considérons p_θ, $\theta \in \Theta$, une famille de densités de probabilité, définie sur un ouvert U de $\mathbb{R}^n$, et W un groupe de Lie agissant sur U par des diffeomorphismes préservant l'orientation $x \in U \mapsto \xi(w,x) = w.x$. La densité image par le difféomorphisme est donnée, pour tout $x \in U$ par :

$$\tilde{p}_{w,\theta}(x) = p_\theta(\xi(w,x)) \det \partial_2\xi(w,x), \tag{10}$$

où ∂_i indique une dérivée partielle par rapport à la i-ième variable. Les dérivées d'ordre supérieur sont notées de même $\partial_{i\ldots i,j\ldots j,\ldots}$ en répétant l'indice k fois pour indiquer une dérivée partielle d'ordre k. On peut alors facilement calculer la métrique d'information de Fisher de la densité image $\tilde{p}_{w,\theta}$ sous sa forme matricielle, en particulier les éléments suivants :

$$g_{w,\theta} = -\int_U \partial_{12} l(x,\theta)\partial_1\xi(e,x)p_\theta(x)dx\, T_w R_{w^{-1}}, \tag{11}$$

$$g_{w,w} = T_w R_{w^{-1}}^T \int_U h_{w,\theta}(x)^T h_{w,\theta}(x) p_\theta(x)dx\, T_w R_{w^{-1}}, \tag{12}$$

où $h_{w,\theta}(x) = \partial_1 l(x,\theta)\partial_1\xi(e,x) + \mathrm{tr}\left(\partial_{12}\xi(e,x)\right)$ et $l(x,\theta)$ désigne la log vraisemblance $p_\theta(x)$. Notons que cette métrique peut s'exprimer comme une métrique invariante à droite sur le

groupe de Lie W. Ce résultat pourra s'appliquer en particulier au cas des familles gamma p_θ, exprimée au moyen des paramètres naturels $\theta = (\eta, \lambda)$. La famille image $\tilde{p}_{w,\theta}$ par le difféomorphisme $\Phi_\beta \colon x \mapsto x^\beta$ correspond à la famille des distributions gamma généralisée.

3 Applications à l'imagerie médicale et au transport aérien

Les distances géodésiques entre distributions sont calculées entre deux points de la variété statistique en résolvant des équations différentielles ordinaires, appelées équations géodésiques. Ces distances seront ensuite implémentées dans un algorithme des k-médoïdes. Une méthode appelée "shooting method" a été sélectionnée pour résoudre ce problème aux limites ; elle converge dans le cas de la variété gamma mais pas forcément dans le cas d'une variété gamma généralisée. Une approximation de la distance géodésique a donc été proposée pour contourner ce problème. Elle est basée sur l'observation que la variété gamma est isométriquement plongée dans la variété gamma généralisée lorsque β est constant. On peut alors obtenir une distance approximative en la décomposant séparément sur les parties verticales et horizontales. Notons $p(\eta_2, \lambda_2, \beta_2)$ et $p(\eta_1, \lambda_1, \beta_1)$ deux densités de probabilité gamma généralisée. On procède alors en deux étapes. Tout d'abord, on calcule la première composante d_1 (partie verticale) sur une ligne reliant deux points β_1 et β_2 pour $\eta = cte, \lambda = cte$. Puis, on calcule la seconde composante d_2 (partie horizontale) en suivant une géodésique sur la sous-variété gamma pour $\beta = \beta_2$. On obtient alors une mesure de similarité $d^2 = d_1^2 + d_2^2$.

Une première application a été menée dans le contexte de l'imagerie médicale pour évaluer si les outils issus de la géométrie de l'information pouvaient améliorer les performances d'algorithmes de classification de la maladie d'Alzheimer. Dans cette étude, les sujets proviennent de la base de données ADNI [Mueller et al., 2005] pour 72 patients atteints de la maladie d'Alzheimer (AD) et 71 sujets sains (HC). La qualité de la classification a été évaluée pour plusieurs types de mesure de dissimilarité : la distance géodésique sur la variété gamma généralisée (DGG1), une approximation de cette distance géodésique (DGG2), la distance géodésique sur la variété gamma (DG), la valeur absolue entre moyennes empiriques (DM) et la divergence de Kullback-Leibler pour les distributions gamma généralisées (KL). Les résultats obtenus ont montré que la variété gamma généralisée (DGG1 et DGG2) discrimine mieux les deux groupes de sujets comparée à la variété gamma (DG) et aux autres méthodes (DM et KL).

Dans le domaine du transport aérien, nous nous sommes intéressés au problème de classification des retards aéroportuaires au départ de 40 aéroports en Europe. Les distances géodésiques ont été implémentées de manière similaire dans un algorithme des k-medoïdes.

Références

[Amari, 2016] Amari, S. (2016). *Information Geometry and Its Applications*. Applied Mathematical Sciences. Springer Japan.

[Amari and Nagaoka, 2007] Amari, S. and Nagaoka, H. (2007). *Methods of Information Geometry*. Translations of mathematical monographs. American Mathematical Society.

[Amoroso, 1925] Amoroso, L. (1925). Ricerche intorno alla curva dei redditi. *Annali di Matematica Pura ed Applicata*, 2(1) :123–159.

[Arwini et al., 2008] Arwini, K., Dodson, C., Doig, A., Sampson, W., Scharcanski, J., and Felipussi, S. (2008). *Information Geometry : Near Randomness and Near Independence*. Information Geometry : Near Randomness and Near Independence. Springer.

[Busovaca et al., 2016] Busovaca, E., Zimmerman, M. E., Meier, I. B., Griffith, E. Y., Grieve, S. M., Korgaonkar, M. S., Williams, L. M., and Brickman, A. M. (2016). Is the Alzheimer's disease cortical thickness signature a biological marker for memory ? *Brain imaging and behavior*, 10(2) :517–523.

[Calin and Udrişte, 2014] Calin, O. and Udrişte, C. (2014). *Geometric Modeling in Probability and Statistics*. Mathematics and Statistics. Springer International Publishing.

[Cercignani et al., 2001] Cercignani, M., Inglese, M., Pagani, E., Comi, G., and Filippi, M. (2001). Mean diffusivity and fractional anisotropy histograms of patients with multiple sclerosis. *American Journal of Neuroradiology*, 22(5) :952–958.

[Cuingnet et al., 2011] Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M.-O., Chupin, M., Benali, H., Colliot, O., and ADNI (2011). Automatic classification of patients with Alzheimer's disease from structural MRI : a comparison of ten methods using the ADNI database. *NeuroImage*, 56(2) :766–781.

[Duistermaat, 2001] Duistermaat, J. (2001). On hessian riemannian structures. *Asian journal of mathematics*, 5 :79–91.

[Lama et al., 2017] Lama, R. K., Gwak, J., Park, J.-S., and Lee, S.-W. (2017). Diagnosis of Alzheimer's Disease Based on Structural MRI Images Using a Regularized Extreme Learning Machine and PCA Features. *Journal of Healthcare Engineering*, 2017.

[Mueller et al., 2005] Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C., Jagust, W., Trojanowski, J. Q., Toga, A. W., and Beckett, L. (2005). The Alzheimer's Disease Neuroimaging Initiative. *Neuroimaging clinics of North America*, 15(4) :869–xii.

[Rebbah et al., 2019] Rebbah, S., Nicol, F., and Puechmorel, S. (2019). The geometry of the generalized gamma manifold and an application to medical imaging. *Mathematics*, 7(8) :674.

[Stacy, 1962] Stacy, E. W. (1962). A generalization of the gamma distribution. *The Annals of Mathematical Statistics*, 33(3) :1187–1192.

Analyse de données d'épidémie de malaria par un modèle de fragilité multivarié à corrélations spatiales

Ajmal Oodally [1] & Klara Goethals [2] & Estelle Kuhn [3] & Luc Duchateau [4]

[1] *ajmal.oodally@inrae.fr*
[2] *klara.goethals@ugent.be*
[3] *estelle.kuhn@inrae.fr*
[4] *luc.duchateau@ugent.be*

Résumé La malaria est une maladie avec un taux de mortalité qui reste élevé en Afrique sub-saharienne. La transmission se fait via un moustique, dont le développement et la reproduction sont favorisés par la présence de plans d'eau. Afin d'étudier l'influence des plans d'eau sur le taux de transmission de la malaria, un riche jeu de données a été constitué dans le secteur du barrage hydroélectrique de Gilgel Gibe dans le sud-ouest de l'Éthiopie. Il est constitué de temps d'infection par la malaria d'enfants répartis en villages, ainsi que de nombreuses covariables. Ces données ont déjà été analysées par un modèle de fragilité structuré selon les villages, incluant la distance entre l'enfant et le barrage comme covariable. Cependant, la proximité entre les enfants n'est pas prise en compte. Afin de mieux prendre en compte cette spécificité liées aux données, nous proposons un modèle de fragilité avec une structure de corrélation spatiale. Les paramètres du modèle sont estimés en maximisant la vraisemblance observée via un algorithme de type Expectation Maximization stochastique. La performance de l'estimateur est évaluée sur des données simulées et des données réelles.

Mots-clés. Modèle de fragilité multivarié, corrélations spatiales, algorithme Expectation Maximization stochastique, incidence de la malaria

Summary

Malaria remains a disease with high morbidity and mortality in Subsaharan Africa, and more specifically in Ethiopia. The mosquito being the vector of this disease, the presence of water bodies, favoring the reproduction and breeding of the mosquito, strongly influences the rate of transmission. A rich dataset has been put together in the area of the Gilgel Gibe hydroelectric dam in south-western Ethiopia, including the malaria infection times of

2037 children situated in different villages as well as many covariates. The data has been analyzed by frailty models introducing the village as cluster to accommodate for the correlation in the data and distance from the dam as main risk factor. However, proximity between the children is not taken into account. In order to consider this specificity in the data, we propose a frailty model with a spatial correlation structure. The parameters of the model are estimated by maximizing the observed likelihood via a stochastic Expectation Maximization algorithm. Parameter estimation is done on the malaria dataset and simulated data to assess the performance of the estimator.

Keywords. Multivariate frality model, spatial correlation, stochastic Expectation Maximization algorithm, malaria incidence

1 Contexte, problématique et données

La malaria est une maladie avec un taux de mortalité qui reste élevé en Afrique sub-saharienne. Le parasite de la malaria est transmis d'homme à homme par le moustique anophèle. Ce moustique est fortement dépendant de l'eau à tous les stades de son développement (œuf, larve, nymphe) et pour sa reproduction. De fait, la présence de plans d'eau, comme ceux liés à la construction de barrages hydrauliques, peut potentiellement avoir un fort impact sur l'incidence de la malaria. En 2003, le barrage hydroélectrique de Gilgel Gibe a été construit dans le sud-ouest de l'Éthiopie. Afin d'étudier l'effet du réservoir d'eau sur la propagation de la malaria, 16 villages à différentes distances du barrage, variant entre 0,26 et 9,05 km, ont été sélectionnés. Au total, 2037 enfants de moins de 10 ans ont fait l'objet d'un suivi hebdomadaire entre juillet 2008 et juin 2010. Les temps d'infection par la malaria ont été observés sur ces enfants avec censure, ainsi que de nombreuses covariables descriptives. Pour plus d'information sur ces données, nous renvoyons aux articles de Yewhalaw et al (2009), Getachew et al (2013). L'hypothèse principale de ces travaux était que plus on s'éloigne du barrage, plus l'incidence de la malaria diminue. Les données ont été analysées via des modèles d'analyse de survie pour mieux quantifier cet effet.

Les modèles de fragilité introduits par Vaupel et al (1979) sont une extension du modèle de Cox (1972) qui modélise le risque de survenue d'un événement comme produit d'une fonction de risque de base et d'une fonction des covariables. Ces modèles permettent en outre de prendre en compte l'hétérogénéité présente dans les données via des effets aléatoires non ob-

servés. Ainsi un modèle de fragilité incluant l'effet "village" comme effet aléatoire et la distance entre l'individu et le barrage comme facteur de risque principal a été utilisé par Getachew et al (2013). Cependant, cette modélisation soulève deux remarques. Premièrement, les résultats du modèle de fragilité ne sont pas fiables lorsqu'il existe une forte corrélation entre la covariable, ici la distance au barrage, et le groupe, ici le village, ce qui est le cas dans ce jeu de données. Deuxièmement, le village est une structure administrative qui ne rend pas forcément compte de la proximité géographique entre individu.

Afin de mieux prendre en compte les spécificités de ces données, nous proposons un modèle de fragilité avec une structure de corrélation spatiale au niveau de l'individu.

2 Modèle de fragilité multivarié à corrélation spatiale

Usuellement, les modèles de fragilité spatiaux modélisent les corrélations entre groupes (cf. Li et Ryan (2002)). Nous proposons un modèle de fragilité multivarié à corrélations spatiales au niveau de l'individu. On considère une population de N individus. Pour $1 \leq i \leq N$, le temps de survenue de l'infection et le temps de censure pour l'individu i sont modélisés par des variables aléatoires notées T_i et C_i respectivement. On observe alors pour $1 \leq i \leq N$ le temps censuré à droite et l'indicateur de censure notés respectivement X_i et Δ_i et définis par $X_i = \min(T_i, C_i)$ et $\Delta_i = \mathbb{1}_{T_i \leq C_i}$. Dans la suite, on note $\mathbf{X} = (X_i)_{1 \leq i \leq N}$ et $\mathbf{\Delta} = (\Delta_i)_{1 \leq i \leq N}$. Le modèle de fragilité multivarié à corrélations spatiales est défini pour $1 \leq i \leq N$ par :

$$h_i(t|b_i) = h_0(t)\ \exp(Z_i^t \beta + b_i) \quad i = 1, \ldots, N \text{ où } \ b = (b_i)_{1 \leq i \leq N} \sim \mathcal{N}_N(0, \Gamma)$$

où $h_i(t|b_i)$ désigne le risque instantané de survenue de l'événement pour l'individu i au temps t, $h_0(t)$ le risque de base au temps t, b_i le vecteur de fragilité de l'individu i, β le vecteur des paramètres de régression inconnu, Z_i le vecteur de covariables associées à l'individu i. Le vecteur de fragilité b suit une distribution normale multivariée centrée avec une matrice de covariance notée $\Gamma = \sigma^2 \Sigma(\rho)$ où σ^2 est un facteur d'échelle et $\Sigma(\rho)$ est la matrice de corrélation structurée paramétrée par $\rho > 0$. Nous considérons les deux

structures suivantes :

$$\Sigma_1(\rho) = \exp(-\rho D) \quad \text{et} \quad \Sigma_2(\rho) = \frac{1}{1+D^\rho} \tag{1}$$

où $D = (d_{ij}) \in R^{N\times N}$ est telle que la composante d_{ij} correspond à la distance entre l'individu i et l'individu j pour $i \neq j$ et $d_{ii} = 0$ par convention.

On suppose par ailleurs que la fonction de risque de base h_0 est paramétrique constante par morceaux pour prendre en compte l'effet saisonnier de l'incidence. La fonction de risque de base est définie par $h_0(t) = h_m$ pour $t \in [\tau_{m-1}, \tau_m[$ pour $m \in [1, M]$ où $(\tau_m)_{m\in[1,M]}$ est une suite strictement croissante et $\tau_0 = 0$. Le modèle s'écrit alors :

$$h(t|b_i) = \sum_{m=1}^{M} h_m \mathbb{1}_{[\tau_{m-1},\tau_m[}(t) \exp(Z_i^t\beta + b_i) \tag{2}$$

Les paramètres à estimer sont $\theta = ((h_m)_{1\leq m\leq M}, \beta, \sigma^2, \rho)$.

3 Estimation des paramètres

Nous estimons les paramètres du modèle par maximum de vraisemblance. La log-vraisemblance complète des données s'écrit :

$$\begin{aligned}\log L_{\text{comp}}(\theta; \mathbf{X}, \mathbf{\Delta}, b) \;=\; & \sum_{m=1}^{M}\sum_{i=1}^{N}\Bigg(\Delta_i\Big(\log(h_m)\mathbb{1}_{[\tau_{m-1},\tau_m[}(X_i) + Z_i^t\beta + b_i\Big) \\ & -H(X_i)\exp(Z_i^t\beta + b_i)\Bigg) + \sum_{i=1}^{N}\log f_\Gamma(b_i)\end{aligned}$$

où le hasard cumulé est noté $H(X_i) = \sum_{l=1}^{M} h_l(\tau_l - \tau_{l-1})\mathbb{1}_{\tau_l\leq X_i} + \sum_{l=1}^{M}(X_i - \tau_{l-1})h_l\mathbb{1}_{\tau_{l-1}\leq X_i<\tau_l}$. La vraisemblance observée est obtenue en intégrant la vraisemblance complète par rapport au vecteur de fragilité b :

$$L_{\text{obs}}(\theta; \mathbf{X}, \mathbf{\Delta}) = \int L_{\text{comp}}(\theta; \mathbf{X}, \mathbf{\Delta}, b)db$$

On définit $\hat{\theta}$ l'estimateur du maximum de la vraisemblance observée par $\hat{\theta} = \text{argmax}\, L_{\text{obs}}(\theta; \mathbf{X}, \mathbf{\Delta})$. Le calcul de cet estimateur ne peut en général pas se faire directement, en particulier lorsque la vraisemblance observée n'admet pas de forme analytique, ce qui est le cas dans le modèle de fragilité défini dans la section 2. En pratique, nous calculons la valeur de l'estimateur via un algorithme itératif stochastique de type Expectation Maximization.

4 Algorithme stochastique d'estimation

On applique l'algorithme MCMC-SAEM introduit par Kuhn et Lavielle (2004) qui combine une méthode de Monte Carlo Markov Chain à l'algorithme d'approximation stochastique de EM. Chaque itération de l'algorithme se compose de trois étapes. A l'itération k :

S-step : une réalisation b^k du vecteur de fragilité non observé est simulé selon le noyau de transition d'une chaîne de Markov convergente $\Pi_{\theta_{k-1}}$ ayant comme distribution stationnaire la distribution conditionnelle du vecteur de fragilité notée $\pi_{\theta_{k-1}}(.|\mathbf{X}, \mathbf{\Delta})$.

$$\pi_{\theta_{k-1}}(.|\mathbf{X}, \mathbf{\Delta}) = \frac{L_{\text{comp}}(\theta; \mathbf{X}, \mathbf{\Delta}, b)}{L_{\text{obs}}(\theta; \mathbf{X}, \mathbf{\Delta})}$$

SA-step : On effectue une approximation stochastique sur la log-vraisemblance complète :

$$Q_k(\theta) = Q_{k-1}(\theta) + \gamma_k(\log L_{\text{comp}}(\theta; \mathbf{X}, \mathbf{\Delta}, b^k) - Q_{k-1}(\theta))$$

où la suite (γ_k) vérifie $0 \leq \gamma_k \leq 1, \sum \gamma_k = +\infty, \sum \gamma_k^2 < +\infty$.

M-step : On met à jour les paramètres selon : $\theta_k = \underset{\theta}{\text{argmax}}\ Q_k(\theta)$

Les quantités Q_0 et θ_0 sont initialisées arbitrairement. Sous des hypothèse de régularité du modèle de fragilité et sous des hypothèses assurant l'ergodicité de la chaine de Markov, la suite $(\theta_k)_k$ générée par l'algorithme décrit ci-dessus converge presque sûrement vers un point critique de la vraisemblance observée (cf. Kuhn et Lavielle (2004)).

5 Expériences numériques

On analyse les temps de survenue de la malaria chez 2037 enfants. Les quatres covariables considérées sont l'âge, le sexe, la structure du toit et la distance entre l'enfant et le barrage. Du fait de la grande dimension du vecteur de fragilité et de l'hétérogénéité de ses composantes, l'étape de simulation est délicate et nécessite l'utilisation de techniques adaptées à ce contexte. Ainsi la simulation du vecteur b^k à l'itération k de l'algorithme se fait via un

Table 1 – Estimations des paramètres et écart type entre parenthèses

β_{sex}	β_{age}	β_{d}	β_{roof}	$(h_1, h_2, h_3, h_4, h_5, h_6) \times 10^{-4}$	σ^2	ρ
-0.0391	-0.0061	0.1057	0.0260	(5.40, 14.3, 3.98, 6.88, 2.42, 2.71)	0.364	0.794
(0.0659)	(0.0201)	(0.140)	(0.0342)	(0.48,0.97,0.22,0.49,0.11,0.18)	(0.088)	(0.11)

algorithme de Gibbs hybride adaptatif (cf. Atchadé et Rosenthal (2005)), assurant un taux d'acceptation homogène selon toutes les composantes.

Nous avons testé la méthode d'estimation sur des données simulées selon le modèle défini dans l'équation (2) et avons obtenu de très bons résultats. Pour l'analyse des données de malaria, nous comparons plusieurs modèles avec différentes fonctions de risque de base h_0 et les structures de corrélation présentées dans la section 2. Les modèles que nous comparons ont le même nombre de paramètres et nous présentons donc les résultats du modèle qui maximise la log-vraisemblance dans le tableau 1. L'estimation de h_2 correspond à la plus forte période de pluie et est associée à une plus grande incidence. La présence de plans d'eau est plus importante pendant la grosse saison de pluie et favorise la reproduction des moustiques et contribue donc à propager plus facilement la malaria. Aussi, nous illustrons graphiquement ce que représente l'estimation du paramètre de corrélation ρ dans la figure 1. Cette estimation semble cohérente avec la distance maximum théorique que parcours le moustique.

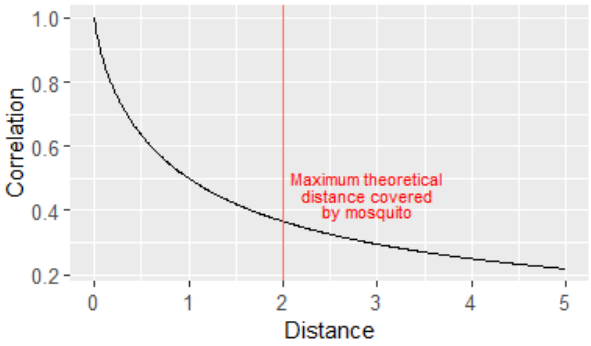


Figure 1 – Représentation graphique de la corrélation en fonction de la distance basée sur l'estimation de $\hat{\rho} = 0.794$

Bibliographie

Cox, D.R. (1972). Regression Models and Life-Tables, *Journal of the Royal Statistical Society*, 34, pp. 187-220. Vaupel, J., Manton, K. et Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality, *Demography*, 16, pp. 439-454. Getachew, Y., Janssen, P., Yewhalaw, D., Speybroeck N. et Duchateau L. (2013). Coping with time and space in modelling malaria incidence: a comparison of survival and count regression models, *Statistics in Medicine*, 32, pp. 3224-3233. Yewhalaw, D., Worku L., Van Bortel W., GebreSelassie S., Kloos H., Duchateau L. et Speybroeck N. (2009). Malaria and water resource development: the case of Gilgel-Gibe hydroelectric dam in Ethiopia, *Malaria Journal*, 8, 21. Dempster, A. P., Laird N. M. et Rubin D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society, Series B*, 39, 1, pp. 1-38. Kuhn, E. et Lavielle, M. (2004). Coupling a stochastic approximation version of EM with an MCMC procedure, *ESAIM: Probability and Statistics*, 8, pp. 115-131. Atchadé, Y. et Rosenthal, J. (2005). On adaptive Markov chain Monte Carlo algorithms, *Bernoulli*, 11. Li, Y. et Ryan, L. (2002). Modeling Spatial Survival Data Using Semiparametric Frailty Models, *Biometrics*, 58, pp. 287-297.

Prévision probabiliste fondée sur l'échangeabilité d'un modèle d'ensemble en environnement.

Éric PARENT [1] & Jacques BERNIER [2]

[1] *UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay, 75005, Paris*
eric.parent@agroparistech.fr
[2] *Ingénieur EDF à la retraite*
jacques.bernier2@orange.fr

Résumé. La prévision *d'ensemble* se fonde sur des scénarios obtenus en modifiant les conditions et les paramètres initiaux d'un code déterministe dynamique décrivant un mécanisme physique (modèles météorologiques régionaux, modèles climatiques, modèles de précipitation et de transport solide ...). Ces scénarios variés traduisent l'espoir du physiciens de représenter l'incertitude due à l'état initial du système (par exemple la température de l'atmosphère, l'humidité du sol, etc.) ou découlant de la connaissance incomplète des paramètres. En propageant les perturbations des conditions limites à travers le code numérique intégro-différentiel complexe simulant le comportement du système, on obtient un ensemble de trajectoires hypothétiques de la variable à prédire (par exemple, la température de l'aéroport, la pluie, le débit des cours d'eau...) également connues sous le nom de *membres*. Dans quelle mesure les informations véhiculées par les différents membres de l'ensemble peuvent-elles aider à construire une prévision probabiliste de la quantité à prévoir? La littérature statistique fourmille de méthodes prêtes à l'emploi, connues sous le nom de techniques de "post-traitement" des ensembles, mais peu d'entre elles traitent formellement des spécificités des ensembles : (1) la trajectoire de la Nature peut ne pas être exactement celle générée par le code déterministe, (2) les membres de l'ensemble proviennent de processus physiques, ce qui devrait permettre un certain apprentissage statistique, (3) les membres de chaque ensemble ont tendance à être, à cause du principe même de leur construction, l' échantillonnage d'un modèle de type *échangeable*.

En statistique, De Finetti et ses successeurs ont démontré que la propriété d'échangeabilité conduit à un modèle à effet aléatoire. L'effet aléatoire résume parcimonieusement l'information véhiculée par les membres de l'ensemblevis à vis la variable à prédire. Nous construisons un modèle normal échangeable pour les températures et suggérons d'autres développements bayésiens hiérarchiques ad hoc : données environnementales avec valeurs nulles (précipitations) ou structures conjointes de séries hydro-météorologiques (pluies-débits). Les performances des modèles proposés sont vérifiées et comparées à celles d'autres techniques de post-traitement sur une longue série d'enregistrements de températures. Les résultats montrent la robustesse des structures prédictives parcimonieuses construites, ce qui plaide en faveur de l'échangeabilité et de la cohérence probabiliste lors de la modélisation des membres d'un ensemble environnemental.

Mots-clés. Modèle bayésien hiérarchique, Prévision Probabiliste, Post-traitement d'ensemble environmental, échangeabilité ...

Abstract. *Ensemble* forecast is based on scenarios obtained by modifying the initial conditions and parameters of a dynamic deterministic code describing a physical mechanism (regional weather models, climate models, precipitation and solid transport models ...). These various scenarios reflect the physicists' hope to represent the uncertainty due to the poorly detremined initial state of the system (e.g. atmospheric temperature, soil moisture, etc.) or resulting from incomplete knowledge of the parameters. By propagating the boundary condition perturbations through the complex integro-differential numerical code mimicking the system behaviour, a bunch of hypothetical trajectories of the variable to be predicted (e.g. airport temperature, rainfall, river flow...) also known as *members of the ensemble* is obtained. To what extent can the information conveyed by the different members of the ensemble help to construct a probabilistic forecast of the quantity to be predicted? Statistical literature offers a wealth of off-the-shelf methods known as "post-processing" techniques for ensembles, but few of them formally address the specific features of an ensemble: (1) the trajectory of Nature may not be exactly the one generated by the deterministic code, (2) the members of the set come from physical processes, which should allow some statistical learning, (3) the members of each ensemble tend to be, because of the very principle of their construction, the sampling of an *exchangeable* model.

In statistics, De Finetti and his successors have shown that the property of exchangeability leads to a random-effect model. The random effect parsimoniously sums up the information conveyed by the members of the ensemble with regard to the variable to be predicted. We construct a normal exchangeable model for temperatures and suggest other ad hoc hierarchical Bayesian developments: environmental data with zero values (precipitation) or joint structures of hydro-meteorological series (rain-flow). The performances of the proposed models are checked and compared with those of other post-processing techniques over a long series of temperature records. The results show the robustness of such statistical structures, which argues in favour of exchangeability and probabilistic consistency when modelling the members of an environmental ensemble.

Keywords. Hierarchical Bayesian model, Probabilistic forecasting, Ensemble post-processing techniques, Exchangeability ...

1 Construction d'une prévision probabiliste fondée sur un ensemble environnemental

Dans le contexte de la prévision d'ensemble, le problème posé, au départ à tout opérateur prévisionniste est la modélisation probabiliste adaptée à la structure des S membres de

l'ensemble $X = \{x_1, x_2, ...x_S\}$ et le passage à la distribution conditionnelle de la cible Y.

Comme dans Courbariaux *et al* (2019), nous travaillons sur les $S = 50$ membres de températures de l'ensemble issu du Centre Européen pour les Prévisions Météorologiques à Moyen Terme pour la station de Vouglans, dans l'Ain. Commençons par distinguer trois types d'échantillons porteurs d'informations:

1. l'échantillon *climatique* (souvent long) où on dispose d'observations de la cible Y, la grandeur température journalière moyenne observée (c'est un long historique journalier (1953 à 2015), d'où son importante pour le calibrage).

2. l'échantillon de *calage* où les modèles marginaux des ensembles X puis le modèle prédictif de $Y|X$ sont estimés (on observe conjointement X et Y de 2005 à 2008, puis de 2011 à 2015),

3. l'échantillon de *vérification- validation* des résultats de la prévision (qui pourra être constitué de sous périodes à extraire de la période conjointe, le reliquat restant pour le calage).

2 Modélisation marginales de l'ensemble X en calage

Notre hypothèse de travail minimale s'appuie sur une certaine symétrie des membres; l'échangeabilité traduira l'invariance de la probabilité de l'ensemble par rapport à toute permutation des S membres. La symétrie acceptée impose, bien sûr, le respect de la stationnarité des séries à traiter, éventuellement obtenue préalablement par une procédure *adhoc* de stationnarisation.

Le théorème de deFinetti (1939), dit de représentation après son extension en 1955 par Savage et Hewitt, stipule que toute séquence *infinie* est échangeable **si et seulement si** il existe "une variable aléatoire Z" telle que l'on a en probabilité, en utilisant la notation *crochet* de Gelfand et Smith (1939) pour les densités :

$$[x_1, ..., x_S] = \int_Z \prod_{s=1}^{S} [x_s|Z][Z]dZ$$

On notera que la réciproque, *seulement si*, s'applique immédiatement même pour tout échantillon fini : tout échantillon d'un *mélange* est échangeable.

Précisons que Z n'est pas n'importe quelle variable aléatoire, elle est étroitement liée à l'échangeabilité des membres, exprimant leur structure particulière et dépend de l'ensemble. Reprenons Bernardo (1996) : *Le théorème de représentation, résultat de pure théorie de probabilité, prouve que, si les observations sont jugées échangeables, alors elles doivent vraiment être un échantillon aléatoire d'un modèle et une distribution a priori sur le paramètre du modèle* (le Z), (imposant de fait une approche bayesienne). Il faut cependant noter que le théorème de représentation est un théorème d'existence: en

géneral il ne spécifie pas le modèle et ne spécifie jamais la distribution a priori requise. Les hypothèses additionnelles qui sont habituellement nécessaires pour spécifier un modèle particulier sont décrites dans les applications. Ici, comme souvent pour les températures, nous invoquons la normalité et proposons le modèle à effets aléatoires construit sur 2 pivôt réels Z_{1t} et Z_{2t}:

$$\begin{aligned} X_{ts} &= \alpha + \beta Z_{1t} + \lambda Z_{2t}^{-0.5} \times \epsilon_{ts} \\ \epsilon_{ts} &\sim N(0,1) \end{aligned} \tag{1}$$

On notera que α, β, λ sont les paramètres de niveau 2 de la structure hiérarchique (cf Bernardo), il faut leur joindre les paramètres des modèles latents des pivôts. Pour des raisons de simplicité statistique, on adoptera l'hypothèse de conjuguées naturelles (Raiffa et Schlaifer, 1961); soit , en notant G la loi gamma:

$$Z_{2t} \sim G(g,1) \tag{2}$$

$$Z_{1t}|Z_{2t} \sim N(0, Z_{2t}^{-0.5}) \tag{3}$$

Ces trois équations définissent un modèles marginal d'ensemble échangeable compatible avec la représentation issue du théorème de de Finetti.

Si, en première analyse on prend pour exemple la période de calage des dennées de température journalières désaisonnalisées (mai-novembre) à Vouglans, on a un échantillon de taille $n > 850$, ce qui nous conduit, d'une part, à pouvoir utiliser des priors non informatifs, et d'autre part, à négliger les incertitudes a posteriori sur les paramètres en les posant égaux à leurs moyennes a posteriori dans les inferences de modèles.

Comme il se doit pour ce modèle normal, la loi de $Z = (Z_1, Z_2)$ conditionnelle à X possède une statistique exhaustive $T(X_t)$ à t fixé : le couple $T(x_t = (\bar{x}_t, s_t^2)$ (moyenne et variance empirique), observable sur chaque ensemble et qui les remplace comme pivôt $Z = (Z_1, Z_2)$ dans toute équation de prévision. Lauritzen (2007) les appelent *résumantes des latentes* Z dans l'expression du théorème de deFinetti. On en déduit les lois marginales conditionnelles de type Fisher (F) et Student (T) pour les composantes du résumé de l'ensemble :

$$\left(\frac{S.s_t^2}{\lambda^2}\right) \sim F(S-1, 2g) \tag{4}$$

$$\left(\frac{\bar{x}_t - \alpha}{k \times s_t}\right) \sim T(S-1) \tag{5}$$

$$\text{avec } k = \left(\sqrt{\frac{\beta^2 + S\lambda^2}{\lambda^2}}\right)$$

3 Modélisation du lien entre la cible Y et l'ensemble X en calage

Pour la phase prédictive, il nous faudra expliciter $[Y|X] = \int [Y|Z][Z|X]dZ$ où l'on peut remplacer les pivôts Z par les statistiques résumantes $T(X)$. On suppose bien sûr que pour chaque temps t, X_t, (qu'on pourrait qualifier d'exhaustif vis à vis de Y) est censé contenir toutes les informations nécessaires à la prévision.

On cherche donc à construire un modèle bivariable $[Y, T(X)]$ qui doit encapsuler les propriétés marginales des résumantes d'ensemble d'un coté et des propriét és climatiques de Y (issues du calibrage) de l'autre coté.

On suggère d'étendre ici une méthode de *copule* normale inspirée de Krzysztofowicz (1997) , en supposant que le résumé statistique s_t^2, la variance de l' ensemble, est un prédicteur plus ou moins relié à la variance de la prédictive. Nous proposont un modèle sur les grandeurs transformées U, V, W à deux paramètres θ et ψ à estimer sur la période de calage, sous la forme:

$$\begin{aligned} V &= \frac{Y - \mathbb{E}(Y)}{\sqrt{\mathbb{V}(Y)}} \sim N(0,1) \\ U &= N\left(T_{S-1}^{-1}\left(\frac{\bar{x}_t - \alpha}{k \times s_t}\right)\right) \\ W &= G_{g,1}\left(F_{S-1,2g}^{-1}\left(\frac{S.s_t^2}{\lambda^2}\right)\right) \\ V_t &= \theta \times U_t + \psi \times \sqrt{W_t} \times \epsilon_t \\ \epsilon_t &\sim N(0,1) \end{aligned} \tag{6}$$

Pour W_t, on pourrait également utiliser une autre transformation d'un prédicteur positif de variance d'erreur de prévision, par exemple une loi normale tronquée sur la demi-droite $0 < W < +\infty$

4 Conclusion

Comparé aux approches classiques en matière de prévisions d'ensemble de Jolliffe *et al* (2012) , le modèle échangeable pour les ensembles environnementaux repose sur une propriété statistique constructive dont le réalisme opérationnel est facilement défendable : l'échangeabilité des trajectoires possibles des grandeurs météorologiques est justement la propriété que souhaitent les physiciens qui établissent les membres de l'ensemble! Quoique son adaptation à certains types de données reste un challenge (le traitement séparé des valeurs nulles des précipitations pose un délicat problème d'introduction de variables latentes, par exemple), les premiers essais de ce modèle présentent des aspects prometteurs

de robustesse et de parcimonie. Il peut sans doute d'ailleurs être perfectionnés de multiples façons: introduction de mémoire autorégressive marginale des latentes d'ensemble, descripteurs d'information supplémentaires au niveau prédictif, adaptabilité par apprentissage séquentiel régulièrement réactualisé, etc.

Remerciements

Ce travail fait suite à la thèse de Marie Courbariaux financée par EDF et Hydro-Québec (contrat AgroParisTech/Innovation 694R) et soutenue en 2016. Les réflexions méthodologiques de cette communication ont été initiées par les échanges soutenus durant cette période et par la suite avec les cadres des services de recherche développement de ces deux sociétés productrices d'hydro-électricité: Rémy Garcon, Luc Perreault et Joël Gailhard.

Bibliographie

Bernardo, J. M. (1996). The concept of Exchangeability and its applications. *Far East Journal of Mathematical Sciences*, 4:111–122.

Courbariaux, M., Barbillon, P., Perreault, L., et Parent, É. (2019). Post-processing multiensemble temperature and precipitation forecasts through an exchangeable normal-gamma model and its tobit extension. *Journal of Agricultural, Biological and Environmental Statistics*, 24(2):309–345.

de Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. In *Annales de l'institut Henri Poincaré*, volume 7, pages 1–68.

Gelfand, A. E. et Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409.

Jolliffe, I. T. et Stephenson, D. B. (2012). *Forecast verification: a practitioner's guide in atmospheric science*. John Wiley & Sons.

Hewitt E et Savage L. J. (1955). Symetric measures on cartesian masures in Trans. Amer. Math. Soc. 80.

Krzysztofowicz, R. (1997). Transformation and normalization of variates with specified distributions. *Journal of Hydrology*, 197(1):286–292.

Lauritzen S. (2007) Exchangeability and de Finetti's Theorem. http://www. stats. ox. ac. uk/˜ steffen/teaching/grad/definetti.pdf.

Schlaifer, R. et Raiffa, H. (1961). *Applied statistical Decision Theory*. Division of Research, Harvard Business School, Boston, MA.

Towards new cross-validation-based estimators for Gaussian process regression: efficient adjoint computation of gradients

Sébastien J. Petit[1,2,⋆] & Julien Bect[1] & Sébastien Da Veiga[3]
& Paul Feliot[2] & Emmanuel Vazquez[1]

[1] *Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des Signaux et Systèmes, Gif-sur-Yvette, France.* ⋆*E-mail :* sebastien.petit@centralesupelec.fr.
[2] *Safran Aircraft Engines, Moissy-Cramayel, France*
[3] *Safran Tech, Châteaufort, France*

Résumé. Nous nous intéressons à l'estimation par validation croisée des paramètres d'une fonction de covariance d'un processus gaussien. Nous suggérons l'utilisation de nouveaux critères de validation croisée dérivés de la littérature des *scoring rules*. Nous proposons de plus une méthode efficace pour le calcul du gradient d'un critère de validation. Cette méthode est plus efficace que ce qui est présenté dans la littérature à notre connaissance, et permet en particulier de réduire la complexité de l'évaluation jointe des critères de validation croisée et des gradients associés.

Mots-clés. Processus gaussien, validation croisée, score de prédiction probabiliste, Mode adjoint

Abstract. We consider the problem of estimating the parameters of the covariance function of a Gaussian process by cross-validation. We suggest using new cross-validation criteria derived from the literature of *scoring rules*. We also provide an efficient method for computing the gradient of a cross-validation criterion. To the best of our knowledge, our method is more efficient than what has been proposed in the literature so far. It makes it possible to lower the complexity of jointly evaluating leave-one-out criteria and their gradients.

Keywords. Gaussian process, cross-validation, scoring rule, reverse-mode differentiation

1 Introduction

Let ξ be a zero-mean Gaussian process indexed by $\mathbb{R}^d$ and denote by k the covariance function of ξ, which is assumed to belong to a parametrized family $\{k_\theta; \theta \in \Theta\}$, where $\Theta \subseteq \mathbb{R}^q$ denotes a q-dimensional space of parameters. We can safely say that the most popular methods for estimating k from data are maximum likelihood and related techniques.

In this article we focus instead on cross-validation methods. Classical cross-validation methods for estimating k are based on the leave-one-out mean squared prediction error or PRESS (Allen, 1974; Bachoc, 2013), and leave-one-out log predictive density (see, e.g., Rasmussen and Williams, 2006). These leave-one-out goodness-of-fit criteria can be computed using closed-form formulas (Dubrule, 1983).

The contribution of this work is twofold. First, we suggest extending the range of classical cross-validation criteria available in the literature of Gaussian processes by using the broad variety of *scoring rules* (see Gneiting and Raftery, 2007, for a survey), such as the continuous ranked probability score (CRPS). Second, we provide an efficient way for computing the gradient of any cross-validation criterion, which can then be used in gradient-based optimization algorithms. The only requirement is for the criterion to be differentiable in closed form with respect to leave-one-out posterior predictive means and variances. The new procedure has a $\mathcal{O}(n^3 + qn^2)$ complexity, against the $\mathcal{O}(qn^3)$ that was deemed "unavoidable" by Rasmussen and Williams (2006).

The article is organized as follows. Section 2 introduces scoring rules and how they can be used for estimating k. Section 3 presents the details of our contribution to the computation of gradients of a cross-validation criterion and Section 4 presents our conclusions and perspectives.

2 Scoring rules and cross-validation criteria

Let $Z_i = \xi(x_i) + \varepsilon_i$, $1 \leq i \leq n$, be some observations of ξ, at points $x_i \in \mathbb{R}^d$, where the ε_is are assumed independent and identically $\mathcal{N}(0, \sigma_\varepsilon^2)$-distributed, with $\sigma_\varepsilon^2 \geq 0$.

The classical framework of Gaussian process regression allows one to build a predictive distribution for an unobserved $\xi(x)$ at $x \in \mathbb{R}^d$ from the Z_is. Criteria for assessing the quality of probabilistic predictions have been studied in depth under the name of *scoring rules* in the seminal article of Gneiting and Raftery (2007). A scoring rule for real variable prediction is a function $S : \mathcal{P} \times \mathbb{R} \to [-\infty, +\infty]$, where $\mathcal{P}$ is a class of probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. For $P \in \mathcal{P}$ and $z \in \mathbb{R}$, $S(P, z)$ measures the goodness of prediction P for z.

Assume that we want to use a scoring rule S for estimating the parameters of a covariance function. Gneiting and Raftery (2007) suggest building a leave-one-out cross-validation criterion L defined as

$$L(\theta) = \frac{1}{n} \sum_{i=1}^{n} S(P_{-i}^{\theta}, Z_i), \tag{1}$$

where P_{-i}^{θ} is the conditional distribution of Z_i given the Z_js, for $1 \leq j \leq n$, $j \neq i$.

In our Gaussian process regression framework, it is well known that P_{-i}^{θ} is a Gaussian distribution $\mathcal{N}(\mu_i, \sigma_i^2)$. Let $K = (k_\theta(x_i, x_j))_{1 \leq i,j \leq n}$ be the covariance matrix of $(\xi(x_1), \ldots, \xi(x_n))^{\mathrm{T}}$, then (Dubrule, 1983; Sundararajan and Keerthi, 2001; Craven and

Wahba, 1979) show that

$$\mu_i = Z_i - \frac{(BZ)_i}{B_{i,i}} \quad \text{and} \quad \sigma_i^2 = \frac{1}{B_{i,i}}, \tag{2}$$

where $B = (K + \sigma_\varepsilon^2 I)^{-1}$ and $Z = (Z_1, \ldots, Z_n)^\mathrm{T}$. Note that (2) still stands true if $\sigma_\varepsilon^2 = 0$.

Remark 1. Craven and Wahba (1979, Lemma 3.1 and 3.2) show that (2) could be generalized to other types of linear predictors, beyond the particular Gaussian process regression framework considered in this article.

The mean squared prediction error and log-predictive density criteria mentioned in Section 1 correspond respectively to the scoring rules $S_1(P, z) = -(\mathbb{E}_{Z\sim P}(Z) - z)^2$ and $S_2(P, z) = \log(f(z))$, where f denotes the density of P with respect to some common measure. A scoring rule is said *strictly proper* if $\mathbb{E}_{Z\sim P}(S(P, Z)) > \mathbb{E}_{Z\sim P}(S(Q, Z))$ for all $P,\ Q \in \mathcal{P}$ with $P \neq Q$. Strict propriety can be viewed as a sanity condition for performing estimation by maximizing (1). Note that S_1 is not strictly proper relative to the class of Gaussian measures whereas S_2 is. A large variety of scoring rules is surveyed by Gneiting and Raftery (2007). We shall use the CRPS in Section 4 for illustration.

3 Efficient computation of the gradient of a leave-one-out criterion

In this section we present our contribution for computing the gradient $\nabla_\theta L$ of (1). Let[1]

$$\begin{cases} \Gamma : \theta \in \mathbb{R}^q \mapsto K \in \mathbb{R}^{n^2}, \\ \varrho : K \in \mathbb{R}^{n^2} \mapsto (\mu, \sigma^2) \in \mathbb{R}^{2n} \text{ according to (2)}, \\ \varphi : (\mu, \sigma^2) \in \mathbb{R}^{2n} \mapsto L \in \mathbb{R} \text{ according to (1)}, \end{cases} \tag{3}$$

where $\mu = (\mu_1, ..., \mu_n)^\mathrm{T}$ and $\sigma^2 = (\sigma_1^2, ..., \sigma_n^2)^\mathrm{T}$, in such a way that $L(\theta) = (\varphi \circ \varrho \circ \Gamma)(\theta)$. Write $w = (\mu, \sigma^2)$ for simplicity. Let $J_{\varphi,w}$, $J_{\varrho,K}$ and $J_{\Gamma,\theta}$ be the $1 \times 2n$, $2n \times n^2$, $n^2 \times q$ Jacobian matrices of φ, ϱ and Γ at w, K and θ respectively. Using the chain rule for derivation we have

$$\nabla_\theta L^\mathrm{T} = J_{\varphi,w}\, J_{\varrho,K}\, J_{\Gamma,\theta}. \tag{4}$$

Rasmussen and Williams (2006) propose an algorithm in $\mathcal{O}(qn^3)$ time for computing $\nabla_\theta L$ from $J_{\Gamma,\theta}$.

Suppose that these Jacobian matrices are already built and stored. Then, computing (4) by multiplying those matrices from the right to the left costs about $2n \cdot n^2 \cdot q + 1 \cdot 2n \cdot q = \mathcal{O}(qn^3)$ additions and multiplications, corresponding to the complexity announced by Rasmussen and Williams (2006). On the other hand, proceeding from the left to the

[1] We identify the space of $n \times n$ matrices with $\mathbb{R}^{n^2}$ and $(\mathbb{R}^n)^2$ with $\mathbb{R}^{2n}$ with a slight abuse of notation.

right costs about $1 \cdot 2n \cdot n^2 + 1 \cdot n^2 \cdot q = 2n^3 + qn^2$ additions and multiplications. (This kind of consideration is a basic illustration of what has been studied in depth in the literature as the *matrix chain multiplication* problem for variable length products of matrices; see, e.g., Hu and Shing, 1982, and references therein.)

Let us now investigate the price paid for building $J_{\varphi,w}$ and $J_{\varrho,K}$. First of all, the computation of $B = (K + \sigma_\varepsilon^2 I)^{-1}$ and then $w = (\mu, \sigma^2)$ from K can be performed in $\mathcal{O}(n^3)$ operations using (2). Moreover, knowing w, L and $J_{\varphi,w}$ can be computed in $\mathcal{O}(n)$ time. In addition, equations used by Sundararajan and Keerthi (2001) show that $J_{\varrho,K}$ can be build from B in $\mathcal{O}(n^3)$ elementary operations. Thus, previous arguments show that it is indeed possible to compute L and $\nabla_\theta L$ from $J_{\Gamma,\theta}$ and K for $\mathcal{O}(n^3 + qn^2)$ elementary operations, thereby avoiding the $\mathcal{O}(qn^3)$ complexity mentioned by Rasmussen and Williams (2006).

Furthermore, available implementations (see, e.g., Bect et al., 2019) show that it is possible build K and $J_{\Gamma,\theta}$ from θ in a $\mathcal{O}(qn^2)$ complexity for the case of an anisotropic stationary covariance with $q = d + 1$ parameters (one variance parameter and q length scales). We see then that our contribution allows us in this case to keep the evaluation of L and $\nabla_\theta L$ from θ in $\mathcal{O}(n^3 + qn^2)$, rather than $\mathcal{O}(qn^3)$.

The main drawback of this scheme is the $2n \times n^2$ storage of $J_{\varrho,K}$. We propose to circumvent this cost by directly implementing the adjoint operators of the differentials of ϱ:

$$\mathcal{L}_\varrho^* : (K, \delta_w) \mapsto J_{\varrho,K}^{\mathrm{T}} \, \delta_w. \tag{5}$$

This can be used to compute $\mathcal{L}_\varrho^*(K, J_{\varphi,w}^{\mathrm{T}}) = J_{\varrho,K}^{\mathrm{T}} J_{\varphi,w}^{\mathrm{T}}$ and then $\nabla_\theta L$ from (4). This way of implementing chain rule derivatives is well known and has been studied under the name of *reverse-mode differentiation*[2] or *backpropagation*, and its paternity can be traced back at least to Linnainmaa (1970).

We propose Algorithm 1 to implement this operator. This algorithm only requires $\mathcal{O}(n^2)$ storage capacity and about $2n^3$ additions and multiplications, thus reducing the burden of storage, while maintaining the global $\mathcal{O}(n^3 + qn^2)$ complexity. (Note that $2n^3$ already corresponds to the cost of matrix multiplication in a "direct" approach that would first build $J_{\varrho,K}$ and then compute $\mathcal{L}_\varrho^*(K, \delta_w)$ by matrix multiplication.)

Remark 2. The algorithm can easily be adapted, through a suitable modification of the matrix B used in Step 6, to any type of linear model for which (2) holds (see Remark 1).

4 Conclusion and perspectives

We suggested using the scoring rules referenced by Gneiting and Raftery (2007) for the estimation of the parameters of a Gaussian process by leave-one-out cross-validation. We also proposed an efficient procedure for computing gradients of cross-validation criteria that is more efficient than what was available in the literature to our knowledge.

[2] In the context of Gaussian process regression, a reverse-mode differentiation approach has been proposed by Toal et al. (2009) for the computation of the likelihood function and its gradient.

Algorithm 1: Implementation of $\mathcal{L}_\varrho^*$ for computing $\delta_K = J_{\varrho,K}^{\mathrm{T}} \delta_w$, from K and δ_w. Inputs at first step refer to what has already been computed for evaluating μ and σ^2. For vectors a and b, $a \oslash b$ and $a \circ b$ denote the Hadamard element-wise division and multiplication respectively.

Input:
K, Z, $B = (K + \sigma_\varepsilon^2 I)^{-1}$, $\alpha = By$, $\kappa = (B_{i,i})_{1 \leq i \leq n}$, $\kappa^{-1} = \mathbb{1} \oslash \kappa$, $\chi = \alpha \circ \kappa^{-1}$, $\delta_w = (\delta_\mu, \delta_{\sigma^2})$
Output: $\delta_K = J_{\varrho,K}^{\mathrm{T}} \delta_w$

$\kappa^{-2} = \kappa^{-1} \circ \kappa^{-1}$
$\delta_\chi = -\delta_\mu$
$\delta_\alpha = \delta_\chi \circ \kappa^{-1}$
$\delta_\kappa = -\delta_\chi \circ \alpha \circ \kappa^{-2} - \delta_{\sigma^2} \circ \kappa^{-2}$
$\delta_B = \delta_\alpha Z^{\mathrm{T}} + \mathrm{diag}(\delta_\kappa)$
$\delta_K = -B^{\mathrm{T}} \, \delta_B \, B^{\mathrm{T}}$

Further work will consist in investigating the properties of these estimators for several scoring rules. For instance, one can choose to use the continous rank probability score (CRPS) defined as $\mathrm{CRPS}(F, z) = -\int_{-\infty}^{+\infty} \left(F(u) - \mathbb{1}_{z \leq u}\right)^2 \mathrm{d}u$, where F is a cumulative distribution function. The CRPS is strictly proper relative to the class of Gaussian measures[3]. An empirical comparison with maximum likelihood for estimating the length scales is presented in Figure 1. Our contribution for computing gradients makes it possible to maintain the same complexity, both in terms of storage and calculation, for the two methods.

References

D. M. Allen. The relationship between variable selection and data agumentation and a method for prediction. *Technometrics*, 16(1):125–127, 1974.

F. Bachoc. Cross validation and maximum likelihood estimation of hyper-parameters of Gaussian processes with model misspecification. *Computational Statistics and Data Analysis*, 66:55–69, 2013.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, USA, 2006.

O. Dubrule. Cross validation of kriging in a unique neighborhood. *Journal of the International Association for Mathematical Geology*, 15:687–699, 1983.

[3] and more generally with respect to the class of all probability measures with finite first order moment (see, e.g. Gneiting and Raftery, 2007, Section 4.2)

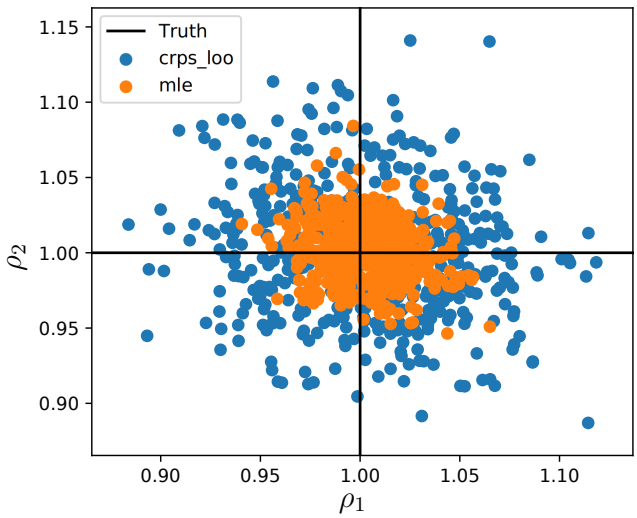


Figure 1: Scatterplots of the estimates of the two length scales of a Gaussian process on $\mathbb{R}^2$. Blue points correspond to CRPS-based cross-validation estimates; orange points correspond to maximum likelihood estimates. True length scales are represented by black lines. Each scatterplot consists of 500 estimations obtained from a space filing design of size $n = 500$. The criteria were optimized using a quasi-Newton type algorithm.

T. Gneiting and A. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102:359–378, 2007.

S. Sundararajan and S. S. Keerthi. Predictive approaches for choosing hyperparameters in Gaussian processes. *Neural Computation*, 13(5):1103–1118, 2001.

P. Craven and G. Wahba. Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, 31:377–404, 1979.

T. C. Hu and M. T. Shing. Computation of matrix chain products. Part I. *SIAM Journal on Computing*, 11(2):362–373, 1982.

J. Bect, E. Vazquez, et al. STK: a Small (Matlab/Octave) Toolbox for Kriging. Release 2.6, 2019. URL http://kriging.sourceforge.net.

D. J. J. Toal, A. I. J. Forrester, N. W. Bressloff, A .J. Keane, and C. Holden. An adjoint for likelihood maximization. *Proceedings of The Royal Society A: Mathematical, Physical and Engineering Sciences*, 465, 2009.

S. Linnainmaa. The representation of the cumulative rounding error of an algorithm as a Taylor expansion of the local rounding errors. Master's thesis, Univ. Helsinki, 1970.

PIntMF : Une méthode de factorisation matricielle pénalisée pour l'intégration de données multi-omiques

Morgane Pierre-Jean [1], Florence Mauger, [1] Jean-François Deleuze [1] et Edith Le Floch [1]

[1] *Université Paris-Saclay, CEA, Centre National de Recherche en Génomique Humaine, 91057, Evry, France. morgane.pierre-jean@cng.fr*

Résumé. La génération de données multi-omiques est en pleine expansion avec l'amélioration des technologies à haut débit. L'intégration au sein d'une seule analyse de plusieurs sources d'information du génome pourrait permettre une meilleure compréhension des maladies ou des systèmes biologiques. Nous proposons ici une méthode non-supervisée de factorisation matricielle pénalisée multi-blocs pour intégrer des données multi-omiques. Cette méthode a pour but d'identifier de nouveaux groupes d'individus au sein d'une même maladie ainsi que d'identifier les variables pertinentes conduisant à cette classification. Nous avons appliqué cette méthode sur des données simulées pour comparer ses performances à des méthodes intégratives non-supervisées existantes et sur des données réelles. Cette nouvelle méthode permet de bien classer les individus et d'identifier les variables liées aux groupes avec plus de précision. Sur les données réelles, la méthode permet d'établir une nouvelle classification qui a un lien avec la survie des patients.

Mots-clés. Méthode non-supervisée, factorisation matricielle, multi-omique, classification

Abstract. The generation of multi-omics data is growing with the improvement of high-throughput technologies. The integration in the same analysis of several levels of the genome could allow a better understanding of diseases or biological systems. Here, we propose a non-supervised penalized matrix factorization method to integrate multi-omics data which aims to identify new groups of individuals within the same disease, as well as relevant markers leading to this classification. We applied this method to simulated data and compared its performances with existing integrative unsupervised methods. Our method leads to a correct clustering of individuals and identifies relevant biomarkers with more precision. The results on real data highlight a new clustering linked to the patient's survival.

Keywords. Unsupervised method, Matrix factorization, multi-omics, clustering

1 Contexte

Grâce au développement de technologies à haut débit, pour un même échantillon, la génération de plusieurs types de données "omiques" (génomique, transcriptomique, épigénomique, protéomique et métabolomique) se développe de plus en plus. En effet, le TCGA (The Cancer Genome Atlas, Weinstein et al. (2013)) a généré, pour un grand nombre de cancers et pour plusieurs échantillons, à la fois des données d'ADN, d'ARN, de méthylation, et même de protéomique. Depuis la dernière décennie, des méthodes d'intégration non-supervisées ont été développées pour analyser les données multi-omiques (Chauvel et al. (2019); Pierre-Jean et al. (2019); Cantini et al. (2020)). D'un point de vue mathématique, les données omiques peuvent être considérées comme des matrices et les variables pertinentes peuvent être extraites à l'aide de méthodes de factorisation matricielle. L'analyse en composantes principales (ACP, Hotelling (1933)) et la factorisation matricielle non négative (NMF (Non-Negative Matrix factorization), Lee and Seung (1999)) sont deux méthodes courantes de factorisation matricielle sous contraintes permettant de classer les échantillons et de mettre en évidence des variables pertinentes (Burstein et al. (2015)). L'ACP est une méthode puissante de réduction de dimensions et de visualisation des données tandis que la NMF qui n'impose pas l'orthogonalité des variables latentes met en évidence des groupes de variables associées aux groupes. Plus récemment, des extensions multi-blocs de NMF ont été développées pour permettre une analyse intégrative des données multi-omiques (Mo et al. (2013); Chalise et al. (2014); Chen and Zhang (2018)).

Ici, nous proposons un méthode intégrative de factorisation matricielle pénalisée (PIntMF : Penalized Integrative Matrix Factorization) dans le but d'intégrer des données multi-omiques. Nous avons comparé

cette méthode à plusieurs méthodes d'intégration existantes sur des simulations et sur des données réelles du TCGA. Un package `R` implémentant la méthode est également disponible sur github.

2 PIntMF : Penalized Integrative Matrix Factorization

2.1 Modèle

Dans cette section, nous décrivons le modèle PIntMF ainsi que sa résolution. On considère K matrices $\mathbf{X}^1, \dots \mathbf{X}^K$ en entrée du modèle. Chaque matrice $\mathbf{X}^k$ est de taille $n \times J_k$ (où n est le nombre d'individus et J_k le nombre de variables dans chaque bloc k. Nous proposons ici un modèle basé sur la factorisation matricielle de chaque bloc de données k i.e :

$$\mathbf{X}^k \approx \mathbf{W}\mathbf{H}^k \tag{1}$$

où $\mathbf{W}$ désigne une matrice de base commune et $\mathbf{H}^k$ une matrice spécifique à chaque bloc k. $\mathbf{W}$ est de taille $n \times P$ et $\mathbf{H}^k$ est de taille $P \times J_k$. Par conséquent, la variable P représente le nombre de variables latentes dans le modèle.

Nous imposons des constraintes de positivité sur les coefficients de la matrice $\mathbf{W}$ (comme dans un modèle NMF classique). Une contrainte de parcimonie a été ajoutée afin d'améliorer l'interprétation du modèle d'un point de vue de la classification et de diminuer sa complexité. La matrice $\mathbf{W}$ est utilisée pour faire une classification des individus en utilisant les K blocs de données omiques simultanément. La classification finale est une classification hierarchique ascendante classique avec la distance de Ward sur la matrice $\mathbf{W}$.

Sur $\mathbf{H}^k$, une contrainte de parcimonie est ajoutée également pour sélectionner des variables pertinentes.

Nous résumons les contraintes décrites précédemment au problème d'optimisation suivant :

$$\min_{\mathbf{W},\mathbf{H}^1,\dots,\mathbf{H}^k} \sum_{k=1}^{K} \|\mathbf{X}^k - \mathbf{W}\mathbf{H}^k\|_F^2 + \lambda_k \|\mathbf{H}^k\|_1 + \sum_{i=1}^{n} \mu_i \|\mathbf{w}_{i\bullet}\|_1 \quad \text{st.} \quad \mathbf{W} \geq 0 \tag{2}$$

where $\|\mathbf{H}^k\|_1 = \sum_{p=1}^{P} \sum_{j=1}^{J_k} |h_{pj}^k|$.

Le problème d'optimisation décrit ci-dessus (Equation 2) n'est pas jointement-convexe sur $\mathbf{W}, \mathbf{H}_1, \dots, \mathbf{H}_k$, mais est convexe séparément sur chacune des matrices. Par conséquent, le problème peut être résolu alternativement sur $\mathbf{W}$, et $\mathbf{H}^1, \dots, \mathbf{H}^k$ jusqu'à convergence de la fonction $g(\mathbf{X}^1, \dots, \mathbf{X}^K, \mathbf{W}, \mathbf{H}^1, \dots, \mathbf{H}^K) = \sum_{k=1}^{K} \|\mathbf{X}^k - \mathbf{W}\mathbf{H}^k\|_F^2$.

2.2 Résolution de W

Dans cette étape, les $\mathbf{H}^k$ sont fixés et l'équation 2 est résolue sur $\mathbf{W}$.

$$\min_{\mathbf{W}} \sum_{k=1}^{K} \|\mathbf{X}^k - \mathbf{W}\mathbf{H}^k\|_F^2 + \sum_{i=1}^{n} \mu_i \|\mathbf{w}_{i\bullet}\|_1 \quad \text{st.} \quad \mathbf{W} \geq 0 \tag{3}$$

Tous les lignes de la matrices $\mathbf{W}$ sont indépendantes quand les $\mathbf{H}^k$ sont fixes. Le problème pour une ligne (individu) i peut être écrit de la façon suivante :

$$\min_{\mathbf{w}_{i\bullet}} \sum_{k=1}^{K} \|\mathbf{x}_{i\bullet}^k - \mathbf{w}_{i\bullet}\mathbf{H}^k\|^2 + \mu_i \|\mathbf{w}_{i\bullet}\|_1 \quad \text{st.} \quad \mathbf{w}_{i\bullet} \geq 0 \tag{4}$$

Le problème d'optimisation décrit par l'équation 4 est un problème Lasso classique avec une contrainte de positivité. Nous avons fixé μ_i à 1 mais il serait intéressant de calibrer la contrainte par validation croisée. Il peut être facilement et rapidement résolu en utilisant le package `glmnet R` développé par Friedman et al. (2010).

2.3 Résolution de $\mathbf{H}^k$

Quand $\mathbf{W}$ est fixé, les $\mathbf{H}^k$ peuvent être résolus indépendamment les uns des autres. Pour faciliter la lecture de cette section, l'indice k a été retiré des équations.

$$\min_{\mathbf{H}} Q(\mathbf{H}) = \min_{\mathbf{H}} \|\mathbf{X} - \mathbf{WH}\|_F^2 + \lambda \sum_{p=1}^{P} \sum_{j=1}^{J} |h_{pj}| \tag{5}$$

$$\begin{aligned} Q(\mathbf{H}) = &\quad \text{Trace}\left\{(\mathbf{X} - \mathbf{WH})(\mathbf{X} - \mathbf{WH})^T\right\} + \lambda \textstyle\sum_{p=1}^{P} \sum_{j=1}^{J} |h_{pj}| \\ = &\quad vec(\mathbf{X} - \mathbf{WH})^T vec(\mathbf{X} - \mathbf{WH}) + \lambda \textstyle\sum_{p=1}^{P} \sum_{j=1}^{J} |h_{pj}| \end{aligned}$$

$$\text{On note } \mathbf{h} = vec(\mathbf{H}) = \begin{pmatrix} \mathbf{H}_{11} \\ \vdots \\ \mathbf{H}_{P1} \\ \vdots \\ \mathbf{H}_{1J} \\ \vdots \\ \mathbf{H}_{PJ} \end{pmatrix} \text{ et } \mathbf{x} = vec(\mathbf{X}) = \begin{pmatrix} \mathbf{X}_{11} \\ \vdots \\ \mathbf{X}_{n1} \\ \vdots \\ \mathbf{X}_{1J} \\ \vdots \\ \mathbf{X}_{nJ} \end{pmatrix}.$$

$$\begin{aligned} Q(\mathbf{H}) = &\quad (\mathbf{x} - vec(\mathbf{WH}))^T (\mathbf{x} - vec(\mathbf{WH})) + \lambda \|\mathbf{h}\|_1 \\ = &\quad (\mathbf{x} - (\mathbb{I}_n \otimes \mathbf{W}) vec(\mathbf{H}))^T (\mathbf{x} - (\mathbb{I}_n \otimes \mathbf{W}) vec(\mathbf{H})) + \lambda \|\mathbf{h}\|_1 \\ Q(\mathbf{h}) = &\quad (\mathbf{x} - \tilde{\mathbf{W}}\mathbf{h})^T (\mathbf{x} - \tilde{\mathbf{W}}\mathbf{h}) + \lambda \|\mathbf{h}\|_1 \end{aligned}$$

où $\mathbf{h} = vec(\mathbf{H})$, $\mathbf{x} = vec(\mathbf{X})$, $\mathbb{I}_n$ est la matrice identité de taille n et $\tilde{\mathbf{W}} = \mathbb{I}_n \otimes \mathbf{W}$

On peut reformuler le problème de la façon suivante : $Q(\mathbf{h}) = \|\mathbf{x} - \tilde{\mathbf{W}}\mathbf{h}\|^2 + \lambda \|\mathbf{h}\|_1$

Minimiser $Q(\mathbf{h})$ revient donc à résoudre un problème Lasso classique. La valeur de λ sera optimisée par validation croisée pour chaque bloc $k = 1, \ldots, K$.

Comme pour $\mathbf{W}$ on utilise le package `glmnet` pour résoudre ce problème.

2.4 Normalisation de W

Afin d'interpréter la matrice $\mathbf{W}$ comme une matrice de poids, nous avons ajouté une contrainte sur la somme des coefficients de chaque ligne de $\mathbf{W}$. Pour éviter des problèmes de convergence ou de non-identifiabilité, la contrainte sur la somme est ajoutée après avoir estimé la matrice $\mathbf{W}$. On divise chaque ligne par sa somme après chaque étape :

$$\mathbf{w}_{i\bullet} = \frac{\mathbf{w}_{i\bullet}}{\sum_{p=1}^{P} \mathbf{w}_{ip}} \tag{6}$$

2.5 Initialisation de l'algorithme

La méthode de Wang et al. (2014) (Similarity Network Fusion : SNF) a déjà montré de bonnes performances pour la classification en multi-omiques. Nous avons choisi d'initialiser $\mathbf{W}$ avec la classfication avec P groupes de cet algorithme. $\mathcal{C}_p$ désigne le groupe p de SNF. Ainsi,

$$w_{ip} = \begin{cases} 1 & \text{si } i \in \mathcal{C}_p \\ 0 & \text{Sinon} \end{cases} \tag{7}$$

Cette initialisation a l'avantage de prendre en compte simultanément les K blocs dans la classification.

3 Résultats

3.1 Simulations

Nous avons évalué les performances du modèle présenté dans la section précédente sur le même schéma de simulation que dans l'article de Pierre-Jean et al. (2019). Ce schéma simule trois blocs de données hétérogènes (sous des distributions différentes : Binaire, Beta et Gaussienne). Quatre groupes déséquilibrés respectivement composés de 25, 20, 5 et 10 individus ont été simulés dans ces trois blocs. PIntMF a été comparé à 6 méthodes existantes qui donnaient les meilleurs résulats dans Pierre-Jean et al. (2019) à savoir : SNF de Wang et al. (2014), intNMF de Chalise et al. (2014), SGCCA de Tenenhaus et al. (2014), MoCluster de Meng et al. (2015), iClusterPlus de Mo et al. (2013), et CIMLR de Ramazzotti et al. (2018). Les performances des méthodes ont été évaluée à la fois sur la capacité des méthodes à retrouver la classification simulée mais également sur la capacité à retrouver les variables simulées associées à la classification.

Évaluation des performances sur le clustering

Les performances de la classification ont été évaluées en utilisant le Rand Index Ajusté (ou Adjusted Rand Index (ARI), Hubert and Arabie (1985)) sur 4 Benchmarks de simulation. L'ARI est égal à 1 si la partition trouvée est la même que celle simulée et 0 si elle est "complètement" différente.

Sur les 4 benchmarks simulés, PIntMF, MoCluster et SNF ont des meilleures performances que les autres méthodes avec un ARI égal à 1 dans la plupart des cas (Fig. 1a). On peut noter que PIntMF ne fait aucune erreur de classification sur les Benchmarks 1, 2 et 4.

Sélection de variables

La performance sur la sélection de variables a été évaluée en utilisant les taux de faux positifs (TFP) et de vrais positifs (TVP). Nous avons résumé les TFP et TVP en traçant des courbes ROC puis en calculant l'AUC (Area under the curve) associé.

Nous ne pouvons pas évaluer les performances SNF car la méthode est basée sur des matrices de similarité qui ne donnent pas de poids aux variables dans les blocs.

Le calcul de l'AUC (Figure 1b) montre que PIntMF est une méthode qui obtient des performances très satisfaisantes et similaires à MoCluster sur les trois types de blocs. SGCCA et CIMLR ont des performances légèrement inférieures en moyenne sur les 3 types de distribution. intNMF ne semble pas adapté aux données simulées sous la distribution beta et iClusterPlus ne parvient pas à retrouver les variables pertinentes sous la distribution Binaire.

3.2 Analyses sur données réelles

Nous avons également utilisé PIntMF sur des données réelles de glioblastome issues de la base de données du TCGA. Ce jeu de données contient 55 individus, avec des données d'expression (1740 gènes), de nombre de copies d'ADN (1599 gènes) et de méthylation (1515 gènes). En accord avec les différentes catégories des tumeurs, nous avons sélectionné 5 variables latentes dans le modèle PIntMF. La heatmap (Figure 1c) représente pour chacun des échantillons (en colonne) la valeur pour chaque variable latente. Nous avons superposé la classification des types de glioblastome et la classification trouvée par PIntMF. La classification est légèrement différente mais semble cohérente avec celle existante dans les données cliniques. Pour les individus non-classés (NA), on pourrait les assigner, grâce à PIntMF, à des groupes déjà existants. Une analyse de survie avec les groupes trouvés par PIntMF a également été réalisée (Figure 1d) et montre que les groupes ont des taux de survies différents (p-value significative à 5%). En particulier, le groupe bleu a un très mauvais pronostic alors que la survie du groupe orange est beaucoup plus favorable.

4 Conclusions et perspectives

Pour conclure, PIntMF donne de bonnes performances sur les simulations que ce soit pour la classification ou la sélection de variables. Sur les données réelles, il semble que les groupes permettent de catégoriser des patients qui n'ont pas été assignés à une classe de tumeur particulière. La méthode révèle aussi des taux survie différents par groupe.

Contrairement aux autres méthodes qui utilise des pénalités (MoCluster, SGCCA, iClusterPlus), un avantage de la résolution de PIntMF est que les valeurs des pénalités sur les matrices $\mathbf{H}^k$ sont automatiquement ajustées. Ainsi, les utilisateurs ont seulement besoin de définir le nombre de variables latentes.

Nous avons également implémenté un package R (nommé PIntMF) pour reproduire les résultats de cet article. Ce package est disponible sur github.

Enfin, il faudrait encore améliorer l'automatisation du choix du nombre de variables latentes. En effet, ceci rendrait la méthode encore plus facile d'utilisation dans le cadre d'applications en génomique. On pourrait également optimiser la contrainte de parcimonie sur la matrice $\mathbf{W}$.

Références

John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, Joshua M Stuart, Cancer Genome Atlas Research Network, et al. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10) :1113–1120, 2013.

Cécile Chauvel, Alexei Novoloaca, Pierre Veyre, Frédéric Reynier, and Jérémie Becker. Evaluation of integrative clustering methods for the analysis of multi-omics data. *Briefings in Bioinformatics*, 2019.

Morgane Pierre-Jean, Jean-Francois Deleuze, Edith Le Floch, and Florence Mauger. Clustering and variable selection evaluation of 13 unsupervised methods for multi-omics data integration. *Briefings in bioinformatics*, 2019.

Laura Cantini, Pooya Zakeri, Celine Hernandez, Aurelien Naldi, Denis Thieffry, Elisabeth Remy, and Anais Baudot. Benchmarking joint multi-omics dimensionality reduction approaches for cancer study. *bioRxiv*, 2020.

Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6) :417, 1933.

Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755) :788, 1999.

Matthew D Burstein, Anna Tsimelzon, Graham M Poage, Kyle R Covington, Alejandro Contreras, Suzanne AW Fuqua, Michelle I Savage, C Kent Osborne, Susan G Hilsenbeck, Jenny C Chang, et al. Comprehensive genomic analysis identifies novel subtypes and targets of triple-negative breast cancer. *Clinical Cancer Research*, 21(7) : 1688–1698, 2015.

Qianxing Mo, Sijian Wang, Venkatraman E Seshan, Adam B Olshen, Nikolaus Schultz, Chris Sander, R Scott Powers, Marc Ladanyi, and Ronglai Shen. Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proceedings of the National Academy of Sciences*, 110(11) :4245–4250, 2013.

Prabhakar Chalise, Devin C Koestler, Milan Bimali, Qing Yu, and Brooke L Fridley. Integrative clustering methods for high-dimensional molecular data. *Translational cancer research*, 3(3) :202, 2014.

Jinyu Chen and Shihua Zhang. Discovery of two-level modular organization from matched genomic data via joint matrix tri-factorization. *Nucleic acids research*, 46(12) :5967–5976, 2018.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1) :1–22, 2010. URL `http://www.jstatsoft.org/v33/i01/`.

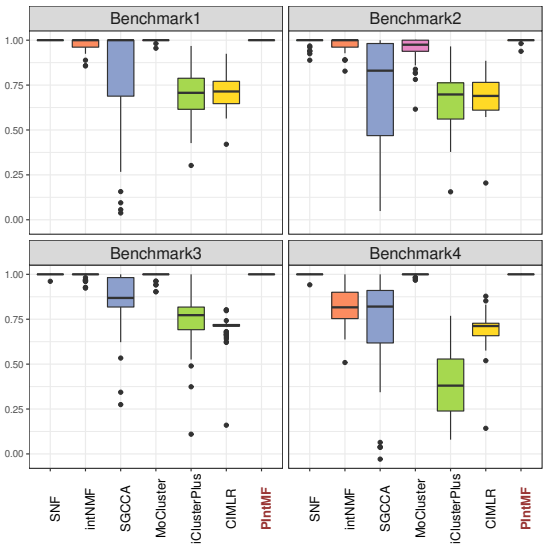
Bo Wang, Aziz M Mezlini, Feyyaz Demir, Marc Fiume, Zhuowen Tu, Michael Brudno, Benjamin Haibe-Kains, and Anna Goldenberg. Similarity network fusion for aggregating data types on a genomic scale. *Nature methods*, 11 (3) :333, 2014.

Arthur Tenenhaus, Cathy Philippe, Vincent Guillemot, Kim-Anh Le Cao, Jacques Grill, and Vincent Frouin. Variable selection for generalized canonical correlation analysis. *Biostatistics*, 15(3) :569–583, 2014.

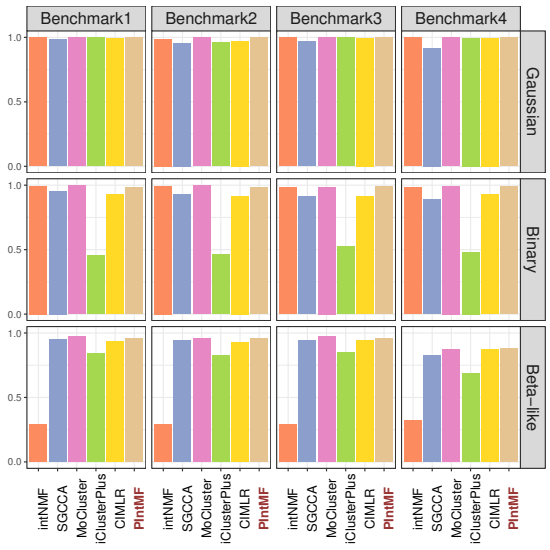
Chen Meng, Dominic Helm, Martin Frejno, and Bernhard Kuster. mocluster : Identifying joint patterns across multiple omics data sets. *Journal of proteome research*, 15(3) :755–765, 2015.

Daniele Ramazzotti, Avantika Lal, Bo Wang, Serafim Batzoglou, and Arend Sidow. Multi-omic tumor data reveal diversity of molecular mechanisms that correlate with survival. *Nature communications*, 9(1) :4453, 2018.

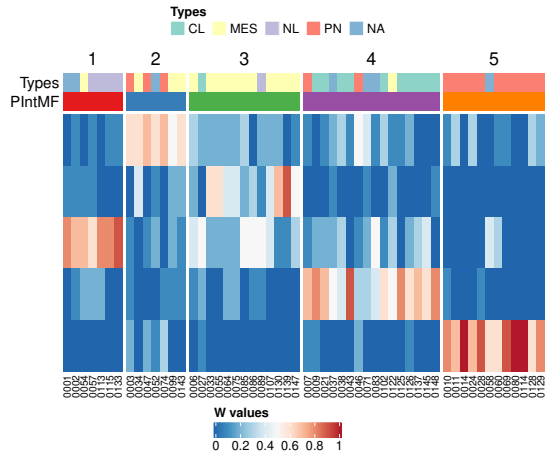
Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1) :193–218, 1985.



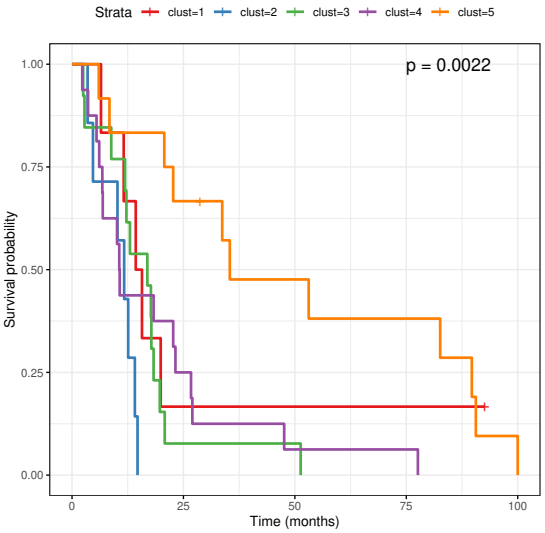
(a) ARI (Adjusted Rand Index)



(b) AUC (Area Under ROC Curve)



(c) Matrice $\mathbf{W}$ de PIntMF



(d) Courbes de survie

FIGURE 1 – (a) Rand Index Ajusté (ARI). (b) Aire sous la courbe ROC pour SNF, intNMF, SGCCA, MoCluster, iClusterPlus, CIMLR et PIntMF sur 4 Benchmarks de simulations. (c) Matrice $\mathbf{W}$ inferrée par PIntMF sur des données de glioblastome. (d) Courbes de survie pour chacun des groupes inferrés par PIntMF.

TEST PAR SIMULATION/CALIBRATION POUR LA SELECTION DE VARIABLES PAR LASSO

Matthieu Pluntz[1], Cyril Dalmasso[2], Pascale Tubert-Bitter[1], Ismaïl Ahmed[1]

[1]Biostatistiques en Grande Dimension, Centre de Recherche en Épidémiologie et Santé des Populations (CESP), INSERM, Université Paris-Saclay, 16, avenue Paul Vaillant-Couturier 94807 Villejuif, France; [2]Laboratoire de Mathématiques et Modélisation d'Evry (LaMME), Université d'Evry Val d'Essonne, CNRS, Evry, France.
E-mails : matthieu.pluntz@inserm.fr, cyril.dalmasso@univ-evry.fr, pascale.tubert@inserm.fr, ismail.ahmed@inserm.fr

Résumé. Dans le contexte des régression linéaires ou linéaires généralisées en grande dimension, le Lasso est une méthode de référence qui permet de sélectionner un nombre restreint de variables. Celles-ci apparaissent successivement dans le modèle au fur à mesure que le paramètre de régularisation de la méthode diminue. Nous proposons un test de significativité empirique qui s'applique à chaccune de ces variables.

L'hypothèse nulle de ce test est paramétrée par un ensemble de variables explicatives et énonce que toutes les variables actives font partie de cet ensemble. Rejeter cette hypothèse conduit à sélectionner la première variable n'appartenant pas à cet ensemble qui soit sélectionnée par le Lasso. La statistique à laquelle s'intéresse le test est la valeur du paramètre de régularisation à laquelle cette variable apparaît. On estime la p-value par une simulation de cette statistique sous l'hypothèse nulle, où intervient une étape de calibration des données simulées sur un modèle cible. Nous prouvons que cette p-value suit une loi uniforme sous l'hypothèse nulle dans le cas gaussien, et les simulations suggèrent que c'est aussi le cas dans le cas binaire logistique. Elles permettent aussi de mesurer les performances du test lorsqu'il est appliqué successivement à une suite croissante densembles de variables dans une optique de sélection de variables, et de comparer ces performances avec d'autres méthodes, dont le covTest de Lockhart et al.

Mots-clés. Régression en grande dimension, Lasso, selection de variables, test statistique, p-value, méthode de Monte-Carlo

Abstract. The Lasso is a popular method of sparse regression which allows variable selection in high-dimensional settings such as genomic or drug exposure data. It produces an order of appearance of successive variables into the model as its regularization parameter decreases. We propose an empirical significance test for each of the appearing variables.
The null hypothesis of the test is that all active variables belong to a given set. Rejecting it means selecting the first variable selected by the Lasso which does not belong to this set. The statistic of interest is the regularization parameter at which any this variable is selected. The p-value is estimated by simulating this statistic under the null hypothesis, which involves a technique where simulated data is calibrated on a given model. We prove that the estimated p-value ais distributed uniformly under the null hypothesis in the Gaussian model, and simulation studies indicate that it is also the case in the logistic models. The studies also allow us to measure the test's performances when it is iterated over a growing sequence of sets in a variable selection purpose, and compare it with other methods, including Lockhart et al's covTest.

Keywords. High dimensional regression, Lasso, variable selection, significance test, p-value, Monte-Carlo method

1. Introduction

Soient $Y \in \mathbf{R}^n$ un vecteur réponse et $X \in \mathbf{R}^{n \times p}$ une matrice de variables explicatives. On considère le modèle linéaire généralisé :

$$E[Y] = f(X\beta).$$

f pourra être l'identité (modèle gaussien $= X\beta + \epsilon,\ \epsilon \sim N(0, \sigma^2 I_n)$) ou une fonction logistique dans le cas d'une réponse binaire.

On estime le vecteur de coefficients $\beta \in \mathbf{R}^p$ par la méthode Lasso (Tibshirani 1996). Étant donné un paramètre de régularisation $\lambda \geq 0$, on définit :

$$\hat{\beta}^{Lasso}(\lambda) = \text{argmin}_\beta \left(-\log \text{likelihood}(y, X\beta) + \lambda \left|\left|\beta\right|\right|_1 \right).$$

Si λ est suffisament élevé, $\hat{\beta}^{Lasso}(\lambda) = 0$: aucune variable n'est sélectionnée. En faisant diminuer λ, des variables sont sélectionnées une à une par la méthode : elles vérifient $\hat{\beta}_j^{Lasso}(\lambda) \neq 0$. Notre objectif est d'utiliser les résultats du lasso pour déterminer si l'association entre Y et chacune de ces variables X_j est statistiquement significative.

Lockhart et al (2014) ont proposé un test de significativité fondé sur la différence de covariance entre Y et ses projections $X\hat{\beta}^{Lasso}(\lambda)$ pour différentes valeur de λ. Leur statistique de test possède une loi asymptotique connue dans le cas gaussien.

Nous proposons un test empirique pour lequel nous déterminons la loi exacte de la statistique de test dans le cas gaussien, et pour lequel nous observons la même loi dans le cas binaire.

2. Méthode

2.1. Notions utilisées

Soit A un sous-ensemble de $\{1, \ldots, p\}$. On suppose que les variables $X_j, j \in A$, sont actives, et on cherche à savoir si d'autres variables le sont aussi. Considérons l'hypothèse nulle et l'hypothèse alternative suivantes :

$$H_0(A) : \forall j \notin A, \beta_j = 0$$

$$H_1(A) : \exists j \notin A, \beta_j \neq 0$$

ainsi que la quantité :

$$\lambda_A = \sup \left\{ \lambda \geq 0 : \exists j \notin A, \widehat{\beta_j}^{Lasso}(\lambda) \neq 0 \right\}.$$

Dans une optique de sélection de variables, on cherche à rejeter $H_0(A)$ lorsque λ_A est trop élevé pour être compatible avec cette hypothèse. Dans ce cas, on conclura que la variable sélectionnée par le Lasso à partir de $\lambda = \lambda_A$, qui est donc responsable de la valeur élevée de λ_A, est une variable active.

$\hat{\beta}(\lambda)$ et λ_A dependent de X et de Y. Dans la suite, X est considéré comme une constante mais Y pourra être remplacé par d'autres vecteurs réponses, c'est pourquoi on note $\lambda_A = \lambda_A(Y)$.

2.2. Simulation-calibration

On s'intéresse à une statistique de test directement liée à λ_A qui approxime la quantité $\mathrm{P}_{H_0(A)}\big(\lambda_A(Y') \geq \lambda_A(Y)\big)$. Il s'agit d'une p-value empirique estimée par une méthode de Monte-Carlo :

$$\widehat{p_A}(Y) = \frac{1}{N}\sum_{l=1}^{N} \mathbf{1}\{\lambda_A\big(Y^{(l)}\big) \geq \lambda_A(Y)\}.$$

Dans la formule ci-dessus, les $Y^{(l)}$ sont des réponses du modèles simulées sous $H_0(A)$. De plus, on applique à ces simulations une étape de *calibration* qui vise à leur faire reproduire la relation linéaire (ou linéaire généralisée) existant entre Y et les $X_j, j \in A$; c'est-à-dire, à faire en sorte que l'estimateur par maximum de vraisemblance du modèle restreint à A $\widehat{\theta_A}$ donne le même résultat quand on l'applique à Y où à l'un des $Y^{(l)}$. Dans le cas gaussien, cela est possible en appliquant une transformation linéaire aux $Y^{(l)}$. Dans le cas binaire, une procedure itérative stochastique est nécessaire.

2.3. Résultats théoriques dans le cas du modèle gaussien

Dans le cas gaussien, sous $H_0(A)$, $\widehat{p_A}(Y)$ suit bien une loi uniforme discrète qui tend vers une loi uniforme sur $[0,1]$ quand on augmente le nombre de simulations. La probabilité de sélectionner à tort une variable inactive en rejetant $H_0(A)$ est donc contrôlée.

De plus, cette probabilité est également contrôlée quand $H_0(A)$ n'est pas vérifiée et qu'il existe des variables actives en dehors de A, à condition que les variables actives soient orthogonales aux variables inactives.

2.4. Procédure itérative de sélection de modèle

En pratique, pour sélectionner un sous-ensemble de variables actives dans le cadre d'une régression en grande dimension, on applique la procédure itérative suivante, qui dépend d'un paramètre α.

1. On effectue la régression Lasso de Y sur X et on pose $A = \emptyset$.
2. De façon itérée :
    1. On extrait la quantité $\lambda_A(Y)$ des résultats du Lasso et on calcule $\widehat{p_A}(Y)$.
    2. Si $\widehat{p_A}(Y) \leq \alpha$, on pose $A = A \cup \{j\}$ où j est la variable sélectionnée par le Lasso à partir de $\lambda = \lambda_A(Y)$, et on répète l'itération.
    3. Si $\widehat{p_A}(Y) > \alpha$, on interrompt la procédure.

3. Étude par simulation

3.1. Démarche

Nous avons mesuré les performances de notre méthode dans 189 situations différentes, caractérisées par les paramètres suivants :

- Type de modèle : gaussien ou binaire, celui-ci se décomposant en binaire “dense” (espérance de Y égale à 1) ou “creux” (espérance égale à 0.1)
- Structure de la matrice : dans tous les cas, X possède $n = 1000$ individus et $p = 500$ variables, chaque variable suit une loi normale centrée réduite et leur matrice de covariance est une matrice de Toeplitz. Les trois structures possibles dépendent de la corrélation entre deux variables adjacentes : 0, 0.9 ou 0.99.
- Nombre de variables actives : 0, 1, 2, 5 ou 10.
- Lorsqu’il y a au moins une variable active, rapport signal/bruit égal à $SNR = \mathrm{Var}(\mathrm{E}[Y|X]) \,/\, \mathrm{E}[\mathrm{Var}(\mathrm{Y}|\mathrm{X})]$.

Dans tous les cas, nous avons effectué le test par simulation-calibration avec $N = 200$ simulations. Nous avons observé :

1. La distribution des $\widehat{p_A}(Y)$ en prenant pour A le véritable ensemble de variables actives, de façon à satisfaire $H_0(A)$;
2. Les résultats de la procédure itérative de sélection de modèle pour des valeurs de α variant sur l’intervalle $[0,0.95]$. Dans chaque cas nous avons calculé la sensibilité et la spécificité de la procédure de sélection façon à décrire sa performance globale par des courbes ROC.

3.2. Résultats

Les $\widehat{p_A}(Y)$ que nous avons simulés sous $H_0(A)$ suivent bien la loi uniforme attendue. Des tests de Kolmogorov-Smirnov ne permettent pas de les distinguer d’un échantillon tiré selon une loi uniforme sur $[0,1]$, et ce, y compris dans le cas des modèles binaires (alors que la théorie l’assure uniquement dans le cas gaussien). Nous observons aussi que l’étape de calibration dans la simulation des $Y^{(l)}$ est nécessaire ; si elle est omise on ne retrouve plus la loi uniforme attendue sous $H_0(A)$.

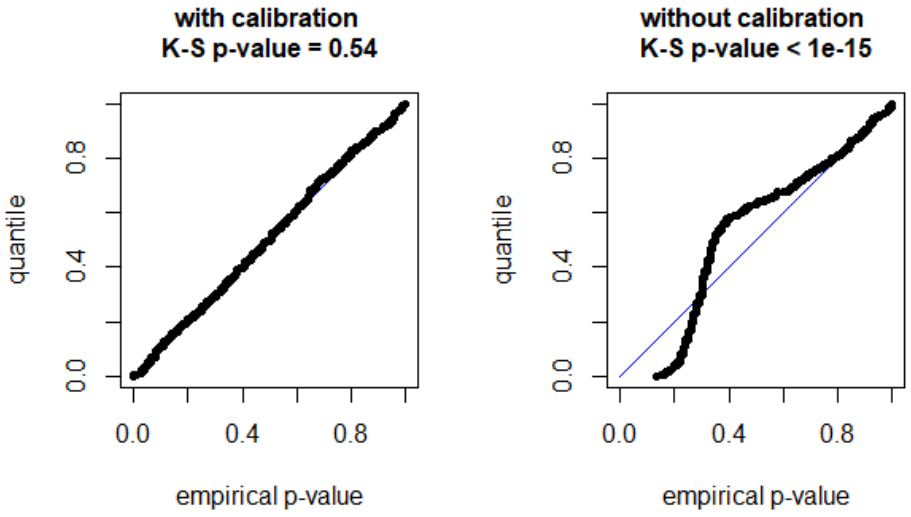


Figure : exemples de graphes quantile-quantile des p-values simulées sous l’hypothèse nulle.

Les courbes ROC obtenues sont satisfaisantes par rapport à celles d'autres méthodes de sélection de variables s'appuyant sur le Lasso : BIC, test de covariance (Lockhart et al 2014), sélection par permutation (Sabourin 2015). Notre méthode apparaît relativement conservative, la probabilité d'obtenir au moins un faux positif étant contrôlée quand les X_j ne sont pas corrélés où quand le signal est assez fort.

Références

Tibshirani, R. (1996). *Regression shrinkage and selection via the lasso*. Journal of the Royal Statistical Society Series B 58 267–288.

Sabourin JA, Valdar W, Nobel AB (2015). *A permutation approach for selecting the penalty parameter in penalized model selection*. Biometrics; doi:10.1111/biom.12359.

Lockhart, R., Taylor, J., Tibshirani, R. and Tibshirani, R. (2014). *A significance test for the lasso (with discussion)*. Ann. Statist. 42 413–468. MR3210970

Spatio-temporal hybrid Geyer point process

Morteza Raeisi [1], Florent Bonneu [1,2] & Edith Gabriel [2]

[1] *LMA EA2151, Avignon University, F-84000 Avignon, France;*
morteza.raeisi@univ-avignon.fr, florent.bonneu@univ-avignon.fr
[2] *INRAE, BioSP, F-84914, Avignon, France; edith.gabriel@inrae.fr*

Résumé. Nous nous intéressons aux semis de points présentant plusieurs structures (agrégation et/ou inhibition) à différentes échelles spatio-temporelles. La méthode d'hydridation permet de construire de tels modèles multi-structures. Nous nous appuyons sur cette approche pour développer le processus de Geyer spatio-temporel. Ce modèle est ensuite utilisé pour modéliser les départs de feux de forêt dans le sud de la France durant plusieurs années.

Mots-clés. Processus ponctuel spatio-temporel, hybridation, processus ponctuel de Gibbs multi-échelles.

Abstract. We focus on spatio-temporal patterns with multi-structure (clustering and/or inhibition) at different spatio-temporal scales. Hybridization is a general way to build processes for modeling such patterns. We investigate the hybridization approach to develop a spatio-temporal Geyer point process. We apply this model to the point pattern of spatio-temporal forest fire occurrences in south of France during several years.

Keywords. Spatio-temporal point process, hybridization, multi-scale Gibbs point process.

1 Introduction on multi-structure point processes

Spatial (and spatio-temporal) point patterns are a collection of events for which locations (and times) of occurrence have been observed in a specified spatial region (and temporal period). Point patterns are often classified into three classes of single interaction structure: randomness, clustering, and inhibition that can be modeled for instance by Poisson process, Cox processes, and Gibbs processes, respectively. These single-structure point process models can be too simplistic to describe some complex phenomena in seismology (Siino et al., 2017; Siino et al., 2018b), epidemiology (Iftimi et al., 2017; Iftimi et al., 2018), and forestry (Gabriel et al., 2017) as they involve several structures at different spatial (and spatio-temporal) scales.

In the last decade, statisticians investigated the methods and models for fitting such patterns (Raeisi et al., 2019b). In the spatial point processes literature, three general approaches are considered for constructing multi-structure point process models: hybridization, thinning and superposition. One important class of multi-structure models is

the family of spatial hybrid Gibbs point process models (Baddeley et al., 2013) that are constructed by hybridization approach. Extension of the hybridization approach to the spatio-temporal framework has recently been considered in Iftimi et al. (2018) and Raeisi et al. (2019a). Here, we introduce the spatio-temporal multi-scale Geyer point process, its simulation algorithm and its estimation methods (Section 2). In Section 3, we apply this model to forest fires in south of France during several years.

2 Spatio-temporal hybrid Geyer point processes

Gibbs point processes offer a large class of models which allow any of the single structure as clustering, randomness, and inhibition. Gibbs point processes are studied by their probability density, defined with respect to a unit rate Poisson point process. In the literature, several spatio-temporal Gibbs point process models have been proposed such as the hardcore (Cronie and van Lieshout, 2015), Strauss (Gonzalez et al., 2016), area interaction (Iftimi et al., 2018), and Geyer (Raeisi et al., 2019a) point processes.

Due to the capability of the Gibbs point processes to cover the prevalent structures, hybridization can be an approach for introducing new Gibbs models which combine several structures at different scales. Given m densities $f_1, f_2, , ...f_m$ of Gibbs point processes, the hybrid density is defined as $f(\mathbf{x}) = cf_1(\mathbf{x}) \times f_2(\mathbf{x}) \times ... \times f_m(\mathbf{x})$ where c is normalization constant and $\mathbf{x}$ is a point pattern (Baddeley et al., 2013).

2.1 Model

The inhomogeneous spatio-temporal hybrid Geyer saturation point process (Raeisi et al., 2019a) has a density defined by

$$f(\mathbf{x}) = c \prod_{(\xi,t)\in\mathbf{x}} \lambda(\xi,t) \prod_{j=1}^{m} \gamma_j^{\min\{s_j, n(C_{r_j}^{q_j}(\xi,t);\mathbf{x})\}}, \tag{1}$$

where $c > 0$ is a normalization constant, λ is a non-negative measurable and bounded function describing the spatio-temporal trend, $r, q > 0$ are the spatial and the temporal radii, s is a saturation parameter, $n(C_r^q(\xi,t);\mathbf{x}) = \sum_{(u,v)\in\mathbf{x}\setminus(\xi,t)} \mathbb{1}\{||u-\xi|| \leq r, |v-t| \leq q\}$ and $n(\mathbf{x})$ is the number of points in point pattern $\mathbf{x} = \{(\xi_1,t_1), ..., (\xi_n,t_n)\}$ in $S \times T \subset \mathbb{R}^2 \times \mathbb{R}$.

The main tool for simulating and fitting Gibbs point processes is the conditional intensity function. For $(u,v) \in S \times T$ it is defined by $\lambda((u,v)|\mathbf{x}) = \frac{f(\mathbf{x}\bigcup(u,v))}{f(\mathbf{x}\setminus(u,v))}$ with $0/0 := 0$ (Cronie and van Lieshout, 2015). The conditional intensity function of hybrid Geyer saturation point process for $(u,v) \in W$ is

$$\lambda((u,v)|\mathbf{x}) = \lambda(u,v) \prod_{j=1}^{m} \gamma_j^{\min\{s_j, n(C_{r_j}^{q_j}(u,v);\mathbf{x})\}} \prod_{(\xi,t)\in\mathbf{x}\setminus(u,v)} \gamma_j^{\min\{s_j, n(C_{r_j}^{q_j}(\xi,t);\mathbf{x}\cup(u,v))\} - \min\{s_j, n(C_{r_j}^{q_j}(\xi,t);\mathbf{x}\setminus(u,v))\}},$$

where its logarithm has a linear form in $\boldsymbol{\theta} = (\log\gamma_1, \log\gamma_2, ..., \log\gamma_m)$, see Raeisi et al. (2019a).

2.2 Simulation algorithm

Gibbs point processes can be simulated by birth-death Metropolis-Hasting algorithm that typically requires only computation of the conditional intensity function. We aim to simulate a spatio-temporal point configuration $\mathbf{x}$ in $S \times T$. The algorithm is as follows.

For $n = 0, 1, ...$, given $X_n = \mathbf{x}$ (e.g. a Poisson process for $n = 0$), generate X_{n+1}:

- Generate two uniform numbers y_1, y_2 in $[0, 1]$
- If $y_1 \leq \frac{1}{2}$ then
  - Generate a new point (u, v) uniformly from a probability density $1/|S \times T|$ where $|S \times T|$ denotes the volume of the spatio-temporal region,
  - Calculate $r((u, v); \mathbf{x}) = \frac{|S\times T|}{n(\mathbf{X})+1}\lambda((u, v)|\mathbf{x}), (u, v) \notin \mathbf{x}$,
    If $y_2 < r((u, v); \mathbf{x})$ then $X_{n+1} = \mathbf{x}\bigcup(u, v)$ else $X_{n+1} = \mathbf{x}$
- If $y_1 > \frac{1}{2}$ then
  - If $\mathbf{x} = \emptyset$ then $X_{n+1} = \mathbf{x}$,
  - Else, select uniformly (ξ, t) from $\mathbf{x}$ according to a discrete probability density $1/n(\mathbf{x})$
    Calculate $r((\xi, t); \mathbf{x}\backslash(\xi, t)) = \frac{|S\times T|}{n(\mathbf{X})}\lambda((\xi, t)|\mathbf{x}), (\xi, t) \in \mathbf{x}$,
    If $y_2 < 1/r((\xi, t); \mathbf{x}\backslash(\xi, t))$ then $X_{n+1} = \mathbf{x}\backslash(\xi, t)$ else $X_{n+1} = \mathbf{x}$.

2.3 Estimation

Gibbs point process models involve two types of parameters: regular and irregular parameters. A parameter is called *regular* if the log likelihood of density is a linear function of that parameter, *irregular* otherwise. In the hybrid Geyer point process (1), λ and γ are regular and r, q, and s are irregular parameters. Irregular parameters can be predetermined by the user. Regular parameters can be estimated using the pseudo-likelihood (Baddeley and Turner, 2000) or logistic likelihood method (Baddeley et al., 2014).

For a Gibbs point process with conditional intensity $\lambda_{\boldsymbol{\theta}}((u, v)|\mathbf{x})$, the log pseudo-likelihood is defined by

$$\log PL(\mathbf{x}; \boldsymbol{\theta}) = \sum_{(\xi,t)\in\mathbf{x}} \log\lambda_{\boldsymbol{\theta}}((\xi, t)|\mathbf{x}) - \int_S\int_T \lambda_{\boldsymbol{\theta}}((u, v)|\mathbf{x})dvdu, \tag{2}$$

and the log logistic likelihood is defined by

$$\log LL(\mathbf{x}, \boldsymbol{d}; \boldsymbol{\theta}) = \sum_{(\xi,t)\in\mathbf{x}} \log \frac{\lambda_{\boldsymbol{\theta}}((\xi,t)|\mathbf{x})}{\lambda_{\boldsymbol{\theta}}((\xi,t)|\mathbf{x}) + \rho(\xi,t)} + \sum_{(u,v)\in\boldsymbol{d}} \log \frac{\rho(u,v)}{\lambda_{\boldsymbol{\theta}}((u,v)|\mathbf{x}) + \rho(u,v)}, \quad (3)$$

where $\boldsymbol{d}$ is a realisation of a Poisson point process with intensity function ρ. By log linearity property for conditional intensity, (2) is a Poisson regression and (3) is a logistic regression (Raeisi et al., 2019a). Hence, estimation can be implemented by using standard software for GLMs. Raeisi et al. (2019a) compare the root of MSE for each parameter estimation obtained from using two methods. They found the advantage of logistic over pseudo-likelihood method for spatio-temporal hybrid Geyer point process, also obtained by Iftimi et al. (2018).

3 Application to forest fire occurences

We consider forest fire occurrences during several years in Southern France. Gabriel et al. (2017) detected space-time interaction structures between fires at different scales. Hence, we consider hybrid Geyer point process (1) to model the forest fire occurrences.

To fit the model, we first need to predetermine irregular parameters. In the literature, there is almost no common method for estimating the irregular parameters in Gibbs point process models. So, we consider ad-hoc values within a reasonable range and their combinations. We then compare the related goodness-of-fit to select the best model.

The next step is to assess the interaction between the forest fires and covariates. In practice, it is particularly useful when the trend can be specified through a regression model on covariates that may vary in space and time (Gonzalez et al., 2016). We consider water coverage, elevation, coniferous cover and building cover as spatial covariates and temperature average, precipitation as spatio-temporal covariates. We express the spatio-temporal inhomogeneity term in Geyer model as $\lambda(x,t) = \beta\mu(x,t)$ where $\log\mu(x,t)$ can be considered as a GLM as follows

$$\log\mu(x,t) = \beta_0 + \sum_{k=1}^{4} \beta_k^S Z_k^S(x) + \sum_{l=1}^{2} \beta_l^{ST} Z_l^{ST}(x,t) + \alpha t. \quad (4)$$

where $Z_k^S(x)$ and $Z_l^{ST}(x,t)$ correspond to the spatial and spatio-temporal covariates.

For model fitting, we use the logistic likelihood (Raeisi et al., 2019a). We fitted the model for a range of values $(r_j, q_j, s_j), j = 1, ..., m$ to choose the optimal combination according to Akaike's Information Criterion (AIC). The goodness-of-fit for a model is accomplished by simulating new point processes from the fitted model. For each simulated realization of points, we compute a spatio-temporal second-order summary statistic. We then compare the distribution of summary statistics for the simulated points and original points by critical envelopes and the local and global test using the Monte Carlo testing procedure (Siino et al., 2018a).

Bibliography

Baddeley, A., Coeurjolly, J.F., Rubak, E. and Waagepetersen, R. (2014). Logistic regression for spatial Gibbs point processes. *Biometrika*, 101(2), pp. 377–392.

Baddeley, A. and Turner, R. (2000). Practical maximum pseudolikelihood for spatial point patterns (with discussion). *Australian and New Zealand Journal of Statistics*, 42, pp. 283–322.

Baddeley, A., Turner, R., Mateu, J. and Bevan, A. (2013). Hybrids of gibbs point process models and their implementation. *Journal of Statistical Software*, 55(11), pp. 1–43.

Cronie, O. and van Lieshout, M.C. (2015). A J-function for inhomogeneous spatio-temporal point processes. *Scandinavian Journal Statistics*, 42(2), pp. 562–579.

Gabriel, E., Opitz, T. and Bonneu, F. (2017). Detecting and modeling multi-scale space-time structures: the case of wildfire occurrences. *Journal of the French Statistical Society*, 158(3), pp. 86–105.

Gonzalez, J.A., Rodriguez-Cortes, F.J., Cronie, O. and Mateu, J. (2016). Spatio-temporal point process statistics: A review. *Spatial Statistics*, 18, pp. 505–544.

Iftimi, A., Montes, P., Mateu, J. and Ayyad, C. (2017). Measuring spatial inhomogeneity at different spatial scales using Hybrids of Gibbs point process models. *Stochastic Environmental Research and Risk Assessment*, 31(6), pp. 1455–1469.

Iftimi, A., van Lieshout, M.C. and Montes, F. (2018). A multi-scale area-interaction model for spatio-temporal point patterns. *Spatial Statistics*, 26, pp. 38–55.

Raeisi, M., Bonneu, F. and Gabriel, E. (2019a). A spatio-temporal multi-scale model for Geyer saturation point process: application to forest fire occurrences. `arXiv:1911.06999`.

Raeisi, M., Bonneu, F. and Gabriel, E. (2019b). On spatial and spatio-temporal multi-structure point process models. *Annales de l'Institut de Statistique de l'Université de Paris*, 63(2-3).

Siino, M., Adelfio, G., Mateu, J., Chiodi, M. and D'Alessandro, A. (2017). Spatial pattern analysis using hybrid models: an application to the Hellenic seismicity. *Stochastic Environmental Research and Risk Assessment*, 31(7), pp. 1633–1648.

Siino, M., Adelfio, G. and Mateu, J. (2018a). Joint second-order parameter estimation for spatio-temporal log-Gaussian Cox processes. *Stochastic Environmental Research and Risk Assessment*, 32, pp. 3525–3539.

Siino, M., D'Alessandro, A., Adelfio, G., Scudero, S. and Chiodi, M. (2018b). Multiscale processes to describe the eastern sicily seismic sequences. *Annals of Geophysics*, 61(2), `DOI:10.441/ag-7711`.

Analyse différentielle de données Hi-C via Classification Ascendante Hiérarchique sous Contrainte de Contiguïté

Nathanaël Randriamihamison [1,2,3] & Marie Chavent [3] & Sylvain Foissac [4] & Nathalie Vialaneix [4] & Pierre Neuvial [3]

[1] *INRAE, UR875 Mathématiques et Informatique Appliquées Toulouse, F-31326 Castanet-Tolosan, France, {nathanael.randriamihamison,nathalie.vialaneix}@inrae.fr*

[2] *Institut de Mathématiques de Toulouse, Univ. Paul Sabatier, UMR 5219, pierre.neuvial@math.univ-toulouse.fr*

[3] *Inria BSO, CQFD Team, CNRS UMR5251 Institut Mathématiques de Bordeaux, marie.chavent@inria.fr*

[4] *GenPhySE, Université de Toulouse, INRAE, ENVT, F-31326, Castanet-Tolosan, France, sylvain.foissac@inrae.fr*

Résumé.

Les données Hi-C mesurent la proximité spatiale entre paires de positions génomiques et donnent des informations sur l'organisation 3D de l'ADN qui, elle-même, a un rôle important dans la régulation de l'expression des gènes. Le but de l'analyse différentielle de données Hi-C est de trouver des différences significatives entre la structure 3D du génome de deux conditions biologiques différentes à partir de plusieurs réplicats d'expériences Hi-C dans chaque condition. Ici, nous proposons une nouvelle méthode d'analyse différentielle basée sur la Classification Ascendante Hiérarchique avec Contrainte de Contiguïté (CAHCC). Celle-ci est utilisée pour représenter la structure hiérarchique des positions génomiques sous la forme d'un arbre binaire et le problème de l'analyse différentielle Hi-C est alors transformé en un problème de comparaison d'arbres, résolu en utilisant des distances entre arbres.

Mots-clés. Classification ascendante hiérarchique, classification ascendante hiérarchique sous contrainte, dendrogramme, données Hi-C, analyse différentielle, distances entre arbres.

Abstract.

The spatial proximity between pairs of genomic positions can be measured by Hi-C experiments, which give insights into the 3D organization of DNA. This organization plays an important role in the regulation of gene expression. The aim of Hi-C differential analysis is to find significant differences in the 3D structure of the genome between two biological conditions from replicates of Hi-C experiments in each condition. Here, we present a new differential analysis method based on Hierarchical Agglomerative Clustering with Contiguity Constraint (CCHAC). CCHAC is used to represent the hierarchical structure of genomic positions in the form of a binary tree. The problem of Hi-C differential analysis is then translated into a tree comparison problem and handled using tree distances.

Keywords. Hierarchical agglomerative clustering, constrained hierarchical agglomerative clustering, dendrogram, Hi-C data, differential analysis, tree distances.

1 Introduction

Données Hi-C et analyse différentielle

Les données Hi-C, issues du séquençage haut-débit de nouvelle génération, sont des données génomiques qui renseignent sur l'organisation spatiale des chromosomes dans le noyau des cellules. Elles nous permettent d'avoir accès à une mesure de la proximité spatiale entre paires de positions à travers l'ensemble du génome et elles ont permis de mettre en évidence l'existence de régions du génome très compactées [Dixon et al., 2012]. Cette organisation spatiale et ses variations ont des répercussions dans la régulation de l'expression des gènes, jouant notamment un rôle déterminant dans le développement de certaines maladies ou malformations [Lupiáñez et al., 2015].

En pratique, les données Hi-C se présentent sous la forme de matrices, $H = (h_{ij})_{i,j}$ (aussi appelées *cartes Hi-C*) dont l'entrée h_{ij} correspond au comptage (obtenu par le séquençage haut débit) du nombre de fois où les intervalles génomiques i et j ont été observés en contact. L'intensité des coefficients décroît avec l'éloignement à la diagonale, ce qui est dû à l'organisation linéaire du génôme dans les molécules d'ADN. Les matrices sont donc des matrices de similarité, carrées, symétriques, à entrées entières et positives. Dans la suite, on notera $(\mathbf{H}^t)_{1\leq t\leq T}$, T matrices Hi-C, de dimension $p \times p$ (p est le nombre d'intervalles génomiques considérés), obtenues dans deux conditions biologiques différentes, $\mathcal{C}_1$ et $\mathcal{C}_2$, telles que $\mathcal{C}_1 \cup \mathcal{C}_2 = \{1, \ldots, T\}$ et $\mathcal{C}_1 \cap \mathcal{C}_2 = \emptyset$, et on notera donc h_{ij}^t le nombre de contacts entre les positions génomiques i et j pour la matrice $\mathbf{H}^t$ [1]. L'objet de ce travail est l'analyse différentielle de données Hi-C, c'est-à-dire, la recherche de différences, avec une garantie statistique, entre les matrices de la condition $\mathcal{C}_1$ et celles de la condition $\mathcal{C}_2$.

État de l'art et limites des méthodes existantes

La plupart des méthodes d'analyse différentielle de données Hi-C sont basées sur des comparaisons indépendantes pour chaque paire d'intervalles génomiques (i, j). Pour (i, j) donnés, elles calculent une statistique de test puis une p-valeur rendant compte de la différence entre les valeurs des $(h_{ij}^t)_t$ entre les deux conditions, ces p-valeurs étant ensuite corrigées pour contrôler le FDR (False Discovery Rate). Parmi ces méthodes, on peut citer celle de [Lun and Smyth, 2015], reposant sur une modélisation des coefficients par une distribution binomiale négative, celle de [Stansfield et al., 2018], basée sur l'utilisation d'un Z-score, ou encore celle de [Djekidel et al., 2018] utilisant un processus spatial de Poisson.

Les limites de telles approches résident dans le fait qu'elles négligent des propriétés importantes des données comme leur structure hiérarchique ou encore la dépendance entre les coefficients. Cela peut engendrer des difficultés pour interpréter les résultats en termes de différences de structure de la chromatine.

1. Pour simplifier le propos, dans ce qui suit, les matrices $(\mathbf{H}^t)_t$ sont considérées ne correspondre qu'à un seul même chromosome.

2 Méthode de comparaison structurelle de données Hi-C

L'approche que nous proposons repose sur une modélisation de la structure hiérarchique de l'organisation spatiale du génome par une méthode de classification ascendante hiérarchique (CAH) avec contrainte de contiguïté entre les classes fusionnées. Les dendrogrammes issus de cette CAH sont alors directement comparés, ce qui permet d'obenir des différences dans l'organisation structurelle et non plus simplement des différences ponctuelles. De manière plus précise, la méthode se déroule en 4 étapes principales :

1. **Étape 1 : Modélisation de la structure hiérarchique sous forme de dendrogramme.** Une adaptation de la Classification Ascendante Hiérarchique avec lien de Ward est utilisée pour obtenir un dendrogramme pour chaque matrice $\mathbf{H}^t$. Cette version permet d'utiliser les entrées log-transformées de la matrice Hi-C comme des similarités sur lesquelles le critère de Ward est calculé et respecte, de plus, l'ordre des positions génomiques grâce à l'imposition d'une contrainte de contiguïté. Les propriétés pratiques et les justifications théoriques de cette méthode, appelée Classification Ascendante Hiérarchique sous Contrainte de Contiguité (CAHCC), ont été discutées dans [Ambroise et al., 2019, Randriamihamison et al., 2019] et elle est implémentée dans le package R **adjclust** (disponible sur le CRAN). À la fin de cette étape, un dendrogramme, $\mathbf{D}^t$, est associé à chaque matrice HiC, $\mathbf{H}^t$, comme illustré dans la Figure 1.

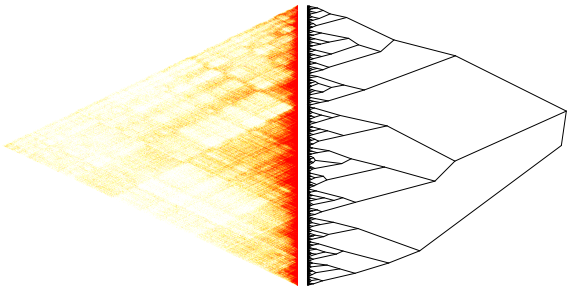


FIGURE 1 – Matrice Hi-C (gauche) et dendrogramme associé (droite).

2. **Étape 2 : Extraction de sous-arbres pour la comparaison.**
Pour déterminer des régions du génome présentant des différences significatives entre conditions, des sous-arbres, basés sur un ensemble commun de feuilles pour tous les dendrogrammes $\mathbf{D}^t$, sont extraits. De manière plus précise, on détermine L sous-ensembles d'intervalles génomiques contigus, $(I_l)_{1\leq l\leq L}$ (par exemple, l'ensemble des intervalles génomiques contigus de taille fixée) et on définit $\mathbf{D}_l^t$ le plus petit sous-arbre induit par $\mathbf{D}^t$ recouvrant I_l dont toutes les branches correspondantes à des feuilles extérieures à I_l sont élaguées (voir figure 2 pour un exemple).

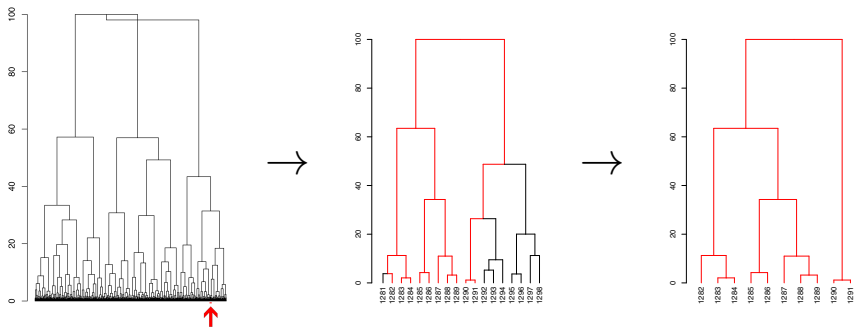


FIGURE 2 – Gauche : Dendrogramme $\mathbf{D}^t$. En rouge est indiquée la zone correspondant à l'intervalle I_l considéré. Centre : Plus petit sous-arbre induit par $\mathbf{D}^t$ recouvrant I_l. En rouge sont indiquées les branches correspondant à des feuilles dans I_l. Droite : Sous-arbre D_l^t obtenu après élagage des feuilles extérieures à I_l.

Des biais techniques variés, liés à l'acquisition des données, induisent des différences dans les hauteurs des dendrogrammes et des sous-arbres associés, qui sont préjudiciables à l'analyse. Ces biais techniques sont corrigés par une phase de normalisation qui consiste à aligner les hauteurs maximales et minimales de l'ensemble des arbres $(D_l^t)_t$ sur des valeurs communes, et uniformes pour tous l.

3. **Étape 3 : Comparaison d'arbres.**
   L'ensemble des paires d'arbres, D_l^t et $D_l^{t'}$ pour un intervalle donné sont alors comparées par le calcul d'une distance entre arbres appelée **weighted Path Difference metric (wPD)** et disponible dans le package R **phangorn**[2]. Cette distance est basée sur l'ensemble des longueurs des plus courts chemins entre deux feuilles le long du dendrogramme (distances cophénétiques pondérée, $\mathbf{d}^t \in \mathbb{R}^Q$ où $Q = n_l(n_l-1)/2$ avec n_l le cardinal de I_l) et correspond à la distance euclidienne des vecteurs de ces longueurs entre les deux dendrogrammes : $\mathbf{wPD}(\mathbf{D}_l^t, \mathbf{D}_l^{t'}) = \|\mathbf{d}_l^t - \mathbf{d}_l^{t'}\|$.
4. **Étape 4 : Obtention de garanties statistiques.**
   Par analogie avec les approches ANOVA, nous proposons de comparer les sous-arbres entre deux conditions par utilisation d'une analogie entre la distance **wPD** et la distance euclidenne et en calculant une statistique de test basée sur le ratio entre inertie inter-conditions et inertie intra-conditions :

$$F_l = \frac{(\mathcal{I}_{inter})/1}{(\mathcal{I}_1 + \mathcal{I}_2)/4}$$

où

2. Pour des questions de place, nous ne présentons pas les distances alternatives qui ont été évaluées.

— $\mathcal{I}_k$ correspond à l'inertie de la condition $\mathcal{C}_k$ calculée à partir des distances **wPD** entre sous-arbres de cette condition ;
— $\mathcal{I}_{\text{inter}}$ désigne l'inertie inter-conditions calculée à partir des distances **wPD** entre sous-arbres de deux conditions différentes.

Pour déterminer les intervalles génomiques significativement différents d'un point de vue structurel, on considère une approximation de la distribution de F_l sous l'hypothèse nulle, obtenue par mélange des conditions biologiques entre échantillons. Des travaux sont en cours pour déterminer une distribution théorique.

3 Exemples de résultats

L'approche proposée a été testée sur un ensemble de 6 matrices Hi-C de 2 conditions différentes (3 par condition) correspondant au chromosome 18 du génome humain pour deux types de cellules différentes. Ces données proviennent de [Sanborn et al., 2015] pour la condition 1 et [Darrow et al., 2016] pour la condition 2 (numéros d'accession GEO GSE63525 et GSE71831). Des exemples de sorties de la méthode sont données dans la figure 3 et illustrent la pertinence de l'approche sur données réelles.

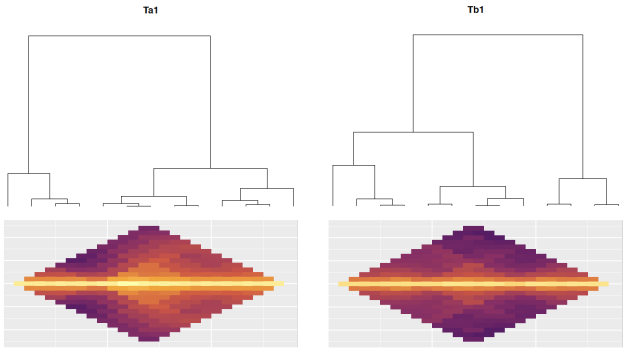


FIGURE 3 – Haut : Exemple de deux sous-arbres identifés associés, respectivement à la condition 1 (gauche) et à la condition 2 (droite). Bas : Sous-matrices Hi-C correspondantes (les niveaux de couleurs correspondent à l'intensité du contact, h_{ij}^t.

4 Conclusion

La méthode d'analyse différentielle présentée dans cette communication est basée sur l'idée de représenter la structure hiérarchique inhérente aux données Hi-C à l'aide d'un arbre (graphe binaire) construit à l'aide de la classification ascendante hiérarchique sous contrainte de contiguïté. On utilise ensuite une distance entre arbre et une statistique de

ratio des inerties inter et intra-conditions pour déterminer des intervalles génomiques où les structures sont significativement différentes.

Cette méthode permet de répondre aux limites des méthodes classiques d'analyse différentielle de données Hi-C qui ne prennent pas en compte la structure hiérarchique des données et les dépendances entre coefficients.

Remerciements

Ce travail a été effectué dans le cadre du projet SCALES, financé par la Mission pour les Initiatives Transverses et Interdisciplinaires du CNRS. La thèse de N.R. est financée par le programme doctoral INRAE/Inria.

Références

[Ambroise et al., 2019] Ambroise, C., Dehman, A., Neuvial, P., Rigaill, G., and Vialaneix, N. (2019). Adjacency-constrained hierarchical clustering of a band similarity matrix with application to genomics. *Algorithms for Molecular Biology*, 14 :22.

[Darrow et al., 2016] Darrow, E. M., Huntley, M. H., Dudchenko, O., Stamenova, E. K., Durand, N. C., Sun, Z., Huang, S.-C., Sanborn, A. L., Machol, I., Shamim, M., Seberg, A. P., Lander, E. S., Chadwick, B. P., and Aiden, E. L. (2016). Deletion of DXZ4 on the human inactive X chromosome alters higher-order genome architecture. *Proceedings of the National Academy of Sciences*, 113(31) :E4504–E4512.

[Dixon et al., 2012] Dixon, J., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485 :376–380.

[Djekidel et al., 2018] Djekidel, M. N., Chen, Y., and Zhang, M. Q. (2018). FIND : difFerential chromatin INteractions Detection using a spatial Poisson process. *Genome Research*, 28(3) :412–422.

[Lun and Smyth, 2015] Lun, A. T. and Smyth, G. K. (2015). diffHic : a Bioconductor package to detect differential genomic interactions in Hi-C data. *BMC Bioinformatics*, 16 :258.

[Lupiáñez et al., 2015] Lupiáñez, D. G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J. M., Laxova, R., Santos-Simarro, F., Gilbert-Dussardier, B., Wittler, L., Borschiwer, M., Haas, S. A., Osterwalder, M., Franke, M., Timmermann, B., Hecht, J., Spielmann, M., Visel, A., and Mundlos, S. (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, 161(5) :1012–1025.

[Randriamihamison et al., 2019] Randriamihamison, N., Vialaneix, N., and Neuvial, P. (2019). Applicability and interpretability of hierarchical agglomerative clustering with or without contiguity constraints. Submitted for publication. Preprint arXiv 1909.10923.

[Sanborn et al., 2015] Sanborn, A. L., Rao, S. S. P., Huang, S.-C., Durand, N. C., Huntley, M. H., Jewett, A. I., Bochkov, I. D., Chinnappan, D., Cutkosky, A., Li, J., Geeting, K. P., Gnirke, A., Melnikov, A., McKenna, D., Stamenova, E. K., Lander, E. S., and Aiden, E. L. (2015). Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proceedings of the National Academy of Sciences*, 112(47) :E6456–E6465.

[Stansfield et al., 2018] Stansfield, J. C., Cresswell, K. G., Vladimirov, V. I., and Dozmorov, M. G. (2018). HiCcompare : an R-package for joint normalization and comparison of Hi-C datasets. *BMC Bioinformatics*, 19 :279.

Inférence de graphe avec contrôle du taux de faux positifs

Tabea Rebafka [1], Etienne Roquain [1]& Fanny Villers [1]

[1] *Sorbonne Université, Université de Paris, CNRS, Laboratoire de Probabilités, Statistique et Modélisation, Paris, France.*
tabea.rebafka@upmc.fr, etienne.roquain@upmc.fr, fanny.villers@upmc.fr

Résumé. Une nouvelle procédure de test multiple est proposée pour inférer un graphe à partir d'une observation bruitée de ce graphe. Dans ce but, nous introduisons d'abord un modèle à blocs stochastiques bruité (NSBM) et développons un algorithme EM variationnel pour estimer les paramètres et calculer un clustering des nœuds. Nous définissons ensuite une nouvelle procédure d'inférence du graphe qui exploite la topologie d'un NSBM. Un résultat théorique montre que notre procédure est asymptotiquement proche de la procédure oracle, qui contrôle le taux de faux positifs tout en maximisant le taux de vrais positifs. Des résultats numériques illustrent les propriétés de notre procédure et montrent qu'elle est plus performante que des procédures de test classiques. Cette note est une version résumée du manuscrit Rebafka et al. (2019).

Mots-clés. Inférence de graphe, modèle à blocs stochastiques, taux de faux positifs, algorithme EM variationnel.

Abstract. A new multiple testing procedure is proposed for the problem of inferring a graph from a noisy observation of that graph. To this end, the so-called noisy stochastic block model (NSBM) is introduced. Parameter estimates and a node clustering is then provided via a variational expectation-maximization algorithm. A new graph inference method is proposed by using a multiple-testing procedure that exploits the graph topology of the NSBM. It can be shown that our procedure asymptotically mimics the oracle procedure that controls the false discovery rate while maximizing the true discovery rate. Numerical experiments illustrate the performance of our test procedure and show that it outperforms classical methods. This note is a summary of the manuscript Rebafka et al. (2019).

Keywords. Graph inference, stochastic block model, false discovery rate, variational expectation-maximization algorithm.

1 Introduction

In many applications, the network of interest is not observed, but only a perturbation of it. An essential task is then to infer a reliable version of the network by removing the edges that are only due to noise.

Graph inference is a long-standing research topic in statistics, especially in the context of estimating marginal or partial correlations between nodes. In this context, a Gaussian graphical model (Lauritzen, 1996) is often used for the correlation or the precision matrix. Inferring the graph is then classically done by some graphical lasso-type approaches (Meinshausen and Bühlmann, 2006; Friedman et al., 2007; Banerjee et al., 2008; Ravikumar et al., 2011).

Once the graph is inferred, it is common that a deeper analysis is carried out to describe the communication structure of the network, for instance by community detection, that is, clustering of the nodes in groups with similar connecting behavior. A popular random graph model for node clustering is the stochastic block model (SBM) (Holland et al., 1983) that models network heterogeneity by varying connecting behavior of different groups of nodes. More precisely, each node is supposed to belong to a group and the edge probability of a pair of nodes depends entirely on the group membership of these two nodes. Thus, clustering becomes the problem of estimating the group memberships in the SBM.

In this work, we propose a procedure for both graph inference and node clustering. To this end, we introduce a variant of the SBM called the noisy stochastic block model (NSBM). In this model, the graph of interest is an SBM, but it is not observed and considered to be a latent structure. The observation is a noisy version of this graph obtained by the following blurring mechanism: in place of missing edges, pure random noise is observed, and in place of present edges, we observe an effect, whose intensity may depend on the block memberships of the nodes in the latent graph. Based on the NSBM, we adapt procedures of the multiple testing literature for mixture models, namely a q-value approach (Storey, 2003; Castillo and Roquain, 2018). Thus, our approach leads to a new procedure for simultaneously inferring a clustering of the nodes and the graph itself, with a clear interpretation in terms of false positives: among the edges discovered by the procedure, there are, on average, at most 5% (say) of errors.

2 Gaussian noisy stochastic block model

Consider an undirected graph with n nodes. Denote $\mathcal{A} = \{(i,j) : 1 \leq i < j \leq n\}$ the set of all possible edges. We observe a symmetric, real-valued matrix $X = (X_{i,j})_{1 \leq i,j \leq n} \in \mathbb{R}^{n^2}$ representing the observed interactions between all node pairs (i,j).

The distribution of X is modeled by a *noisy stochastic block model* (NSBM) defined as a perturbation of a standard binary stochastic block model (SBM) . More precisely, we suppose that X is a noisy version of a binary adjacency matrix $A = (A_{i,j})_{1 \leq i,j \leq n} \in \{0,1\}^{n^2}$, where $A_{i,j} = 1$ if and only if there is an edge between node i and node j. We assume that A is a binary SBM and our aim is to derive A from the observation X.

Let Q be the number of latent node blocks. The NSBM is defined by the following random layers. First, generate a vector $Z = (Z_1, \ldots, Z_n)$ of block memberships of the nodes such that Z_i, $1 \leq i \leq n$ are i.i.d. taking their values in $\{1, 2, \ldots, Q\}$ with proba-

bilities $\pi_q = \mathbb{P}(Z_1 = q)$ for $q \in \{1, \dots, Q\}$ with parameter $\pi = (\pi_q)_{q \in \{1,\dots,Q\}} \in [0,1]^Q$ such that $\sum_{q=1}^{Q} \pi_q = 1$. Then, conditionally on Z, the variables $A_{i,j}$, $(i,j) \in \mathcal{A}$ are independent and each $A_{i,j}$ follows a Bernoulli variables with parameter w_{Z_i,Z_j}, that is,

$$A_{i,j} \mid Z \sim \mathcal{B}(w_{Z_i,Z_j}),$$

for some symmetric parameter $w = (w_{q,\ell})_{q,\ell \in \{1,\dots,Q\}} \in [0,1]^{Q^2}$, that is, $w_{q,\ell} = w_{\ell,q}$ for all $q, \ell \in \{1, \dots, Q\}$, as the graph here is assumed to be undirected. Note that only $A_{i,j}$, $(i,j) \in \mathcal{A}$ are sampled randomly and we set $A_{j,i} = A_{i,j}$ for all $(i,j) \in \mathcal{A}$ and $A_{i,i} = 0$ for $i \in \{1, \dots, n\}$. Finally, conditionally on (Z, A), the observed variables $X_{i,j}$, $(i,j) \in \mathcal{A}$ are independent Gaussian variables and every $X_{i,j}$ has the marginal distribution given by

$$X_{i,j} \mid (Z, A) \sim (1 - A_{i,j})\mathcal{N}(0, \sigma_0^2) + A_{i,j}\mathcal{N}(\mu_{Z_i,Z_j}, \sigma^2_{Z_i,Z_j}),$$

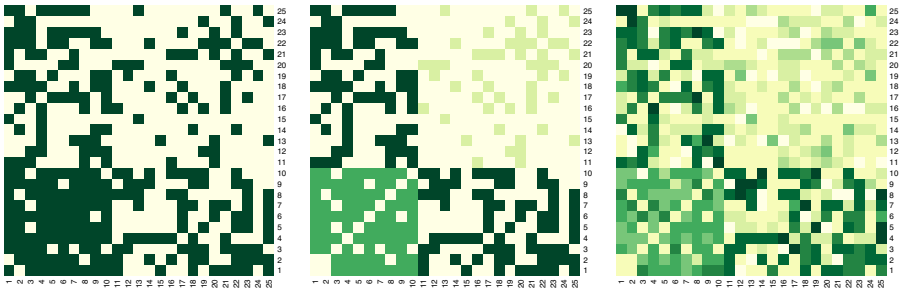
with parameters $\sigma_0^2 > 0, \mu_{q,\ell} \in \mathbb{R}$ and $\sigma^2_{q,\ell} > 0$. The unknown global model parameter is $\theta = (\pi, w, \sigma_0^2, \mu, \sigma)$ with $\mu = (\mu_{q,\ell})_{q \le \ell}$ and $\sigma^2 = (\sigma^2_{q,\ell})_{q \le \ell}$. The observation is X, while both Z and A are unobserved, that is, (Z, A) are the latent variables of this model.

The rationale behind this model is that, in place of missing edges ($A_{i,j} = 0$) pure random noise modeled by the null distribution $\mathcal{N}(0, \sigma_0^2)$ is observed, and instead of present edges ($A_{i,j} = 1$), we observe an effect, whose distribution $\mathcal{N}(\mu_{Z_i,Z_j}, \sigma^2_{Z_i,Z_j})$ depends on the block membership of the interacting nodes in the underlying SBM. While we focus here on Gaussian distributions for brevity, this model can be extended to any parametric model.

Figure 1 provides an illustration with two latent blocks: one is a community, the other has few connections among nodes in the same block and more connexions with nodes of the other block, see (a). In (b), the associated Gaussian means are depicted: 0 if there is no edge, i.e. $A_{i,j} = 0$, and some non zero mean $\mu_{q,\ell}$ otherwise, depending on the group memberships of the interacting nodes. Finally, (c) displays the observations $X_{i,j}$ that are random perturbations of the means in (b). To to recover A from X, the fundamental idea of our method is that learning the block memberships helps in the decision about the presence or absence of edges. For instance, from (c) we see that when $X_{i,j}$ is associated with two nodes in the first group there is few ambiguity about the underlying edges. However, for values $X_{i,j}$ associated with two nodes belonging to the second block this is a more difficult decision. Indeed, the good interpretation of a intermediate, light green value of $X_{i,j}$ depends on the group memberships of the interacting nodes.

Our results on parameter estimation

It can be shown that the NSBM is **identifiable** up to label swapping under classical assumptions. Furthermore, the maximum likelihood estimator of θ can be approached by a **variational EM-algorithm**. The development of this algorithm is involved, but similar to other SBM-type models. As a byproduct, the variational EM-algorithm also provides a clustering $\hat{Z}$ of the nodes into Q classes. See Rebafka et al. (2019) for details.



(a) Latent graph A. (b) Gaussian means $(\mu_{Z_i,Z_j})_{i,j}$. (c) Observations X.

Figure 1: Gaussian NSBM with two groups. (a) Latent adjacency matrix A (b) Matrix of Gaussian means according to presence/absence of edge $A_{i,j}$ and group membership (c) observed random perturbation of the matrix in (b).

3 Test procedure

We now turn to the main objective of this note: graph inference. More precisely, the aim is to recover the adjacency matrix A from the observation X. A natural approach is to decide, for each node pair (i, j), whether $A_{i,j}$ is zero or not by using a testing procedure. This corresponds to simultaneously test $H_{0,i,j} : A_{i,j} = 0$, that is, no edge between i and j in the latent graph, against $H_{1,i,j} : A_{i,j} = 1$, that is, there is an edge between i and j in the latent graph, for all pairs (i, j).

Formally, a multiple testing procedure is any measurable function $\varphi(X) \in \{0, 1\}^{\mathcal{A}}$, with the convention that $\varphi_{i,j}(X) = 1$ if and only if $H_{0,(i,j)}$ is rejected. The false discovery rate (FDR) of a given multiple testing procedure $\varphi(X)$ is the average proportion of errors among the discovered edges. The power of test or true discovery rate (TDR) is the average proportion of discovered edges in the underlying latent graph. A good testing procedure detects a maximum number of significant edges, without making too many false detections. In this sense, for a given level $\alpha \in (0, 1)$, we aim at finding a testing procedure $\varphi = \varphi_\alpha$ such that for all θ,

$$\mathrm{FDR}(\theta, \varphi) \leq \alpha, \quad \text{with } \mathrm{TDR}(\theta, \varphi) \text{ as large as possible.}$$

That is, the FDR is controlled at level α, while many true edges are discovered.

3.1 Oracle procedure

To infer the latent graph A in the NSBM the best classification rule is the Bayes rule, which is based on the posterior distribution of A, that is, the distribution of A given X. However, this distribution is intractable like in numerous other latent variable models.

So, to start with, we assume that the latent clustering Z is known. Then it is natural to consider as test statistics the posterior probabilities of $A_{i,j}$ given X and Z, that is

$$\ell_{i,j}(X, Z, \theta) = \mathbb{P}_\theta(A_{i,j} = 0 \mid X, Z) = \frac{(1 - w_{Z_i,Z_j}) f_{\mathcal{N}(0,\sigma_0^2)}(X_{i,j})}{(1 - w_{Z_i,Z_j}) f_{\mathcal{N}(0,\sigma_0^2)}(X_{i,j}) + w_{Z_i,Z_j} f_{\mathcal{N}(\mu_{Z_i,Z_j},\sigma^2_{Z_i,Z_j})}(X_{i,j})}.$$

We refer to the $\ell_{i,j}(X, Z; \theta)$ as the *ℓ-values*, which corresponds to the well known local FDR in multiple testing literature Efron et al. (2001)). A convenient multiple testing procedure rejects $H_{0,i,j}$ provided that $\ell_{i,j}(X, Z; \theta) \leq t$ for some threshold t. This threshold t should be chosen such that the FDR is lower than or equal to α. Since this FDR value is concentrated around

$$Q_\theta(t) = \frac{\mathbb{E}_\theta\left[\sum_{(i,j)\in\mathcal{A}} (1 - A_{i,j}) \mathbb{1}\{\ell_{i,j}(X, Z, \theta) \leq t\}\right]}{\mathbb{E}_\theta\left[\sum_{(i,j)\in\mathcal{A}} \mathbb{1}\{\ell_{i,j}(X, Z, \theta) \leq t\}\right]},$$

it is natural to reject $H_{0,i,j}$ whenever $q_{i,j}(X, Z; \theta) = Q_\theta(\ell_{i,j}(X, Z, \theta)) \leq \alpha$. The quantities $q_{i,j}(X, z; \theta)$ are referred to as the *q-values*, a term that goes back to Storey (2003). Thus, we define the *oracle multiple testing procedure* as

$$\varphi^*_{i,j} = \mathbb{1}\{q_{i,j}(X, Z; \theta) \leq \alpha\}, \ \ (i, j) \in \mathcal{A},$$

which depends on two unknown quantities of the NSBM: the global parameter θ and the latent clustering Z.

Our results on oracle graph inference

We prove that the oracle procedure maximizes the TDR among all procedures controlling the marginal FDR at level α (under appropriate assumptions), see Rebafka et al. (2019) for details.

3.2 New test procedure

Obviously, the oracle procedure is unknown, but it can be approximated using the estimator $\widehat{\theta} = (\widehat{\pi}, \widehat{w}, \widehat{\sigma}_0, \hat{\mu}, \hat{\sigma}^2)$ of θ and the estimated clustering $\widehat{Z}$ obtained by the VEM algorithm for the NSBM. Thus, our final graph inference procedure is defined by

$$\varphi^{\text{VEM}}_{i,j} = \mathbb{1}\{q_{i,j}(X, \widehat{Z}; \widehat{\theta}) \leq \alpha\}, \quad (i, j) \in \mathcal{A}.$$

Our results on graph inference

We show that the procedure φ^{VEM} is asymptotically mimicking the oracle φ^* both in terms of FDR and TDR, under appropriate assumptions. The proof relies on the regularity of the

Gaussian model, combined with concentration inequalities. In addition, these theoretical findings are supported via simulation results. While our procedure maintains the FDR close to the nominal value α, it achieves the highest TDR among various other test procedures, in a large variety of scenarios, see Rebafka et al. (2019) for details.

Acknowledgments

This work has been supported by the grants ANR-16-CE40-0019 (SansSouci), ANR-17-CE40-0001 (BASICS), ANR-18-CE02-0010-01(EcoNet) and by the GDR ISIS through the "projets exploratoires" program (project TASTY).

References

Banerjee, O., El Ghaoui, L., and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.*, 9:485–516.

Castillo, I. and Roquain, E. (2018). On spike and slab empirical Bayes multiple testing. *arXiv e-prints*, page arXiv:1808.09748.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.*, 96(456):1151–1160.

Friedman, J., Hastie, T., and Tibshirani, R. (2007). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Holland, P., Laskey, K., and Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137.

Lauritzen, S. L. (1996). *Graphical models*, volume 17 of *Oxford Statistical Science Series*. The Clarendon Press, Oxford University Press, New York. Oxford Science Publications.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34(3):1436–1462.

Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Stat.*, 5:935–980.

Rebafka, T., Roquain, E., and Villers, F. (2019). Graph inference with clustering and false discovery rate control.

Storey, J. D. (2003). The positive false discovery rate: a Bayesian interpretation and the q-value. *Ann. Statist.*, 31(6):2013–2035.

Estimateurs du Maximum de Vraisemblance explicites pour le modèle linéaire généralisé dans le cas de covariables catégorielles

Tom Rohmer [1], Alexandre Brouste [2] & Christophe Dutang [3]

[1] *GenPhySE, Université de Toulouse, INRAE, ENVT, F-31326*
24, Chemin de Borde Rouge, 31326 Castanet Tolosan, France
[2] *Institut du Risque de l'Assurance, Laboratoire Manceau de Mathématiques*
Le Mans Université, Avenue Olivier Messiaen, 72085 LE MANS, France
[3] *CEREMADE, CNRS, Univ. Paris-Dauphine*
Place du Maréchal de Lattre de Tassigny, 75016 PARIS, France

Résumé.

Dans cette contribution, en reprenant l'approche des modèles linéaires généralisés, nous étudions la modélisation de variable d'intérêt, conditionnellement à des covariables qualitatives. Nous déterminons une forme explicite pour l'estimateur du maximum de vraisemblance (MLE) lorsque les covariables sont catégorielles : dans le cas du modèle avec une covariable catégorielle et dans le cas du modèle avec deux (ou plus) covariables catégorielles avec interactions. Ces estimateurs sont largement plus rapide à obtenir que les algorithmes Iterative Weighted Least Square (IWLS). Dans le cas du modèle sans interaction, le MLE n'a en général pas de forme explicite cependant nous proposons un estimateur de type Least Square possédant les mêmes propriétés asymptotiques que le MLE.

Mots-clés. Modèle linéaire généralisé, MLE explicites

Abstract.

In this paper, we study the modelling of variables of interest, conditional on qualitative covariates, using the generalized linear model approach. On the one hand, we determine a general explicit form for the Maximum Likelihood Estimator (MLE) when the covariates are categorical ; in the case of the model with one categorical explanatory variable and in the case of more of one categorical explanatory variable and interaction terms. These estimators are far less computer intensive than the Iterative Weighted Least square (IWLS) algorithm. In the case of the model without interaction terms, the MLE does not have explicit solution. We propose an alternative explicit Least Square type estimator with same asymptotic properties as MLE.

Keywords. Generalized linear model, explicit MLE

1 Introduction

Les modèles linaires généralisés (GLM) sont introduits dans Nelder et Wedderburn (1972) et ont été largement popularisés notamment au travers du livre de McCullagh et Nelder (1989). Les modèles de régressions s'appuient sur des échantillons $Y_1, \dots Y_n$ non identiquement distribués, en considérant par exemple des vecteurs de covariables explicatives $(X_1^{(1)}, \dots, X_1^{(p)})$, $\dots$, $(X_n^{(1)}, \dots, X_n^{(p)})$, $p \geq 1$ Dans le cas des GLM, pour i variant de 1 à n, les Y_i sont supposés appartenir à la famille exponentielle de paramètre $\lambda_i \in \Lambda$, où $\lambda_i = \lambda(\boldsymbol{\vartheta}, \boldsymbol{X}_i)$ est une fonctionnelle du vecteur de covariable $(X_i^{(1)}, \dots, X_i^{(p)})$ et d'un vecteur de paramètre $\boldsymbol{\vartheta} \in \Theta \subset \mathbb{R}^p$ inconnu. Plus précisément, la log vraisemblance associée à l'expérience statistique générée par Y_i, $i = 1, \dots, n$ s'écrit :

$$\log L(\boldsymbol{\vartheta}|y_i) = \frac{\lambda_i y_i - b(\lambda_i)}{a(\phi)} + c(y_i, \phi), \quad y_i \in \mathbb{Y} \subset \mathbb{R}, \tag{1}$$

et $-\infty$ si $y_i \notin \mathbb{Y}$ où $a : \mathbb{R} \to \mathbb{R}$, $b : \Lambda \to \mathbb{R}$ et $c : \mathbb{Y} \times \mathbb{R} \to \mathbb{R}$ sont des fonctions mesurables, connues et ϕ un paramètre de dispersion.

Le GLM est alors défini en supposant un lien, caractérisé par la fonction bijective continue et dérivable $g : b'(\Lambda) \to \mathbb{R}$, entre l'espérance $\mathbb{E}_\vartheta Y_i$ et le prédicteur linéaire $\eta_i = \langle x_i, \vartheta \rangle$:

$$g(\mathbb{E}_\vartheta Y_i) = \langle \boldsymbol{x}_i, \boldsymbol{\vartheta} \rangle = \eta_i, \quad \forall \boldsymbol{\vartheta} \in \Theta,$$

avec dans le cas de la famille exponentielle $\mathbb{E}_\vartheta Y_i = b'(\lambda_i)$, pour b définie dans (1) ; en d'autre mots, en notant $\ell = (b')^{-1} \circ g^{-1}$, le modèle satisfait $\lambda_i = \ell(\eta_i)$. Le vecteur des paramètres ϑ est généralement estimé de telle façon à maximiser la log- vraisemblance de l'expérience statistique engendrée par $(Y_1, \dots Y_n)$, c'est-à-dire de telle sorte à résoudre les équations du score $S_j(\boldsymbol{\vartheta}) = 0$, $j = 1, \dots, p$ où

$$S_j(\boldsymbol{\vartheta}) = \frac{1}{a(\phi)} \sum_{i=1}^{n} x_i^{(j)} \ell'(\eta_i) \left(y_i - b'\left(\ell(\eta_i)\right)\right).$$

On notera lorsqu'il existe, $\widehat{\boldsymbol{\vartheta}}_n = \arg\max_{\boldsymbol{\vartheta}} L(\boldsymbol{\vartheta}|y_1, \dots, y_n)$. Lorsque le modèle est identifiable, on peut montrer que la séquence $(\widehat{\boldsymbol{\vartheta}}_n)_{n \geq 1}$ existe asymptotiquement et est un estimateur consistent de $\boldsymbol{\vartheta}$ (on pourra se référer par exemple à Fahrmeir et Kaufmann (1985)). Néanmoins dans le cas général, l'existence du MLE pour des petites tailles d'échantillons n'est pas garantie et s'il existe n'a en général pas de forme explicite dû à la non-linéarité du système $S_j(\boldsymbol{\vartheta}) = 0$. Des méthodes d'algorithme de descentes de gradients de type Newton-Raphson sont généralement employées pour obtenir une valeur approchée de $\widehat{\boldsymbol{\vartheta}}_n$.

2 Résultats

Dans cette contribution, nous nous sommes intéressés au modèle GLM pour lequel les vecteurs de covariables $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ sont catégorielles. Nous détaillons les conditions

d'existence non-asymptotique du MLE, qui dépendent de la distribution mais également du choix de la fonction de lien g. Lorsque ce dernier existe, nous donnons une forme explicite de cet estimateur du paramètre $\boldsymbol{\vartheta}$.

Commençons par présenter le cas **d'une variable catégorielle**. Dans le modèle, $p = 2$, $x_i^{(1)} = 1$ pour tout $i = 1, \dots, n$ (intercept) et $x_i^{(2)}$ possède d modalités notées $(v_1, \dots, v_d)$. À la place de $(1, x_i^{(2)})$, nous considérons

$$\boldsymbol{x}_i = (1, x_i^{(2),1}, \dots, x_i^{(2),d}), \qquad x_i^{(2),j} = 1_{\{x_i^{(2)} = v_j\}}.$$

Le GLM considéré est donc le suivant :

$$g\left(\mathbf{E}Y_i\right) = \vartheta_{(1)} + \sum_{j=1}^{d} x_i^{(2),j} \vartheta_{(2),j}, \quad i = 1, \dots, n,$$

où $\boldsymbol{\vartheta} = (\vartheta_{(1)}, \vartheta_{(2),1}, \dots, \vartheta_{(2),d})$ est le vecteur de paramètre inconnu à estimer de $\mathbb{R}^{d+1}$.

Le modèle étant surparamétrisé, nous devons imposer une contrainte linéaire sur $\boldsymbol{\vartheta}$:

$$\langle \boldsymbol{R}, \boldsymbol{\vartheta} \rangle = 0, \quad \text{où } \boldsymbol{R} \text{ est ici un vecteur de dimension } d+1 \text{ vérifiant} \quad R_1 - \sum_{j=2}^{d+1} R_j \neq 0.$$

Définissons pour $j = 1, \dots, d$, $\bar{y}^{(j)}$ la moyenne des y_i calculée sur la $j$ème modalité de $x_i^{(2)}$: $\bar{y}^{(j)} = \frac{1}{m_j} \sum_{i=1 | x_i^{(2),j}=1}^{n} y_i$, où $m_j = \#\{i; x_i^{(2),j} = 1\} > 0$ est le nombre d'occurrence de la $j$ème catégorie.

Nous montrons dans ce cas (Brouste et al. 2019) que le MLE existe si pour $i = 1, \dots, n$ Y_i est à valeur dans l'espace $b'(\Lambda)$. Dans ce cas, le MLE de $\boldsymbol{\vartheta}$ est donné par la formule explicite

$$\widehat{\boldsymbol{\vartheta}}_n = (\boldsymbol{Q}'\boldsymbol{Q} + \boldsymbol{R}'\boldsymbol{R})^{-1} \boldsymbol{Q}' g(\bar{\boldsymbol{Y}}), \tag{2}$$

où $g(\bar{\boldsymbol{Y}})$ est le vecteur $(g(\bar{Y}^{(j)}))_{j=1,\dots,d}$ et Q est la matrice de dimension $d \times d+1$ donnée par

$$\boldsymbol{Q} = \begin{pmatrix} 1 & 1 & & 0 \\ \vdots & & \ddots & \\ 1 & 0 & & 1 \end{pmatrix}.$$

Considérons à présent le cas **de deux variables catégorielles**. Le cas de $p > 2$ variables catégorielles étant similaire nous ne le présenterons pas dans cette contribution.

Notons d_2 et d_3 le nombre de modalités de la première et de la seconde variable catégorielle respectivement. Pour $i = 1, \dots, n$, $k \in K = \{1, \dots, d_2\}$ et $l \in L = \{1, \dots, d_3\}$ soit $x_i^{(2),k}$ (resp. $x_i^{(3),l}$) valant 1 si la première covariable prend la $k$ème modalité (resp. valant 1 si la seconde covariable prend la $l$ème modalité) et $x_i^{(k,l)} = x_i^{(2),k} x_i^{(3),l}$. Dénotons

$m_{k,l}$ le nombre d'occurrences simultanées de la $k$ème et $l$ème catégorie de la première et seconde variable. Soit $KL = K \times L$ et de la même façon que précédemment, dénotons $\overline{y}_n^{(k,l)}$ la moyenne des y_i prise sur la $k$ème et $l$ème categorie. Pour simplifier nous présentons uniquement ici le cas $m_{k,l} > 0$; une légère adaptation permet de traiter également le cas le cas où $m_{k,l} \geq 0$.

Le GLM considéré est le suivant (modèle avec interaction) :

$$g\left(\mathbf{E}Y_i\right) = \vartheta_1 + \sum_{k=1}^{d_2} x_i^{(2),k} \vartheta_{(2),k} + \sum_{l=1}^{d_3} x_i^{(3),l} \vartheta_{(3),l} + \sum_{(k,l)\in K\times L} x_i^{(k,l)} \vartheta_{kl}$$

où $\boldsymbol{\vartheta} = (\vartheta_{(2)}, (\vartheta_{(2),k})_{k\in K}, (\vartheta_{(3),l})_{l\in L}, (\vartheta_{kl})_{(k,l)\in KL})$ est le vecteur de taille $d_2 + d_3 + d_2 d_3 + 1$ de paramètre à estimer.

Comme précédemment, nous avons à imposer des contraintes linéaires

$$\langle \boldsymbol{R}, \boldsymbol{\vartheta} \rangle = \mathbf{0}_q. \tag{3}$$

où $\boldsymbol{R}$ est ici une matrice de taille $q \times (1 + d_2 + d_3 + d_2 d_3)$, avec $q \geq 1 + d_2 + d_3$ et $rank(\boldsymbol{R}) = 1 + d_2 + d_3$.

Avec les mêmes considérations d'existence que dans le cas d'une variable explicative, dès que la matrice $(\boldsymbol{Q}'\, \boldsymbol{R}')$ est de rang plein (c.-à-d. $d_2 d_3$), où $\boldsymbol{Q}$ est la matrice de dimension $d_2 d_3 \times 1 + d_2 + d_3 + d_2 d_3$ de la forme

$$\boldsymbol{Q} = \begin{pmatrix}
1 & 1 & & 1 & & 0 & 1 & & & & & \\
\vdots & & & & \ddots & & & \ddots & & & & \\
\vdots & 1 & 0 & 0 & & 1 & & & \ddots & & & \\
\vdots & 0 & 1 & 1 & & 0 & & & & \ddots & & \\
\vdots & \vdots & \vdots & & \ddots & & & & & & \ddots & \\
1 & 0 & 1 & 0 & & 1 & & & & & & 1
\end{pmatrix},$$

le MLE est donné par

$$\widehat{\boldsymbol{\vartheta}}_n = (\boldsymbol{Q}'\boldsymbol{Q} + \boldsymbol{R}'\boldsymbol{R})^{-1}\boldsymbol{Q}'\boldsymbol{g}(\bar{\boldsymbol{Y}}), \tag{4}$$

avec $g(\bar{\boldsymbol{Y}}) = (g(\bar{Y}^{(k,l)}))_{(k,l)\in KL}$.

Si l'on considère le modèle sans interaction,

$$g\left(\mathbf{E}Y_i\right) = \vartheta_1 + \sum_{k=1}^{d_2} x_i^{(2),k} \vartheta_{(2),k} + \sum_{l=1}^{d_3} x_i^{(3),l} \vartheta_{(3),l}$$

le MLE **n'a pas de forme explicite** dans le cas général, cependant nous proposons un estimateur de type Least Square de la forme (4) et possédant les mêmes propriétés asymptotiques que le MLE.

A titre d'exemple, pour le modèle Pareto avec la fonction de lien canonique, pour $n = 1400$ et une variable catégorielle avec $d = 5$ modalités, le nombre d'opérations en virgule flottante (FLOP) utilisées dans le calcul de l'estimateur exact est environ 5000 fois moins important qu'en utilisant un algorithme IWLS (qui utilise 5 itérations pour obtenir la solution). Ces simulations peuvent être retrouvées dans (Brouste et al. 2019, section 6).

Bibliographie

Brouste, A., Dutang, C., Rohmer, T. (2019). Closed form Maximum Likelihood Estimation for Generalized Linear Models in the case of categorical explanatory variables : application to insurance loss modelling, *Computational Statistics, 35, 689-724, p. 1-36 (2020).*

Fahrmeir, L., Kaufmann, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models, The Annals of Statistics pp. 342–368.

McCullagh, P., Nelder, J. A. (1989). Generalized linear models, Vol. 37, CRC press.

Nelder, J. A., Wedderburn, R. W. M. (1972). Generalized linear models, Journal of the Royal Statistical Society. Series A **135***(3), 370–384.*

Sensibilité des classements vis-à-vis des paramètres : l'exemple de Parcours Sup

Antoine Rolland

IUT Lumière Lyon II, 69676 BRON, antoine.rolland@univ-lyon2.fr

Résumé. Le classement des candidats en entrée de formation post-bac via la plateforme "Parcours Sup" est régulièrement critiqué pour son opacité. Nous analysons ici, à travers l'exemple d'un DUT STID, la sensibilité du classement vis-à-vis de la pondération choisie de chacune des variables servant à calculer le score des futurs étudiants. Nous utilisons pour cela un Indice de Sensibilité du Classement inspiré de l'analyse de variance.

Mots-clés. Parcoursup, Classement, Sensibilité

Abstract. In France, future students have to apply for graduate studies via a web platform named "Parcoursup", which is regularly criticized for its opacity. Each formation must rank its candidates, generally through a multivariate analysis. We analyse in this paper how the weights given to the variables can affect the final ranking, through the example of a DUT STID (Statistics and Business Intelligence). We use in this purpose a dedicated Sensitivity Ranking Index inspired by ANOVA.

Keywords. Parcoursup, Ranking, sensitivity

1 La sensibilité des classements

La production de classements ou de palmarès à partir de plusieurs critères (ou variables) est un exercice à la fois très répandu, en particulier dans les médias, et pourtant délicat. Le "palmarès des villes où il fait bon vivre/étudier/travailler", le palmarès des hôpitaux, sans oublier le fameux palmarès mondial des universités dit "de Shangaï", sont autant de marronniers qui reviennent quasiment chaque année. Ces palmarès sont très souvent obtenus par l'utilisation d'une moyenne pondérée, les poids étant fixés de manière discrétionnaire par les auteurs du palmarès. Récemment, Rolland et Cugliari (2020) ont proposé un indice permettant de mesurer la sensibilité du classement final vis-à-vis du choix des paramètres. Après avoir rappelé brièvement comment est calculé cet indice, nous nous attacherons à présenter une application concrète dans le cas du classement des candidats via le processus Parcours Sup dans le cadre du DUT STID de l'université Lumière Lyon II.

1.1 Notations

On note $\mathcal{X}_m$ un ensemble de m individus décrits par p variables $X_1, \ldots, X_p$. On suppose que $\forall i \in 1, \ldots, p,\ X_i = \mathbb{I} \subseteq \mathbb{R}$, i.e. les X_i sont tous identiques (toutes les variables utilisent la même échelle). On note x_{ij} la valeur de la j^{eme} variable pour l'individu i, avec $i = 1, \ldots, m$ et $j = 1, \ldots, p$.

On note également $\mathcal{W}$ le simplexe sur $\mathbb{R}^p$, i.e. l'ensemble des vecteurs de probabilité de p valeurs positives réelles :$\mathcal{W} = \{(w_1, \ldots, w_p) \in \mathbb{R}^p | w_j \geq 0, j = 1, \ldots p; \sum_{j=1}^p w_j = 1\}$.

Une moyenne pondérée f_w, $w \in \mathcal{W}$, est une fonction d'agrégation $f_w : \mathbb{I}^p \mapsto \mathbb{I}$ telle que $\forall x = (x_1, \ldots, x_p),\ f_w(x) = \sum_{j=1}^p w_j x_j =< w, x >$. Une moyenne pondérée donnée dépend donc d'un vecteur de paramètres donné $w = (w_1, \ldots, w_p)$.

On peut alors utiliser le résultat de la moyenne pondérée f_w pour déterminer un classement sur les m individus, prenant en compte les valeurs prises par les individus sur les variables et le vecteur poids. La question est alors la suivante : étant donné un ensemble d'individus, et une famille de vecteurs de poids $\{f_w, w \in \mathcal{W}\}$, quelle est l'influence du paramètre w sur le classement final? En d'autres terme, le classement sur les individus est-il sensible à une modification des poids dans la fonction d'agrégation?

Rolland et Cugliari (2020) répondent à cette question en proposant un indice global mesurant la sensibilité du classement d'un jeu de données spécifique vis-à-vis du changement des paramètres à travers un indice de sensibilité du classement (ISC) de $\mathcal{X}_m$ fondé sur l'utilisation d'une approche par analyse de variance.

1.2 Indice de sensibilité du classement.

Le classement des individus peut donc varier en fonction de deux sources : (1) les valeurs prises par les m individus x sur chaque variable $j \in \{1, \ldots, p\}$, et (2) le vecteur de poids w.

Soit W une variable aléatoire prenant ses valeurs sur le simplexe $\mathcal{W}$. La loi de W est notée $\mathcal{L}(W)$. Pour chaque individu $x_i \in \mathcal{X}_m$ on définit une moyenne pondérée aléatoire $Y_i =< W, x_i >,\ i = 1, \ldots, m$. On s'intéresse au classement de ces moyennes pondérées aléatoires $\{\rho_i := \text{rank}(Y_i), i = 1, \ldots, m\}$.

Sous des conditions simples, par exemple si $\mathcal{L}(W)$ admet une densité de probabilité continue, alors on peut définir l'espérance $\bar{\rho_i} = \mathbb{E}_W[\rho_i]$ et la variance $var_i = \mathbb{E}_W[\rho_i - \bar{\rho_i}]^2$.

On définit également deux valeurs non aléatoire : la moyenne des rangs $\bar{\rho} = m^{-1} \sum_{i=1}^m \rho_i = \frac{m+1}{2}$, et la somme totale du carré des variations

$$SSD_{\text{total}}^{\mathcal{X}_m, W} = \sum_{i=1}^{m} (\rho_i - \bar{\rho})^2.$$

On peut noter que cette somme du carré des écarts peut se décomposer de manière classique comme

$$\mathbb{E}_W[\rho_i - \bar{\rho}]^2 = var_i + \mathbb{E}_W[\bar{\rho_i} - \bar{\rho}]^2.$$

On peut alors définir l'indice de sensibilité du classement (ISC) de $\mathcal{X}_m$ vis-à-vis de $\mathcal{L}(W)$, noté $RSI(\mathcal{X}_m)_{\mathcal{L}(W)}$, comme la part restante de l'erreur carrée moyenne expliquée par les variations du rang moyen, i.e.

$$ISC(\mathcal{X}_m)_{\mathcal{L}(W)} = 1 - \frac{\sum_{i=1}^{m} var_i}{SSD_{\text{total}}^{\mathcal{X}_m,W}} \tag{1}$$

Dans le cas d'un classement total sans *ex aequo*, $SSD_{\text{total}}^{\mathcal{X}_m,W} = \sum_{i=1}^{m}(i - \frac{m+1}{2})^2 = \frac{m(m^2-1)}{12}$. ISC s'écrit donc comme suit :

$$ISC(\mathcal{X}_m)_{\mathcal{L}(W)} = 1 - \frac{12\sum_{i=1}^{m} var_i}{m(m^2-1)} \tag{2}$$

Le cas le plus intéressante est celui où $\mathcal{L}(W)$ est la distribution uniforme sur $\mathcal{W}$. $ISC(\mathcal{X}_m)_{\mathcal{L}(W)}$ sera alors noté simplement $ISC_{\mathcal{X}_m)}$. On peut alors procéder à une estimation de $ISC(\mathcal{X}_m)_{\mathcal{L}(W)}$ par une méthode de Monte-Carlo, ce qui donne de très bons résultats d'après Rolland et Cugliari (2020). L'indice $ISC_{\mathcal{X}_m}$ s'interprète alors comme suit :

- si $ISC_{\mathcal{X}_m} = 1$, cela signifie que $\forall i \in 1, \ldots, m$, $var_i = 0$, autrement dit que le rang de x_i est toujours le même quel que soit le jeu de poids, et donc que le classement sur $\mathcal{X}_{\Updownarrow}$ ne dépend pas du jeu de poids : toute l'information est dans les données
- si $ISC_{\mathcal{X}_m} = 0$, cela signifie que $\forall i \in 1, \ldots, m$, $var_i = var(1, \ldots, m)$, autrement dit le rang de x_i est complètement dépendant du jeu de poids choisi.

Dans une perspective d'analyse, plus l'indice est proche de 1, moins il dépend du choix (parfois arbitraire) du jeu de poids de la moyenne pondérée et donc plus le classement est robuste vis-à-vis du jeu de poids.

1.3 Top-k et middle-k listes

On peut généraliser l'indice $ISC_{\mathcal{X}_m}$ aux classements partiels. En effet, il arrive que l'intérêt d'un classement réside simplement dans les k premiers ou les k derniers éléments; typiquement, lors d'une recherche internet, seuls les 10 premiers résultats (la première page) sont intéressants. A la suite de Fagin *et al.* (2003), nous nous intéresserons aux top-k listes, mais aussi aux bottom-k listes ainsi qu'aux middle-k listes. De manière formelle, une top-k liste R est une bijection d'un domaine D (ici les membres de la liste) vers l'ensemble $\{1, \ldots, k\}$. Rolland et Cugliari (2020) proposent une extension de cette définition, utilisant l'approche "optimiste" de Fagin *et al.* (2003) : la représentation d'une top-k liste sur un ensemble de m éléments suppose que tous les $m - k$ éléments de la liste qui ne sont pas dans le top-k sont tous classés *ex-aequo* à la $k + 1^{eme}$ position. Une top-k liste R est alors une bijection de $\mathcal{X}_m$ vers l'ensemble $\{1, 2, 3 \ldots, k, k + 1, \ldots, k + 1\}$.

De même, une bottom-k liste représente une situation où seul le classement des derniers de la liste est d'importance. Une bottom-k liste R est alors une bijection de $\mathcal{X}_m$ vers l'ensemble $\{m-k-1,\ldots,m-k-1,m-k,m-k+1,\ldots,m-1,m\}$.

Enfin, une middle-k liste représente une situation où l'intérêt se porte sur les individus classés au milieu de la liste. Une middle-k liste R est en fait définie par une bijection de $\mathcal{X}_m$ vers l'ensemble $\{d-1,\ldots,d-1,d,d+1,\ldots,f-1,f,f+1\ldots,f+1\}$, avec $f-d=k$.

Dans ces trois cas il est donc nécessaire de revenir à la définition de $ISC(\mathcal{X}_m)$ tel qu'exprimé dans l'équation 1, en calculant pour chaque cas la quantité $SSD_{\text{total}}^{\mathcal{X}_m,W} = \sum_{i=1}^{m}(\rho_i-\bar{\rho})^2$.

2 Application à Parcoursup

2.1 Le classement en entrée du DUT STID

La plate-forme Parcoursup permet depuis 2018 aux futurs étudiants de candidater aux différentes formations de l'enseignement supérieur qui les intéressent. Réciproquement, cette même plate-forme Parcoursup permet aux formations d'instruire les dossiers de candidature puis de classer les candidats dans un ordre de préférence. Les candidats se voient ensuite proposé dans cet ordre une place dans la formation, qu'ils acceptent ou refusent. Le processus continue ensuite jusqu'à épuisement soit de la capacité d'accueil de la formation, soit de la liste des candidats.

Le DUT STID de l'IUT Lumière Lyon II est une formation sélective, qui reçoit (bien) plus de candidatures qu'elle n'a de places. Il s'agit donc pour l'équipe pédagogique d'analyser le niveau et la motivation des étudiants suivant un processus utilisant trois informations : une note reflétant le *niveau scolaire* de l'étudiant, estimé à travers ses bulletins de notes; une note reflétant la *connaissance* de la formation et de ses débouchés, obtenue via le remplissage d'un dossier écrit spécifique; une note reflétant la *motivation* de l'étudiant et l'adéquation de son projet aux spécificités de l'IUT[1], obtenue suite à un entretien individuel avec le candidat.

Le classement final est obtenu suite au calcul d'une moyenne pondérée de ces trois notes. La question se pose donc de la sensibilité du classement final, conditionnant l'acceptation ou non dans la formation, au choix de la pondération des trois notes. Nous sommes donc exactement dans le cas d'utilisation de l'indice de sensibilité du classement tel que défini supra.

Il y a eu, pour la campagne de recrutement 2019, 184 candidats qui ont obtenu une note aux entretiens individuels[2], pour environ 54 places dans la formation. Il est certain que les 54 premiers du classement sont certains d'être pris : il est donc inutile de s'attarder

[1]Toutes les formations de l'IUT Lumière Lyon II se déroulent exclusivement en alternance

[2]Plusieurs candidats n'ont pas été convoqués aux entretiens par faute d'un dossier suffisant. D'autres candidats, convoqués, ne se sont pas présentés aux entretiens.

sur le classement exact de ces 54 premiers. De même, l'expérience des années précédentes montre que l'on descend jusqu'au 100^{eme} ou 110^{eme} candidat à peu près avant de remplir les 54 places. Le classement relatif des candidats classés au delà de la 100^{eme} place n'a donc pas d'importance, puisqu'ils ne seront pas admis. Nous sommes donc typiquement dans le cas d'un classement en trois tiers : le premier tiers du classement est assuré d'être pris, et le rang n'a donc pas d'importance. Le dernier tiers est assuré de ne pas être pris. Reste donc à étudier la sensibilité du classement du tiers du milieu, où une place de différence peut faire basculer du statut d'admis à celui de refusé. Nous sommes donc formellement en présence d'une middle-k liste, avec $d = 60$, $f = 124$ et $m = 184$. Les valeurs de d et f sont arbitraires et d'autres valeurs de d et f pourraient tout à fait être choisies.

2.2 Calcul de l'indice de sensibilité

2.2.1 Classement entier

Nous pouvons calculer tout d'abord une estimation de l'indice de sensibilité $ISC_{\mathcal{X}_m}$ par simulation sur n jeux de poids obtenus par tirage alétoire suivant une loi de Dirichlet de paramètres $(1/3; 1/3; 1/3)$, correspondant à la distribution uniforme sur le simplexe. Une rapide simulation sur 100 jeux de poids nous donne une estimation de $\widehat{ISC} \approx 0,80$. Une répétition de 1000 simulations nous permet d'obtenir une estimation de l'écart-type de l'estimateur de ISC, qui est très faible (0,013).

2.2.2 Middle-k liste

Le calcul de l'indice de sensibilité $ISC_{d,f}$ d'une middle-k liste nécessite de calculer dans un premier temps la valeur exacte de $SSD_{\text{total}}^{\mathcal{X}_m,W} = \sum_{i=1}^{m}(\rho_i - \bar{\rho})^2$. Nous calculons ensuite une estimation de $ISC_{d,f}$ en tirant n jeux de poids. Nous choisissons d'effectuer une rapide simulation sur 100 jeux de poids, ce qui nous donne une estimation de $\widehat{ISC}_{60,124} \approx 0,75$. Une répétition de 1000 simulations nous permet d'obtenir une estimation de l'écart-type de l'estimateur de $ISC_{60,124}$, qui est également très faible (0,013).

Il est également possible de regarder les valeurs prises par ISC en faisant varier d et f. La figure 1 montre les estimations de ISC pour des valeurs de (d, f) allant de $(50, 134)$ à $(70, 114)$, autrement dit les estimations de ISC pour les k classements centraux, k variant entre 44 et 84. Ces valeurs ont été obtenues par la méthode de Monte-Carlo en calculant, pour chaque valeur de d, 1000 occurences de $\widehat{ISC}$ obtenues à partir de 100 jeux de poids différents.

2.2.3 Interprétation

L'indice de sensibilité du classement des candidats à l'entrée du DUT STID se situe autour de 0,75 pour le classement des k candidats du milieu de la liste, et à 0,80 pour le

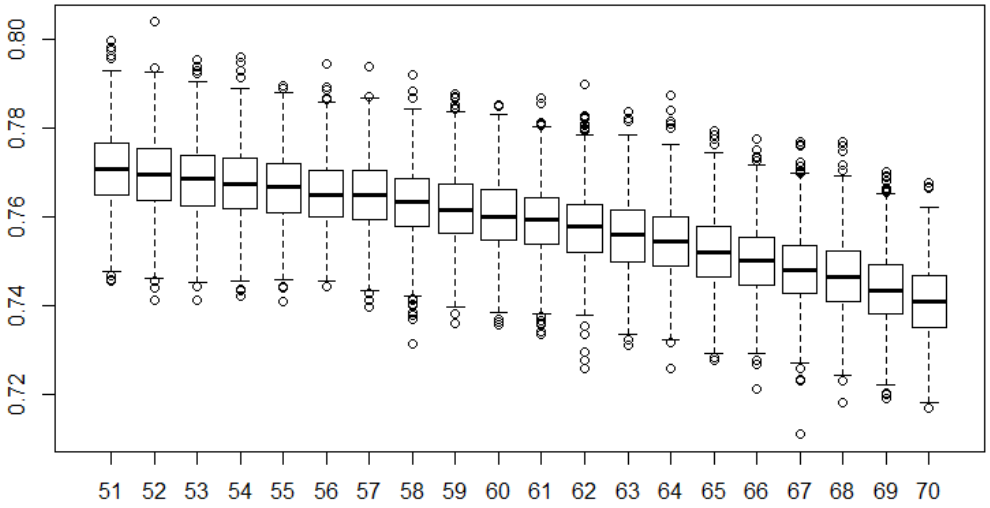


Figure 1: Les 1000 estimations de ISC

classement entier : cela signifie que le classement obtenu est très robuste vis-à-vis d'un changement de poids des critères dans la définition du rang. Autrement dit, le classement final est très déterminé par les notes obtenues par les candidats sur les trois critères, et peu déterminé par le poids de ces critères. Cela est une relativement bonne nouvelle, en particulier pour le cas des étudiants en milieu de classement : un changement de poids des différents critères de sélection n'aurait pas fondamentalement changé l'ordre des candidats, candidats qui ne doivent donc pas leur classement uniquement aux poids des critères mais bien aux valeurs obtenues sur ces mêmes critères.

Remerciements

L'auteur tient à remercier Anne Perrut pour lui avoir suggéré l'idée de cette analyse.

Bibliographie

R. Fagin, R. Kumar, and D. Sivakumar (2003), Comparing top k lists, *J. Disc. Math.* 17, pp. 134–160.

A. Rolland and J. Cugliari (2020), Sensitivity index to measure dependence on parameters for rankings and top-k rankings, *Journal of Applied Statistics*.

Inférence bayésienne de l'évolution de l'atrophie cérébrale et de plages de leucopathie à partir de séquences IRM 3D non homogénéisées

Julien Roussel[1], Sophie Ancelet[2], Samy Djazoubi[1], Geoffray Brelurut[1], Cécilia Damon[2], Monica Ribeiro[3], Nadya Pyatigorskaya[3], Marie-Odile Bernier[2], Damien Ricard[4] & Nicolas Bousquet [1,5]

[1] *Quantmetry, Paris. email: jroussel@quantmetry.com*
[2] *Institut de Radioprotection et de Sûreté Nucléaire, Fontenay-aux-Roses*
[3] *Hôpital de la Pitié-Salpêtrière, 47 boulevard de l'Hôpital, 75013 Paris*
[4] *HIA Percy, Clamart & UMR 8257 MD 4 « Cognac-G », Paris*
[5] *Sorbonne Université, LPSM, 4 place Jussieu, 75005 Paris*

Résumé. Ce travail décrit un algorithme synchrone de Gibbs réalisant une segmentation bayésienne non supervisée de régions du cerveau à partir de modèles de Markov cachés et de séquences IRM 3D non homogénéisées, issues d'examens IRMs réalisés dans le cadre d'un protocole non standardisé. L'inférence de ce type de modélisation bien connu pose encore de nombreuses difficultés, liées aux biais d'illumination et aux temps de calculs. Elle permet néanmoins d'estimer rapidement l'évolution spatio-temporelle de marqueurs radiologiques de la neurotoxicité d'un traitement par radiothérapie (e.g., oedème cérébral, atrophie cérébrale, plages de leuco-encéphalopathie). L'application porte sur les données IRMs de la cohorte prospective bicentrique EpiBrainRad incluant des patients atteints de gliomes cérébraux de haut grade, traités par radiothérapie et suivis en routine par IRM.

Mots-clés. Segmentation d'image, IRM cérébrales, champs de Markov cachés, temps de calcul, statistique bayésienne, neurotoxicité de la radiothérapie

Abstract. This work describes a synchronous Gibbs algorithm that allows fitting a hidden Markov random field addressing the problem of unsupervised Bayesian brain areas segmentation, using non-homogenized 3D MRI sequences collected in the context of a non-standardized protocol. While the principle of such a segmentation is well known, the inference task still poses many difficulties related to illumination biases and computational time. The proposed approach makes it possible to rapidly estimate the spatio-temporal evolution of imaging markers of neurotoxicity from radiotherapy (e.g., cerebral edema, brain atrophy, leukoencephalopathy areas). It was implemented on the MRIs from the prospective and bicentric cohort EpiBrainRad including newly diagnosed glioblastoma patients, treated by radiotherapy and routinely monitored by MRI.

Keywords. Image segmentation, brain MRI, hidden Markov random fields, computational time, Bayesian statistics, neurotoxicity of radiotherapy

1 Contexte et objectif

La radiothérapie occupe une place importante dans le traitement des tumeurs cérébrales. L'augmentation du temps de survie des patients peut néanmoins s'accompagner du développement à plus ou moins long terme de complications neurologiques post-radiques comme une atrophie du cerveau et/ou une leuco-encéphalopathie. Cette dernière est une anomalie progressive et diffuse de la substance blanche. Cliniquement, elle se manifeste par des troubles de l'attention, de la mémoire, un ralentissement et/ou des troubles exécutifs. Malgré l'augmentation des études sur le sujet, il n'existe pas de consensus sur l'évolution générale des fonctions cognitives au cours d'une leuco-encéphalopathie et les processus en jeu demeurent mal compris.

Dans ce contexte, le projet EpiBrainRad (Durand *et al.* 2015) s'intéresse à l'analyse de données cliniques, biologiques et radiologiques issues d'une cohorte prospective bicentrique de 250 patients traités par radiothérapie pour un gliome cérébral de haut grade. Les patients sont inclus avant le début de la radiothérapie puis bénéficient d'un suivi classique : évaluations cliniques et neuropsychologiques pré/post-radiothérapie et examens d'IRM tous les 2-3 mois après radiothérapie. L'objectif principal du projet est d'évaluer l'incidence et l'évolution de troubles cognitifs radio-induits dans cette cohorte, notamment par l'analyse de l'évolution des lésions mentionnées précédemment.

En pratique, l'appréciation visuelle, par le neuroradiologue, de l'atteinte de la substance blanche est subjective, basée sur des échelles qualitatives de scores relativement simples mais trop peu précises pour pouvoir évaluer l'évolution de cette atteinte en fonction de la dose reçue par une structure cérébrale spécifique. De même, l'appréciation visuelle de l'atrophie cérébrale est souvent globale et reflète mal les atteintes spécifiques. Un des axes de recherche du projet EpibrainRad est de développer un outil automatique robuste de segmentation et de quantification précise de l'évolution spatio-temporelle de l'atrophie cérébrale et des plages de leucopathie, à partir de séquences d'IRMs 3D déformées et hétérogènes, issues d'un protocole non standardisé. Cet outil devra être utilisable par les radiologues pour les aider à objectiver l'évaluation des stades d'évolution de ces lésions et devra permettre de renseigner les études épidémiologiques ultérieures.

2 Difficultés techniques

Les marqueurs radiologiques définis pour caractériser l'évolution des lésions post-radiques sont : le volume du cerveau, de la substance blanche, de la substance grise et du liquide céphalo-rachidien ainsi que le volume des plages de leucopathie. Le calcul de ces indicateurs repose sur l'estimation du nombre de voxels correspondant à chacune de ces régions sur une séquence IRMs 3D, ce qui nécessite de segmenter ou classifier ces régions, en fonction des caractéristiques d'illumination et d'homogénéité de chaque voxel.

Les plages de leucopathie se caractérisent par une forte intensité (*hypersignal*) de la substance blanche sur des images FLAIR. Dans le cas d'étude considéré, l'apprentissage à

mener, non supervisé, ne peut s'appuyer sur une connaissance *a priori* forte de la présence de ces lésions (comme des données de contour extrêmement longues à produire par des neuroradiologues). Les techniques reposant sur des réseaux de neurones convolutionnels ne peuvent donc être utilisées. L'inférence bayésienne de modèles de champs de Markov cachés à l'aide de MCMC, qui fait autorité dans le cadre non supervisé (Moores *et al.* 2015), semble plus appropriée. Cette approche a été adaptée pour répondre aux deux difficultés suivantes : 1) la nécessité de forts recalages et débruitages, liés en particulier au fait que les données proviennent de machines différentes, utilisées par des opérateurs différents ; 2) permettre la construction d'un pipeline automatisé de traitement des historiques de séquences en un temps raisonnable, dans un esprit similaire à Guizard *et al.* (2012), par le biais d'accélérations algorithmiques.

L'utilisation conjointe de plusieurs modalités d'IRM issues du même examen est essentielle pour limiter l'impact des bruits et décalages provenant d'examens réalisés en routine. Comme l'illustre la figure 1, les histogrammes d'intensité des IRMs T1 et FLAIR acquises dans la cohorte EpiBrainRad diffèrent grandement, de par leur unimodalité, des IRMs classiquement rencontrées dans les études cliniques très contrôlées et dont la multimodalité permet de différencier les types de régions du cerveau. L'utilisation simultanée des IRMs T1 et FLAIR permet néanmoins de pallier aux difficultés induites par le manque de contraste (notamment sur les plages de leucopathie) et l'hétérogénéité des données.

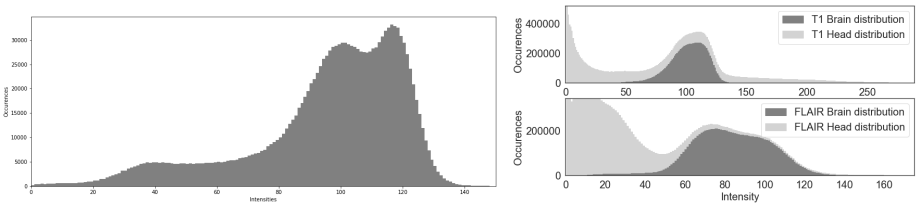


Figure 1: À gauche : Histogramme d'intensité d'une IRM T1 "propre". À droite : Cas des IRMs T1 et FLAIR acquises lors d'un même examen dans EpiBrainRad.

3 Modélisation bayésienne d'une IRM cérébrale 3D

On définit une image 3D de résolution $N_x \times N_y \times N_z$ comme une hypermatrice $\mathbf{y} \in \Omega^{N_x \times N_y \times N_z}$ où $\Omega = \mathbb{R}_+$ désigne l'ensemble des intensités possibles pour chaque voxel de l'image. Soit $\mathcal{S} = \{1, \ldots, N_x\} \times \{1, \ldots, N_y\} \times \{1, \ldots, N_z\}$ l'ensemble des voxels d'une image IRM 3D. Le but de la segmentation est d'étiqueter chaque voxel $s \in \mathcal{S}$ avec un label entier $x_s \in \mathcal{A} = \{1, \ldots, D\}$, où $D = 4$ est le nombre des différentes régions d'intérêt du cerveau. Les nombres 1, 2, 3 et 4 représentent respectivement le liquide céphalo-rachidien, la substance grise, la substance blanche et les plages de leucopathie. Le calcul de la distribution de probabilité conditionnelle $\pi(\mathbf{X}|\mathbf{Y} = \mathbf{y}^*)$ d'un vecteur de labels $\mathbf{X} = (X_s)_{s \in \mathcal{S}}$

(avec $X_s \in \mathcal{A}$), connaissant l'image $\mathbf{y}^*$, repose sur les ingrédients suivants :

1 - *Vraisemblance* $\pi(\mathbf{Y}|\mathbf{X})$: pour une seule modalité d'IRM, les intensités Y_s des voxels sont supposées conditionnellement indépendantes (sachant $\mathbf{X}$) et suivre une distribution gaussienne (tronquée en pratique) d'espérance μ_{X_s} et de variance $\sigma^2_{X_s}$:

$$\pi(\mathbf{Y} = \mathbf{y}|\mathbf{X} = \mathbf{x}, \theta) = \prod_{s\in\mathcal{S}} \pi(Y_s = y_s|X_s = x_s, \theta) = \prod_{s\in\mathcal{S}} \frac{1}{\sqrt{2\pi}\sigma_{x_s}} \exp\left(-\frac{(y_s - \mu_{x_s})^2}{2\sigma^2_{x_s}}\right).$$

avec: $\theta = (\mu_1, \sigma_1, \cdots, \mu_D, \sigma_D) \in \Theta = \mathbb{R}^{2D}_+$.
2 - *Loi a priori* $\pi(\mathbf{X})$: suivant les choix proposés dans (Leemput, Weisenfeld *et al.* 1999) , un modèle de champ de Markov caché, appelé modèle de Potts, est considéré :

$$\pi(\mathbf{X}) = Z^{-1} \exp(-U(\mathbf{X})), \qquad U(\mathbf{x}) = \sum_{s\sim s'} v(x_s, x_{s'})\, S(s, s'),$$

où on somme sur les couples de voxels voisins et $Z > 0$ est une constante de normalisation. U pénalise les voisinages entre voxels de labels différents. Cette pénalité dépend de la surface de contact S entre les deux voxels, et de leurs labels respectifs. En effet, l'affinité entre régions est définie par $D \times (D-1)/2$ hyperparamètres $v(k, k')_{k,k'\in\mathcal{A}}$ qui ont été fixés empiriquement (entiers entre 1 et 10) à partir de connaissances métiers (e.g., la substance grise n'est pas voisine de plages de leucopathie). Six voxels voisins ont été considérés pour chaque voxel de l'image (réduits à 5, 4 ou 3 sur les bords de l'image 3D) – un choix courant (Moores *et al.* 2015) imposé par la complexité du problème.
3 - *Loi a priori* $\pi(\theta)$: Afin de bénéficier de propriétés de conjugaison, des lois *a priori* informatives gaussiennes ont été choisies pour les paramètres μ_k et inverse-Gamma pour les paramètres σ^2_k avec $k = \{1, \ldots, D\}$, guidées par une analyse manuelle d'IRMs cérébrales.

3.1 Algorithme synchrone de Gibbs

Un algorithme MCMC peut être implémenté pour échantillonner selon la loi *a posteriori* $\pi(\mathbf{X}, \theta|\mathbf{Y} = \mathbf{y}^*)$. Si la mise à jour des paramètres (μ_k, σ^2_k) peut être réalisée à l'aide d'un pas de Gibbs, par conjugaison, celle du vecteur des labels $\mathbf{X}$ se révèle très coûteuse en calculs, notamment lorsque celle-ci est réalisée composante par composante.

Aussi, un échantillonnage plus performant du vecteur des labels, appelé algorithme *synchrone* de Gibbs (Gonzalez *et al.* 2011), a été implémenté et testé. Il repose sur l'observation que les sites $s \in \mathcal{S}$ peuvent être séparés en deux classes $\mathcal{S}_+$ et $\mathcal{S}_-$ selon un damier tridimensionnel ($s_{i,j,k} \in \mathcal{S}_+ \Leftrightarrow i+j+k$ est pair; $s_{i,j,k} \in \mathcal{S}_- \Leftrightarrow i+j+k$ est impair), et en utilisant la formule explicite suivante:

$$\begin{aligned}\pi(\mathbf{X}_+ = \mathbf{x}_+|\mathbf{Y} = \mathbf{y}^*, \theta, \mathbf{X}_- = \mathbf{x}_-) &\propto \pi(\mathbf{Y}_+ = \mathbf{y}^*_+|\mathbf{X}_+ = \mathbf{x}_+, \theta)\pi(\mathbf{X}_+ = \mathbf{x}_+|\mathbf{X}_- = \mathbf{x}_-) \\ &\propto \prod_{s\in\mathcal{S}_+} \pi(\mathbf{Y}_s = \mathbf{y}^*_s|\mathbf{X}_s = \mathbf{x}_s, \theta)\, \pi(\mathbf{X}_s = \mathbf{x}_s|\mathbf{X}_- = \mathbf{x}_-)\end{aligned}$$

où $\mathbf{X}_+ = (X_s)_{s\in\mathcal{S}_+}$ et $\mathbf{X}_- = (X_s)_{s\in\mathcal{S}_-}$. Les formules restent inchangées en inversant les signes + et −. Par ailleurs, les lois conditionnelles complètes $\pi(\mathbf{X}_+ = \mathbf{x}_+|\mathbf{Y} = \mathbf{y}^*, \theta, \mathbf{X}_- = \mathbf{x}_-)$ et $\pi(\mathbf{X}_- = \mathbf{x}_-|\mathbf{Y} = \mathbf{y}^*, \theta, \mathbf{X}_+ = \mathbf{x}_+)$ sont simples à simuler: il s'agit de distributions catégorielles indépendantes. L'échantillonneur synchrone s'écrit alors :

- `Fixons` $X^{-1,J-1} = \mathbf{x}^*$,
- `Pour` i `de` 0 `à` $I-1$:
  1. `simuler` $\theta^i \sim \pi(\theta|\mathbf{Y} = \mathbf{y}^*, \mathbf{X} = \mathbf{X}^{i-1,J-1})$;
  2. `Pour j allant de` 0 `à` $J-1$:
     (a) `simuler` $\mathbf{X}_+^{i,j} \sim \pi(\mathbf{X}_+|\mathbf{Y} = \mathbf{y}^*, \theta^i, \mathbf{X}_- = \mathbf{x}_-^{i,j-1})$
     (b) `simuler` $\mathbf{X}_-^{i,j} \sim \pi(\mathbf{X}_-|\mathbf{Y} = \mathbf{y}^*, \theta^i, \mathbf{X}_+ = \mathbf{x}_+^{i,j})$
     (c) `définir` $\mathbf{X}^{i,j}$ `par combinaison de` $\mathbf{X}_+^{i,j}$ `et` $\mathbf{X}_-^{i,j}$
- `Produire l'échantillon` $(\mathbf{X}^{i,j}, \theta^i)_{0\leqslant i<I, 0\leqslant j<J}$.

Gonzalez *et al.* (2011) montrent en fait que l'on obtient des résultats équivalents en remplaçant $\mathbf{x}_+^{i,j}$ par $\mathbf{x}_+^{i,j-1}$ dans (b), permettant de mettre à jour la totalité de $\mathbf{X}$ de façon synchrone. Il s'agit de la version implémentée en pratique.

4 Résultats

La figure 2 présente les résultats de segmentation obtenus en appliquant le pipeline proposé à un des patients de la cohorte EpibRainRad, atteint de leucopathie. La plage de leucopathie en hypersignal (blanc) sur l'IRM a été labellisée comme telle (en vert). La figure 3 présente l'estimation bayésienne des volumes des différentes régions d'intérêt à chacun des examens réalisés par ce même patient. L'évolution temporelle des volumes (en variation relative par rapport à l'examen initial) met en évidence l'aggravation de la leucopathie au fil des mois.

5 Conclusion et perspectives

Les résultats de quantification obtenus jusqu'à présent sont prometteurs, confortant la vision qualitative des médecins. De plus, une séquence IRM peut être traitée en 5 à 10 minutes sur un unique CPU (8 Go, 2.3 GHz). Un travail en cours est d'améliorer la spécification des lois *a priori* afin de mieux segmenter la substance grise et la substance blanche, et mieux estimer l'évolution temporelle du volume de cette dernière. Il consiste à définir une loi *a priori* locale sous la forme d'un atlas.

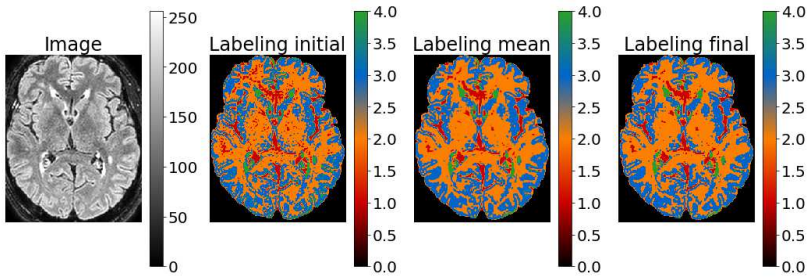


Figure 2: De gauche à droite : Section axiale, label initial, moyenne *a posteriori* du label en chaque site (traité comme un entier), échantillon final du label.

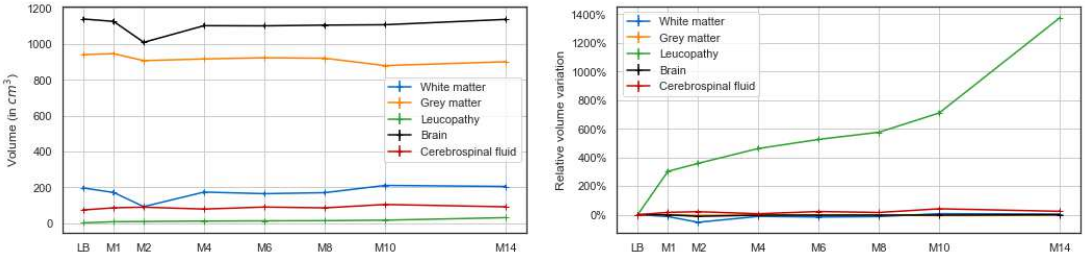


Figure 3: évolution temporelle des volumes pour les différentes zones labelisées. (Gauche) Volumes en cm^3. (Droite) Variation relative des volumes par rapport à l'examen initial.

Bibliographie

Durand, T. *et al.* (2015), EpiBrainRad: an epidemiologic study of the neurotoxicity induced by radiotherapy in high grade glioma patients, *BMC Neurology*, 15, pp. 261.
Gonzalez, J. *et al.* (2011). Parallel Gibbs sampling: From colored fields to thin junction trees. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (pp. 324-332).
Guizard, N. *et al.* (2012), Robust individual template pipeline for longitudinal MR images, *Spatio-temporal Image Analysis for Longitudinal and Time-Series Image Data.*
Moores, M.T. *et al.* (2015). Pre-processing for approximate Bayesian computation in image analysis, *Statistics & Computing*, 25, pp. 23-33.
Van Leemput, K. *et al.* (1999). Automated model-based tissue classification of MR images of the brain, *IEEE Transactions on Medical Imaging*, 18, pp. 897-908.
Weisenfeld, N.I. and Warfield, S.K. (2009). Automatic segmentation of newborn brain MRI. *Neuroimage*, 47, pp. 564-572.

Locally asymptotically efficient test for detecting a threshold effect in the integer-valued AR(1) models

Mohamed Djemaà SADOUN[1,2] & Mohamed BENTARZI[2]

mo-hamedsadoun@outlook.fr *mohamedbentarzi@yahoo.fr*

[1] *Centre for Research in Applied Economics for Development (CREAD), Algiers, Algeria*

[2] *Operational Research Department, University of USTHB, Algiers, Algeria*

Résumé. L'objectif principal de ce travail est de proposer un test efficace capable de détecter l'existence d'un effet de seuil dans un modèle autoregressive à valeurs entières d'ordre 1, basé sur un opérateur d'amincissement généralisé et conduit par une suite de variables aléatoires indépendantes suivant une distribution non spécifiée. Nous commençons d'abord par établir la propriété de normalité asymptotique locale (LAN), tout en exploitant une certaine représentation de l'espérance conditionnelle des transitions des scores due à Drost *et al* (2008) pour le modèle sous-jacent. Ensuite, nous montrons la propriété de linéarité asymptotique locale concernant la suite centrale de ce même modèle sous-jacent. En second lieu, en utilisant ces résultats, nous construisons un test localement asymptotiquement efficace, pour l'hypothèse nulle d'un processus $(GINAR\,(1))$ classique contre l'hypothèse alternative d'un processus $GINAR$ à seuil d'ordre 1 avec deux régimes $(SET-GINAR\,(2,1))$. Les performances du test établi sont montrées via une étude de simulation intensive et une application sur un ensemble de données réelles.

Mots-clés. Processus à valeurs entières de comptage, modèle $SET-GINAR(2,1)$, propriété (LAN), propriété de linéarité asymptotique, test efficace.

Abstract. The main aim in this work is to propose an efficient test able to detect the existence of a threshold effect in a first-order generalized integer-valued autoregressive $(GINAR\,(1))$ model, based on general thinning operator, and driven by a sequence of independent random variables with an unspecified distribution. We establish firstly, while verifying the conditional expectation representation of the transitions scores due to Drost *et al* (2008) to the underlying model, the local asymptotic normality (LAN). Then we show the local asymptotic linearity property concerning the central sequence of our underlying model. Secondly, using these results, we construct an efficient locally asymptotically test, for the null hypothesis of classical $(GINAR\,(1))$ process against an alternative hypothesis of a self-exciting threshold generalized integer-valued autoregressive process of order one with two regimes $(SET-GINAR(2,1))$. The performances of the established test are shown via an intensive simulation study and an application on real data set.

Keywords. Count integer-valued process, $SET-GINAR(2,1)$ model, (LAN) property, asymptotic linearity property, efficient test.

1 Introduction and notations

The self-exciting threshold integer-valued autoregressive $SETINAR(2,1)$ process of first order with two regimes, has been introduced by Monteiro *et al* (2012) to model phenomena with non-negative integer values that evolve over time, including in their structure a large exceeding of the high threshold value appearing in clusters. The distribution of a $SETINAR$ process is mainly described by two parameters: a vector of auto-regression coefficients and a probability distribution on the non-negative integers, called an innovation distribution. This work focuses on the obtaining local asymptotic normality (LAN) property in order to test the presence of a threshold effect in a first-order generalized integer-valued autoregressive ($GINAR(1)$) model. By following the definition in Monteiro *et al* (2012), we can extend, without loss of generality, the $SETINAR(2,1)$ process based on binomial thinning operator to a $SETINAR(2,1)$ process based on the generalized thinning operator, $SET-GINAR(2,1)$, in the following way

$$y_t = (\varphi_1 * y_{t-1} + \varepsilon_{t,1})\,\mathbb{I}_{t-1,1} + (\varphi_2 * y_{t-1} + \varepsilon_{t,2})\,\mathbb{I}_{t-1,2}, \quad t \in \mathbb{Z}, \tag{1.1a}$$

where $\mathbb{I}_{t-1,2} = 1 - \mathbb{I}_{t-1,1}$ which is defined by:

$$\mathbb{I}_{t-1,1} = \begin{cases} 1 & if\ y_{t-1} \leq c, \\ 0 & if\ y_{t-1} > c \end{cases} .$$

The innovation processes, $\{\varepsilon_{t,j}, t \in Z, j = 1,2\}$, are a sequence of independent non-negative integer-valued random variables, with some discrete distribution belonging to the parametric family $\{\mathbb{G}_{\underline{\alpha}_j} | \underline{\alpha}_j = (\alpha_{j,1}, \alpha_{j,2}, \ldots, \alpha_{j,q})' \in A \subset R_+^q\}$, where A is an open, convex subset of $\mathbb{R}_+^q$, and where " $*$ " stands, for the generalized Steutel-Van Harn thinning operator (Latour (1997)), which is defined, for the integer stochastic process y_{t-1} and two counting sequences of independent and identically distributed non-negative integer-valued random variables $\{W_{i,t,j}, i \in \mathbb{N}, t \in \mathbb{Z}, j = 1,2\}$ with finite mean $\varphi_j \in \mathbb{R}_+$ by:

$$\varphi_j * y_{t-1} = \begin{cases} \sum_{i=1}^{y_{t-1}} W_{i,t,j}, & if\ y_{t-1} > 0, \\ 0, & if\ y_{t-1} = 0 \end{cases} , \tag{1.2b}$$

which for each t, $\{W_{i,t,j}\}_{i\in\mathbb{N},t\in\mathbb{Z},j=1,2}$ are independent of y_{t-1}. Moreover, for $j \in \{1,2\}$ being fixed, the innovation process $\{\varepsilon_{t,j}, t \in \mathbb{Z}\}$ is supposed to be independent of y_{t-1} and $\varphi_j \circ y_{t-1}$, and also we have $\{\varepsilon_{t,1}\}$ and $\{\varepsilon_{t,2}\}$ are mutually independent. It is worth to mention that the autoregressive parameters φ_1 and φ_2, the innovation parameters $\underline{\alpha}_j$ are positive and unknown, but the threshold parameter c is assumed to be known and positive. Letting $\underline{\alpha}_j = (\alpha_{j,1}, \alpha_{j,2}, ..., \alpha_{j,q})'$ be the $q-$column vector of the parameters of the innovation law and defining the $q+1-$column vector $\theta_j = \left(\varphi_j; \underline{\alpha}_j'\right)' = \left(\varphi_j, \alpha_{j,1}, \ldots, \alpha_{j,q}\right)' \in (0,1) \times A \subset \mathbb{R}_+ \times \mathbb{R}_+^q$. Let g the point mass function of the discrete distribution $\mathbb{G}$, and denoting $H_g^{(n)}(\underline{\theta})$ a sequence of null hypotheses under which $\left\{y_t^{(n)}, t \in \mathbb{Z}\right\}$ be a sequence of realizations of an integer-valued process satisfying the model

$(1.2a)$, with a $(1+q)\,2-$vector parameters $\underline{\theta} = (\theta_1', \theta_2')'$. Similarly, denoting $H_g^{(n)}\left(\underline{\theta}^{(n)}\right)$ the sequence of alternative hypotheses under which the sequence $\left\{y_t^{(n)}, t \in \mathbb{Z}\right\}$ be a sequence of realizations of a process satisfying the self-exciting threshold generalized integer-valued autoregressive model $(SET - GINAR(2,1))$, with a $(1+q)\,2-$vector parameters $\underline{\theta}^{(n)} = \left(\theta_1^{(n)\prime}, \theta_2^{(n)\prime}\right)'$, where $\varphi_j^{(n)} = \varphi + \dfrac{\lambda_0^{(n)}}{\sqrt{n}} + \dfrac{h_{j,0}^{(n)}}{\sqrt{n}}$ and $\alpha_i^{(n)} = \alpha_i + \dfrac{\lambda_i^{(n)}}{\sqrt{n}} + \dfrac{h_{j,i}^{(n)}}{\sqrt{n}}$, $j = 1, 2$, and $i = 1, ..., q$, with the identification condition $h_{1,l}^{(n)} = -\,h_{2,l}^{(n)}$, $l = 0, 1, ..., q$, or equivalently $\underline{\theta}^{(n)} = \underline{\theta} + \mathbf{K}\, n^{-1/2} \underline{\boldsymbol{\tau}}^{(n)}$, where the $[(1+q)\,2] \times [(1+q)\,2]$ matrix $\mathbf{K}$ is given by

$$\mathbf{K} = \begin{pmatrix} \mathbf{I}_{(1+q)\times(1+q)} & \mathbf{I}_{(1+q)\times(1+q)} \\ \mathbf{I}_{(1+q)\times(1+q)} & -\mathbf{I}_{(1+q)\times(1+q)} \end{pmatrix},$$

where $\mathbf{I}_{q\times q}$ denotes the identity matrix of dimension q. Let $\underline{\boldsymbol{\tau}}^{(n)}$ be the $(1+q)\,2-$vector column which is given by $\underline{\boldsymbol{\tau}}^{(n)} = \left(\underline{\boldsymbol{\lambda}}^{(n)\prime}; \underline{\mathbf{h}}_1^{(n)\prime}\right)'$ where $\underline{\boldsymbol{\lambda}}^{(n)} = \left(\lambda_0^{(n)},\, \lambda_1^{(n)}, ...,\, \lambda_q^{(n)}\right)' \in \mathbb{R}^{1+q}$, $\underline{\mathbf{h}}_1^{(n)} = \left(h_{1,0}^{(n)}, h_{1,1}^{(n)},\, ...,\, h_{1,q}^{(n)}\right)' \in \mathbb{R}^{1+q}$, such that $\sup_n \left(\underline{\boldsymbol{\tau}}^{(n)\prime}\, \underline{\boldsymbol{\tau}}^{(n)}\right) < \infty$. The global local perturbations of the parameters $\varphi, \alpha_1, ..., \alpha_q$ can be decornmposed in to two component types, namely : the local simple perturbations $\lambda_0^{(n)}, \lambda_1^{(n)}, ..., \lambda_q^{(n)}$ and the quantities $h_{j,0}^{(n)}, h_{j,1}^{(n)}, ..., h_{j,q}^{(n)}$, $j = 1, 2$, which can be interpreted as local perturbations due of the regime j of the parameters $\varphi, \alpha_1, ..., \alpha_q$, respectively. Letting $\underline{\nu}^{(n)}$ be the $(1+q)\,2 \times (1+q)\,2$ matrix given by $\underline{\nu}^{(n)} = n^{-1/2}\,\mathbf{K}$, one can easily rewrite the sequence of alteative hypotheses in the form $H_g^{(n)}\left(\underline{\theta} + \underline{\nu}^{(n)} \underline{\boldsymbol{\tau}}^{(n)}\right)$.

The efficient test and estimation procedures turn over to the first work of Stein (1956) in which the author gave a condition so that a model, of a symmetrical position, can be estimated or testing in an adaptive way. We recall that an adaptive test is efficient for a model when the distribution of the errors, f, is only specified partially (continuous and symmetric sense). It is worth noting that the quasi-totality of the existing results, in the literature of the time series analysis concerning the adaptive test, were obtained for real-valued time series models with time-invariant coefficients (Allal and Melhaoui (2006), Benghabrit and Hallin (1998), Bentarzi and Hallin (1996), and many others). In spite of, on one hand, the (asymptotic) optimality of the efficient test under a name of adaptive and, on the other hand, the utility of detecting the existence of a threshold effect in time series modeling.

The present paper contributes to constructing, while following the Le Cam (1960) methodology, an efficient parametric test of the existence of a threshold effect in the $GINAR\,(1)$ process. This Le Cam's methodology allows for more precise and more general results under milder technical assumptions than, for instance, traditional Lagrangian multiplier testing procedures.

2 Local asymptotic normality

Several researchers were interested in derivation of the LAN property for various time series models (see, Koul and Schick (1997), Linton (1993)). To obtain the (LAN) property concerning our $SET-GINAR(2,1)$ process, while following the same framework in Sadoun and Bentarzi (2019). We need the following definitions and notations.
Let $\underline{y}^{(n)}=\left(y_1^{(n)},\ldots,y_n^{(n)}\right)$ be a realization of a finite size n of a stationary integer-valued process $\{y_t,t\in\mathbb{Z}\}$ satisfying $(1.2a)$. Denote by $\Lambda_g^{(n)}\left(\underline{\theta}^{(n)}\right)=\Lambda_g^{(n)}(\underline{\theta}+\underline{\nu}^{(n)}\underline{\boldsymbol{\tau}}^{(n)})$, the logarithm of the likelihood ratio of the calculated conditional likelihood $L_n\left(\underline{\theta}|y_0,\underline{y}^{(n)}\right)$ under $H_g^{(n)}\left(\underline{\theta}\right)$ versus $L_n\left(\underline{\theta}^{(n)}|y_0,\underline{y}^{(n)}\right)$ under $H_g^{(n)}\left(\underline{\theta}^{(n)}\right)$, which is given by:

$$\begin{aligned}\Lambda_g^{(n)}\left(\underline{\theta}+\underline{\nu}^{(n)}\underline{\boldsymbol{\tau}}^{(n)}\right)&=\sum_{t=1}^{n}\log P^{\underline{\theta}^{(n)}}_{(y_{t-1}),y_t}-\sum_{t=1}^{n}\log P^{\underline{\theta}}_{(y_{t-1}),y_t}+o_P\left(1\right),\\&=\sum_{t=1}^{n}\log\left(\sum_{k=0}^{\overline{u}}\left[\left(b_{y_{t-1},\varphi_1^{(n)}}\left(k\right)g_{\underline{\alpha}_1^{(n)}}\left(y_t-k\right)\mathbb{I}_{t-1,1}\right)+\left(b_{y_{t-1},\varphi_2^{(n)}}\left(k\right)g_{\underline{\alpha}_2^{(n)}}\left(y_t-k\right)\mathbb{I}_{t-1,2}\right)\right]\right)\\&-\sum_{t=1}^{n}\log\left(\sum_{k=0}^{\overline{u}}\left[\left(b_{y_{t-1},\varphi_1}\left(k\right)g_{\underline{\alpha}_1}\left(y_t-k\right)\mathbb{I}_{t-1,1}\right)+\left(b_{y_{t-1},\varphi_2}\left(k\right)g_{\underline{\alpha}_2}\left(y_t-k\right)\mathbb{I}_{t-1,2}\right)\right]\right)+o_P\left(1\right),\end{aligned}$$

where $\overline{u}=\min\left(y_{t-1},y_t\right)$ and for $j=1,2$, b_{y_{t-1},φ_j} which stands for the point mass function of the discrete distribution of the $\varphi_j*y_{t-1}|y_{t-1}$ random variable.

Proposition 2.1. *Under some regularity conditions and under* $H_g^{(n)}\left(\underline{\theta}\right)$, *we have*:
i) Taylor expansion in probability of $\Lambda_g^{(n)}\left(\underline{\theta}^{(n)}\right)$

$$\Lambda_g^{(n)}\left(\underline{\theta}+\underline{\nu}^{(n)}\underline{\boldsymbol{\tau}}^{(n)}\right)=\underline{\tau}^{(n)'}\underline{\Delta}^{(n)}\left(\underline{\theta}\right)-\frac{1}{2}\underline{\tau}^{(n)'}\Gamma^{\underline{\Delta}^{(n)}}\left(\underline{\theta}\right)\underline{\tau}^{(n)}+o_P\left(1\right),$$

where the $(1+q)\,2\times(1+q)\,2$ *square matrix* $\Gamma^{\underline{\Delta}^{(n)}}\left(\underline{\theta}\right)$ *is the variance matrix of the score vector (also called central sequence)* $\underline{\Delta}^{(n)}\left(\underline{\theta}\right)$.
ii) Local Asymptotic Normality of the central sequence $\underline{\Delta}^{(n)}\left(\underline{\theta}\right)$

$$\underline{\Delta}^{(n)}\left(\underline{\theta}\right)\xrightarrow{d}N_{(1+q)2}\left(\underline{0},\Gamma^{\underline{\Delta}^{(n)}}\left(\underline{\theta}\right)\right).$$

where $\Lambda_g^{(n)}\left(\underline{\theta}+\underline{\nu}^{(n)}\underline{\boldsymbol{\tau}}^{(n)}\right)=\log\left(dP^{(n)}_{\underline{\theta}+\underline{\nu}^{(n)}\underline{\boldsymbol{\tau}}^{(n)}}\Big/dP^{(n)}_{\underline{\theta}}\right)$ *which represent the Radon-Nikodym derivative.is nonsingular and where* $\Lambda_g^{(n)}(\underline{\theta}+\frac{1}{\sqrt{n}}\underline{\tau}^{(n)})=\log\left(dP^{(n)}_{\underline{\theta}+\underline{\nu}^{(n)}\underline{\boldsymbol{\tau}}^{(n)}}\Big/dP^{(n)}_{\underline{\theta}}\right)$ *which represent the Radon-Nikodym derivative.*

3 Local asymptotic efficient test

Among the different consequence of the (LAN) property, we have the following result:

$$\begin{aligned}&\underline{\Delta}^{(n)}\left(\underline{\theta}\right)\Rightarrow N_{(1+q)2}\left(\underline{0},\Gamma^{\underline{\Delta}^{(n)}}\left(\underline{\theta}\right)\right)\text{ under }H_g^{(n)}\left(\underline{\theta}\right),\\&\underline{\Delta}^{(n)}\left(\underline{\theta}\right)\Rightarrow N_{(1+q)2}\left(\Gamma^{\underline{\Delta}^{(n)}}\left(\underline{\theta}\right)\underline{\tau}^{(n)},\Gamma^{\underline{\Delta}^{(n)}}\left(\underline{\theta}\right)\right)\text{ under }H_g^{(n)}\left(\underline{\theta}+\underline{\nu}^{(n)}\underline{\tau}^{(n)}\right).\end{aligned}$$

Let us note $\eta = \Gamma^{\Delta}(\underline{\theta}) \begin{pmatrix} \underline{\boldsymbol{\lambda}} \\ \underline{\mathbf{h}} \end{pmatrix}$, with $\underline{\boldsymbol{\lambda}}^{(n)} \to \underline{\boldsymbol{\lambda}}$ and $\underline{\mathbf{h}}^{(n)} \to \underline{\mathbf{h}}$ as $n \to \infty$, then the testing problem of the null hypothesis $H_g^{(n)}(\underline{\theta})$ versus the local alternative $H_g^{(n)}\left(\underline{\theta} + \underline{\nu}^{(n)}\underline{\boldsymbol{\tau}}^{(n)}\right)$, i.,e., testing a time-invariant $GINAR(1)$ model, against a $SET - GINAR(2,1)$ model given by $(1.2a)$, becomes, quite simply, a testing problem tied to the experiment of Gaussian position. More precisely : testing the null hypothesis

$$H_{0,g} : N\left(\eta_0, \Gamma^{\Delta}(\underline{\theta})\right), \quad \left(\eta_0 = \Gamma^{\Delta}(\underline{\theta}) \begin{pmatrix} \underline{\boldsymbol{\lambda}} \\ \underline{\mathbf{0}} \end{pmatrix}\right),$$

versus the alternative one

$$H_{1,g} : N\left(\eta, \Gamma^{\Delta}(\underline{\theta})\right), \quad \left(\eta = \Gamma^{\Delta}(\underline{\theta}) \begin{pmatrix} \underline{\boldsymbol{\lambda}} \\ \underline{\mathbf{h}} \end{pmatrix}, \underline{\mathbf{h}} \neq 0\right).$$

The following proposition establish a locally asymptotically optimal test (so called most stringent test) to test $H_{0,g}$ versus $H_{1,g}$.

Proposition 3.1. *Under the condition* $(A.1) - (A.6)$, *the test rejecting the null hypothesis* $H_g^{(n)}(\underline{\theta})$ *whenever:*

$$\widehat{Q}_g^{(n)}\left(\widehat{\underline{\theta}}^{(n)}\right) = \widehat{\underline{\Delta}}^{(n)\prime}\left(\widehat{\underline{\theta}}^{(n)}\right)\left(\Gamma^{\underline{\Delta}^{(n)}}\left(\widehat{\underline{\theta}}^{(n)}\right)\right)^{-1}\widehat{\underline{\Delta}}^{(n)}\left(\widehat{\underline{\theta}}^{(n)}\right) > \mathcal{X}^2_{(1+q)2,1-\alpha},$$

is such that :

(i) *has asymptotic level* α *under* $H_g^{(n)}(\underline{\theta})$,

(ii) *has asymptotic power:*

$$1 - \mathcal{F}\left(\mathcal{X}^2_{1-\alpha}; (1+q)\,2; \underline{h}'\Gamma^{\underline{\Delta}^{(n)}}\left(\widehat{\underline{\theta}}^{(n)}\right)\underline{h}\right), \text{ under } H_g^{(n)}\left(\underline{\theta} + \underline{\nu}^{(n)}\underline{\boldsymbol{\tau}}^{(n)}\right),$$

where $\mathcal{F}\left(\mathcal{X}^2_{1-\alpha}; r; v\right)$ *denotes the non central chi-square distribution function with* r *degrees of freedom and non-centrality parameter* v;

(iii) *is locally asymptotically most stringent test against* $H_g^{(n)}\left(\underline{\theta} + \underline{\nu}^{(n)}\underline{\boldsymbol{\tau}}^{(n)}\right)$.

4 Numerical illustration

The performance of the constructed efficient test is shown by simulation studies. Two $SET-GINAR(2,1)$ data-generator processes, $M1-M2$, and $GINAR(1)$ process $M3$ are used to simulate time series of small, moderate and relatively large sizes $(n = 100 - 800)$, based on two different thinning operators, namely: the binomial thinning operator, and the negative binomial thinning operator (Ristić (2009)). The sets of parameter values are chosen such that the underlying models are strictly stationary. Where the stationary condition applies here is the sufficient condition: $\varphi_1 < 1, \varphi_2 < 1$ and $\varphi_1 + \varphi_2 < 1$. The model $M3$ (non threshold $GINAR(1)$) is used with the aim of calculating the empirical levels of the obtained test. For each data-generating process, we consider 1000 Monte Carlo replications and report frequencies of the case where the threshold effect is (correctly for $M1 - M2$ and erroneously for $M3$) identified. The true parameter values of these models are given as follows :

$$\text{Model } M1 : \underline{\theta} = [(\varphi_1, \varphi_2; \alpha_1, \alpha_2; c)]' = [(0.3, 0.7; 6, 2; 13)]',$$

Model $M2 : \underline{\theta} = [(\varphi_1, \varphi_2; \alpha_1, \alpha_2; c)]' = [(0.2, 0.1; 3, 3.5; 7)]'$,
Model $M3\ y_t = 0.8 \circ y_{t-1} + \varepsilon_t$ with $\varepsilon_t \rightsquigarrow \mathcal{P}(4)$,

We stress that for all models which were considered here, $\varepsilon_{t,1} \rightsquigarrow \mathcal{P}(\alpha_1)$ and $\varepsilon_{t,2} \rightsquigarrow \mathcal{P}(\alpha_2)$. We have reported in the Table 1, brief results to illustrate our theoretical results, where $\varphi\circ$ and $\varphi*$ denotes the binomial and the negative binomial thinning operators, respectively.

Table 1
Empirical powers of the efficient tests ϕ *for the level* 5%

n		100	150	200	300	400	600	800
		ϕ	ϕ	ϕ	ϕ	ϕ	ϕ	ϕ
$M1$	$\varphi\circ$	.9150	.9580	.9880	.9993	1	1	1
	$\varphi*$	.9460	.9760	.9974	1	1	1	1
$M2$	$\varphi\circ$	.8250	.9305	.9760	.9911	.9940	.9988	.9992
	$\varphi*$	.8522	.9389	.9515	.9772	.9955	.9993	.9999

Bibliographie

Bentarzi, M. and Hallin, M. (1996). Locally Optimal Tests Against Periodic Autoregression. *Econometric Theory* 12 pp. 88 − 112.

Benghabrit, Y. and Hallin, M. (1998). Locally asymptotically optimal Tests for $AR(p)$ against diagonal bilinear dependence. *Journal of Statistical Planning and Inference* 68, pp. 47 − 63.

Koul, H.L. and Schick, A. (1997). Efficient estimation in non-linear autoregressive time series models. *Bernoulli.* 3, pp. 247 − 277.

Latour, A. (1997). The Multivariate $GINAR(p)$ Process. *Adv. Appl. Prob.* 29, pp. 228 − 248.

Le Cam, L. (1986). Asymptotic Methods in Statistical Decision. *Theory. New York: Springer-Verlag.*

Linton, O. (1993). Adaptive Estimations in $ARCH$ Models. *Econometric Theory.* 9(4) pp. 539 − 569.

Monteiro, M. Scotto, M.G. and Pereira, I. (2012). Integer-Valued Self-Exciting Threshold Autoregressive process. *Journal of Communication in statistics-Theory and Methods* 41, pp. 2717 − 2737.

Ristić, M.M, Bakouchb, H. and Nastića, A. (2009). A newgeometricfirst-orderinteger-valuedautoregressive ($NGINAR(1)$) process. *Journal of Statistical Planning and Inference.* 139, pp. 2218 − 2226.

Sadoun, M. and Bentarzi, M. (2019). Efficient estimation in periodic $INAR(1)$ model :parametric case. *Communications in Statistics-Simulation and Computation*, pp. 1 − 21.

Stein, C. (1956). Efficient nonparametric testing and estimation. *In: Proceeding of the Third Berkeley Sympsium on Mathematical Statistic and Probability.* Berkley, CA: University of California press, Vol. 1, pp.187 − 196.

Trend detection in extremes: pointwise and spatial approaches by Peaks-Over-Thresholds. Application to extreme temperature and precipitation in Burkina Faso

Béwentaoré Sawadogo[1,2] & Liliane Bel[1] & Diakarya Barro[3]

[1] *Université Paris Saclay, INRAe, AgroParisTech, UMR MIA-Paris, 75005, Paris, France, liliane.bel@agroparistech.fr*

[2] *LANIBIO, UJKZ, UFR-SEA, BP: 7021, Ouagadougou 03, Burkina Faso, sbewentaore@yahoo.fr*

[3] *UFR-SEG, Université Ouaga II, 12 BP: 417 Ouagadougou 12, Burkina Faso, dbarro2@gmail.com*

Abstract. Modelling extremes of climate variables in the framework of climate change is a particularly difficult task, since it implies taking into account spatio-temporal non-stationarities. The first issue addressed in this study is the temporal evolution of extremes handled pointwise by the approach of threshold exceedances. The trend in parameters of the distribution of excesses and the intensity of the occurrences of extreme events for several meteorological variables are described as linear or quadratic functions of time. Then the spatial dimension is taken into account using generalized ℓ-Pareto processes in a non-stationary framework. This allows us to produce both return levels map and their confidence intervals from the extrapolation of identified trends. Finally, our approach is applied to temperature and rainfall data in Sub-Saharan Africa, particularly in Burkina Faso.

Keywords: Non-stationary POT model, Generalized ℓ-Pareto process, Space-time Extremes, Climate Change.

Résumé. La modélisation des extrêmes des variables climatiques dans le cadre du changement climatique est une tâche particulièrement difficile, car elle implique la prise en compte des non-stationnarités spatio-temporelles. La première question abordée dans cette étude est l'évolution temporelle des extrêmes traitée ponctuellement par l'approche des dépassements de seuils. L'évolution des paramètres de la distribution des dépassements et de l'intensité des occurrences d'événements extrêmes pour plusieurs variables météorologiques sont décrites comme des fonctions linéaires ou quadratiques du temps. Ensuite, la dimension spatiale est intégrée à l'aide de processus ℓ-Pareto généralisés dans un cadre non stationnaire. Cela nous permet de produire une carte des niveaux de retour avec leurs intervalles de confiance à partir de l'extrapolation des tendances identifiées. Enfin, notre approche est appliquée aux données de température et de précipitations en Afrique subsaharienne, en particulier au Burkina Faso.

Mots-clés: Modèle POT non stationnaire, Processus ℓ-Pareto généralisé, Extrêmes Spatio-temporels, Changement climatique.

1 Introduction

Modelling extremes of climate variables in the context of climate change is a particularly difficult task, as it implies taking into account spatio-temporal non-stationarities. According to the latest report of the Intergovernmental Panel on Climate Change (IPCC, 2018, 2019), occurrences of climate extremes will increase on the horizon 2050. Extreme Value Theory provides a rigorous mathematical tool for modelling and detecting recent trends in the extremes of climate variables and extrapolating these events beyond observed data in a spatial and temporal framework.

The spatial POT approach was introduced in a stationary framework by Ferreira and de Haan[7], and generalized by Dombry and Ribatet[3] who defined the family of generalized ℓ-Pareto process. Neglecting the non-stationarity of extremes would not only provide a poor description of the data but also have dramatic consequences on the estimation of return levels. Although it is relatively simple to construct non-stationary models in the uni-variate framework, for example by letting the parameters of the marginals depend on time and other covariates ([4], [8], [9] and [10]), it is more difficult to model spatial and temporal trends in the dependence structure. In addition, even if a parametric family can be identified, it can be very difficult to make an inference if the data set is not spatially rich enough to be used for the inference. Our approach takes into account non-stationarity in the parameters of the marginal distributions and models the spatial dependence using ℓ-Pareto processes, following the inference procedure deveopped by de Fondeville et al. in a sationnary framework for the Brow-Resnick process [1], [2].

2 Methodology and data

2.1 Data set

This study uses time series of daily temperature and precipitation measurements from 1957 to 2016 provided by ten synoptic stations extracted from the Burkina Faso climatological database. These stations have been selected to ensure good spatial uniformity and representativeness of different climatic regimes and data quality.

2.2 Non-stationary ℓ-Pareto process

Let S be a compact subset of $\mathbb{R}^m$ and T be a compact subset of $\mathbb{R}^+$ denoting respectively the space and time domains, $\mathcal{C}(S \times T)$ is the set of continuous real functions on $S \times T$ provided with the uniform norm $\| \, . \, \|_\infty$ and $\mathcal{C}_+(S \times T)$ is the restriction of $\mathcal{C}(S \times T)$ to non-negative functions. We designate by $\mathcal{C}_0(S \times T)$ the subset of $\mathcal{C}_+(S \times T)$ containing only non-negative functions, i.e. $\mathcal{C}_0(S \times T) = \mathcal{C}_+(S \times T) \setminus \{0\}$, thus avoiding the appearance of degenerated limit probability measures.

Let $Z = \{Z(s,t)\}_{s\in S, t\in T}$ be a space-time stochastic process defined on $\mathcal{C}_+(S\times T)$. The peak-over threshold model has been extended to continuous processes ([2], [3], [5], [7]) as part of regular functional variation. A stochastic process Z is regularly varying, if there existe a sequence : $a_n : \mathcal{C}_+(S\times T) \to \mathbb{R}_+$ of continuous functions such that

$$nP\left(a_n^{-1}(s,t)Z(s,t) \in A\right) \longrightarrow \Lambda(A), \quad n \longrightarrow +\infty, \tag{1}$$

we note $Z \in RV\{\mathcal{C}_0(S\times T), a_n(s,t), \Lambda\}$, where Λ is a measure on $\mathcal{C}_0(S\times T)$ such that $\Lambda(aA) = a^{-1}\Lambda(A)$ for any $a > 0$ and any borelian $A \in \mathcal{B}(\mathcal{C}_0(S\times T))$. In the multivariate and spatial framework, a threshold exceedance is defined by Dombry and Ribatet[3] as an event

$$\{\ell(Z(s,t)) > u_n\} \quad \text{such that } \lim_{n\to+\infty} P(\ell(Z(s,t)) > u_n) \longrightarrow 0,$$

where u_n is a sequence of thresholds and ℓ is a continuous non-negative, homogeneous function, i.e. $\ell(aZ(s,t)) = a\ell(Z(s,t)), a > 0$. The risk function ℓ determines the type of extreme events we are interested in. For example, such a function could be the maximum, minimum, average, or value at a specific point $(s_0,t_0) \in S\times T$.

For given a ℓ risk function, the ℓ-excess of Z^* obtained by transforming Z into standard Pareto margins converge weakly in $\mathcal{C}_+(S\times T)$ to a process $W_\ell^*(s,t)$ called standard ℓ-Pareto process. The generalized form of ℓ-Pareto processes is given in [7] by:

$$W_\ell(s,t) = \begin{cases} \mu(s,t) + \sigma(s,t)\left\{W_\ell^*(s,t)^{\gamma(s,t)} - 1\right\}/\gamma(s,t), & \gamma(s,t) \neq 0 \\ \mu(s,t) + \sigma(s,t)\log W_\ell^*(s,t), & \gamma(s,t) = 0 \end{cases}, \tag{2}$$

where $\sigma(s,t) > 0$, $\mu(s,t)$ and $\gamma(s,t)$ are continuous functions taken from $\mathcal{C}_+(S\times T)$. A pseudo-polar decomposition of W_ℓ in (2) leads to the following formulation ([1], [3])

$$W_\ell = R\frac{Q}{\ell(Q)} \tag{3}$$

where R is a unit Pareto random variable of index γ_R representing the intensity of process, and Q is stochastic process denoted the angular component with state space S and taking values in $\mathscr{S}_\ell = \{f \in \mathcal{C}_0(S) : \ell(f) \geq 1, \| f \|_1 = 1\}$.

In order to make inferences and model the dependence structure of ℓ-Pareto processes, we choose to model its angular component in (3) taking a Brown-Resnick model, in this case we are able to estimate the dependence parameters from the density λ of the Λ measure. To better capture the possible dependence structure, we use anisotropic semi-variogram models in the definition of the Brow-Resnick process and estimate the different parameters using the gradient scoring rule method[1].

3 Main Results

In this study, we consider that daily temperature and precipitation were generated by a stochastic process $Z = \{Z(s,t)\}_{s\in S, t\in T}$ taking its values in the space of continuous functions $\mathcal{C}_+(S \times T)$, where $S \times T$ designates a subset of $\mathbb{R}^2 \times \mathbb{R}^+$. First we studied the local behavior of the tail of the daily precipitation distribution by fitting a non-stationary generalized Pareto distribution for each station $s \in S$, i.e. :

$$P\left(Z(s,t) - u(s,t) > z \mid Z(s,t) > u(s,t)\right) = \left(1 + \gamma_u(s,t)\frac{z}{\sigma_u(s,t)}\right)_+^{-\frac{1}{\gamma_u(s,t)}}, \quad x \geq 0$$

Trends are identified by incorporating non-stationarity into the parameters of the distribution of excesses and the intensity of the occurrences of extreme events for several meteorological variables are described as linear or quadratic functions of time, i.e:

$$\left.\begin{aligned} \sigma_u(s,t) &= \exp\left(\sigma_0 + \sigma_1 t + \sigma_2 t^2\right) \\ \lambda(s,t) &= \alpha_0 + \alpha_1 t + \alpha_2 t^2 \\ \gamma_u(s,t) &= \gamma_0 \end{aligned}\right\} s \in \mathcal{S} \subset \mathbb{R}^2.$$

The shape parameter and threshold are kept constant so as not to increase the uncertainty in the estimation of parameters and extreme quantiles. In the pointwise POT approach the results of the best trends (see figure 1) adjusting the data for climates variables are obtained using the maximum likelihood ratio test at risk 5%. The optimal trends

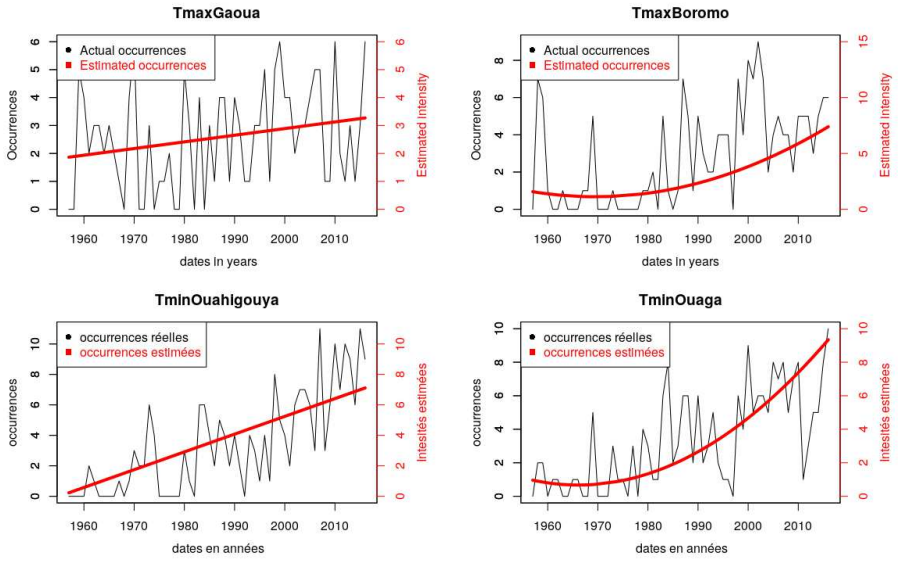


Figure 1: Optimal trends in Tmax and Tmin occurrences.

identified for minimum and maximum temperatures show an increasing frequency of heat

waves in recent years. This allowed us to produce return levels maps by kriging and their confidence intervals for several climate variables from the extrapolation of identified trends.

The inference of the results of standardized ℓ-Pareto with $\ell(Z) = \max_{s \in S} Z(s)$ gives an anisotropic variogram (figure 2a) which an advanced analysis shows that extreme rainfall is spatially correlated in the East-West direction. There is no spatial dependence in the other directions.

The extremogram $\pi(h) = P\left(Z(s+h) > u \mid Z(s) > u\right)$ (figure 2b) shows an extremal dependence up to 200km that stabilizes to 0.25. The fitted model represented by the green curve in figure 2b follows the cloud of estimated conditional exceedance probabilities fairly well and captures the general trend.

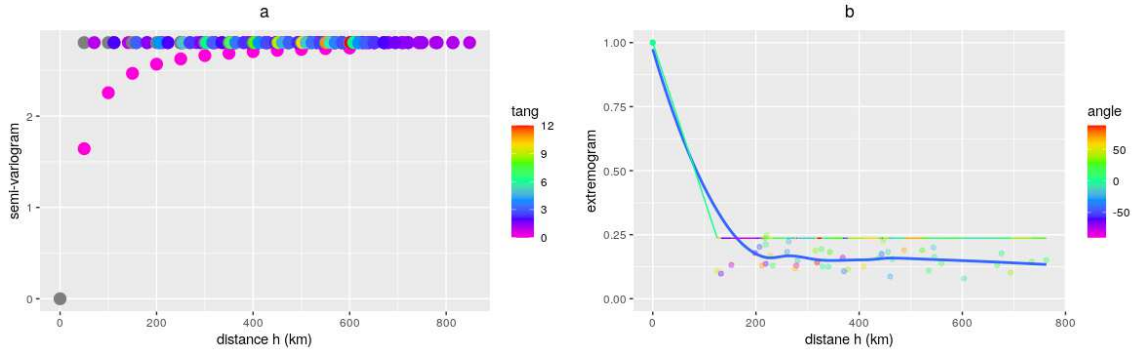


Figure 2: Anisotropic variogram estimated (figure 2a) and extremogram $\pi(h)$ (figure 2b) for risk functional $\ell(Z) = \max_{s \in S} Z(s)$ of distance between locations s and $s+h$.

4 Conclusion

In this study we modelled recent temperature and precipitation trends in Burkina Faso using non-stationary ℓ-Pareto processes. Over the period from 1957 to 2016, we note an increasing trend in the intensity of extreme temperature occurrences. Extremes precipitation are spatially correlated in the East-West direction.

References

[1] de Fondeville, R. and Davison, A. C. (2018). High-dimensional peaks-over-threshold inference *Biometrika*, 105(3):575-592.

[2] de Fondeville, R., Davison, A.C., 2020. Functional peaks-over-threshold analysis. arXiv preprint arXiv:2002.02711

[3] Dombry, C. et Ribatet, M. (2015). Functional regular variations, Pareto processes and peaks over threshold. *Stat. Interface*, 8(1):9–17.

[4] Economou, T. and David, B. (2014). Spatio–temporal Modelling of Extreme Storms. *Annals of Applied Statisitcs*, 8(4):2223–2246.

[5] Engelke, Sebastian, Raphaël De Fondeville, and Marco Oesting. *Extremal Behaviour of Aggregated Data with an Application to Downscaling*. Biometrika 106.1 (2019): 127-44.

[6] Engelke , S., M Alinowski, A., K Abluchko , Z., S Schlather, M. (2015). Estimation of Hüsler–Reiss distributions and Brown–Resnick processes. J. R. Statist. Soc. B 77, 239–65.

[7] Ferreira, A. et de Haan, L. (2014). The generalized Pareto process ; with a view towards application and simulation. *Bernoulli*, 20(4):1717–1737.

[8] Huser, R., Genton, M.G. Non-Stationary Dependence Structures for Spatial Extremes. *JABES* 21, 470–491 (2016). https://doi.org/10.1007/s13253-016-0247-4.

[9] Ferreira, A., Friederichs, P., de Haan, L., Neves, C., and Schlather, M. (2017). *Estimating space-time trend and dependence of heavy rainfall.* arXiv preprint arXiv:1707.04434.

[10] Parey S., Thu Huong Hoang T. , Dacunha-Castelle D., (2010b), Different ways to compute temperature return levels in the climate change context, *Environmetrics* 2010; 21: 698–718, DOI: 10.1002/env.1060.

Reconstruction de la connectivité fonctionnelle en Neurosciences: une amélioration des algorithmes actuels

Gilles Scarella [1] & Cyrille Mascart [2] & Alexandre Muzy [3] & Tien Cuong Phi [4] & Patricia Reynaud-Bouret [5]

[1] *Université Côte d'Azur, CNRS, LJAD/I3S - gilles.scarella@univ-cotedazur.fr*
[2] *Université Côte d'Azur, CNRS, I3S - mascart@i3s.unice.fr*
[3] *Université Côte d'Azur, CNRS, I3S - alexandre.muzy@cnrs.fr*
[4] *Université Côte d'Azur, CNRS, LJAD - Tien_cuong.phi@univ-cotedazur.fr*
[5] *Université Côte d'Azur, CNRS, LJAD - reynaudb@univ-cotedazur.fr*

Résumé. Afin d'identifier la connectivité fonctionnelle entre neurones, des travaux précédents (dans [2] notamment) ont utilisé des processus de Hawkes pour modéliser les intensités conditionnelles des trains de spikes et ont reconstruit la connectivité fonctionnelle par moindres carrés pénalisés. On propose ici une nouvelle méthode de construction des matrices du même problème Lasso obtenu et l'utilisation d'un autre solveur, qui semble plus efficace, pour une amélioration du temps de calcul.

Mots-clés. Neurosciences, processus de Hawkes, Lasso, connectivité fonctionnelle

Abstract. To identify the functional connectivity between neurons, previous works (see [2] in particular) have used Hawkes processes to model conditional intensities of spike trains and have reconstructed functional connectivity by penalized least square method. We propose here a new method to construct the matrices in the same resulting Lasso problem and the choice of another solver, which seems to be more efficient, to improve computational times.

Keywords. Neuroscience, Hawkes processes, Lasso, functional connectivity

1 Présentation du problème

Le but est d'étudier la connectivité fonctionnelle entre neurones, qui est un enjeu important en Neurosciences. La présente étude est basée sur l'enregistrement simultané des temps d'émission des potentiels d'action, ou spikes, de M neurones. Comme cela a été présenté dans [2], on considère M trains de spikes simultanés, modélisés comme des processus de Hawkes multivariés. L'intensité du i-ième train de spikes N^i a la forme suivante

$$\lambda_i(t) = \left(\nu_i + \sum_{j=1}^{M} \sum_{T \in N^j, T<t} h_{j \to i}(t-T) \right)_+, \quad \forall\, t, \quad \forall i \in [\![1, M]\!].$$

Le coefficient ν_i est le taux de décharge spontané du i-ième train de spikes, donnant la fréquence moyenne d'apparition d'un nouveau spike en l'absence d'excitation ou d'inhibition, et la fonction $h_{j\to i}$, qui dépend du temps, modélise l'interaction excitatrice ou inhibitrice du j-ième train sur le i-ième. On suppose que les fonctions $h_{j\to i}$ sont constantes par morceaux sur une partition de K intervalles de taille δ, telles que:

$$h_{j\to i} = \sum_{k=1}^{K} a_{j\to i}^{k}\, \varphi_k \quad \text{où } \varphi_k = \mathbb{1}_{((k-1)\delta, k\delta]}.$$

Les coefficients (ν_i) et $(a_{j\to i}^{k})$ doivent être estimés, pour tous $(i,\, j) \in [\![1, M]\!]^2$ et $k \in [\![1, K]\!]$.

Le code *neuro-stat*, introduit dans [2], permet de retrouver une estimation *sparse* des coefficients (ν_i) et $(a_{j\to i}^{k})$ et donc du graphe de connectivité fonctionnelle (où l'existence d'une connexion entre j et i correspond à la non nullité de $h_{j\to i}$). De plus, le code permet aussi, en fonction de différentes phases de l'enregistrement, d'estimer différents jeux de paramètres et différents graphes (typiquement on peut alors associer un graphe à un comportement animal).

Ici nous nous intéressons à l'amélioration de l'algorithme sur une plage de temps fixée $(T_{\min}, T^{\max}]$.
De manière générale, on se focalise sur le critère des moindres carrés suivant, que nous pénaliserons ensuite:

$$\int_{T_{\min}}^{T^{\max}} \overline{\lambda_i}^2(t)\, dt - 2 \int_{T_{\min}}^{T^{\max}} \overline{\lambda_i}(t)\, dN_t^i$$

où $\overline{\lambda}_i$ est un candidat intensité de la forme

$$\begin{aligned} \overline{\lambda_i}(t) &= \overline{\nu_i} + \sum_{j=1}^{M} \sum_{T\in N^j, T<t} \overline{h}_{j\to i}(t-T), \quad \forall\, t, \quad \forall i \in [\![1, M]\!], \\ \text{avec } \overline{h}_{j\to i} &= \sum_{k=1}^{K} \overline{a}_{j\to i}^{k}\, \varphi_k \end{aligned}$$

Soit $\psi_t^l(\varphi_k)$ la fonction prévisible définie par:

$$\psi_t^l(\varphi_k) = \int_{-\infty}^{t^-} \varphi_k(t-u)\, dN_u^l = \sum_{T<t, T\in N^l} \mathbb{1}_{((k-1)\delta, k\delta]}(t-T), \tag{1}$$

on en déduit de (1) que $\overline{\lambda}_i$ s'écrit donc sur le dictionnaire des fonctions prévisibles sous la forme suivante

$$\overline{\lambda}_i(t) = \overline{\nu}_i + \sum_{j=1}^{M} \sum_{k=1}^{K} \overline{a}_{j\to i}^{k}\, \psi_t^j(\varphi_k)$$

En introduisant comme dans [2] les matrices b et G, de dimension $(1+MK)$-by-M (resp. $(1+MK)$-by-$(1+MK)$), et en notant $\alpha(l,k) = 1+(l-1)\,K+k$, l'indice dans $[\![2, 1+M\,K]\!]$, pour $k \in [\![1,K]\!]$ et $l \in [\![1,M]\!]$, on voit que le critère des moindres carrés devient

$$-2\,{}^T\mathbf{b}^i\,\beta + {}^T\beta\,\mathbf{G}\,\beta \quad \forall i \text{ avec}$$

$$\begin{aligned}
\mathbf{b}^i_{\alpha(l,k)} &= \int_{T_{\min}}^{T^{\max}} \psi_t^l(\varphi_k)\,dN_t^i \quad \text{et} \qquad \mathbf{b}^i_1 = \#\left\{T \in N^i,\ \ T \in (T_{\min}, T^{\max}]\right\},\\
\mathbf{G}_{\alpha(l_1,k_1),\alpha(l_2,k_2)} &= \int_{T_{\min}}^{T^{\max}} \psi_t^{l_1}(\varphi_{k_1})\,\psi_t^{l_2}(\varphi_{k_2})dt, \quad \mathbf{G}_{\alpha(l,k),1} = \int_{T_{\min}}^{T^{\max}} \psi_t^{l}(\varphi_{k})\,dt \text{ et } \mathbf{G}_{1,1} = T^{\max} - T_{\min}
\end{aligned}$$

En suivant [1], on pénalise par une norme l_1 à poids tels que $\mathbf{d}$ est une matrice de dimension $(1+MK)$-by-M définie à partir de $\mu_{\mathbf{2}}$ (de même dimension) et $\mu_{\mathbf{A}}$ vecteur de taille $(1+MK)$.

On utilise la même définition que [2] pour $\mathbf{d}$, dans laquelle on prend $\gamma = 3$.

$$\begin{aligned}
\mathbf{d}^i &= \sqrt{2\gamma\, c_{log}\, \mu_{\mathbf{2}i}} + \frac{\gamma}{3} c_{log}\,\mu_{\mathbf{A}},\ \forall i \in [\![1, 1+MK]\!],\ \ c_{log} = \log((1+MK)M)\\
(\mu_{\mathbf{A}})_{\alpha(l,k)} &= \sup_{t\in(T_{\min},T^{\max}]} |\psi_t^l(\varphi_k)|,\ \ (\mu_{\mathbf{A}})_1 = 1,\\
(\mu_{\mathbf{2}})^i_{\alpha(l,k)} &= \int_{T_{\min}}^{T^{\max}} \left(\psi_t^l(\varphi_k)\right)^2\, dN_t^i,\ \ (\mu_{\mathbf{2}})^i_1 = \#\left\{T \in N^i,\ \ T \in (T_{\min}, T^{\max}]\right\}.
\end{aligned}$$

On suit la même procédure que dans [2] et l'on doit résoudre un ensemble de problèmes Lasso de dimension $(1+M\,K)$, pour tout $i \in [\![1,M]\!]$, pour obtenir les coefficients $a^k_{j\to i}$

$$\begin{aligned}
\mathbf{a}^i_{BL} = \arg \min_{\beta\in\mathbb{R}^{(1+MK)}} & -2\,{}^T\mathbf{b}^i\,\beta + {}^T\beta\,\mathbf{G}\,\beta + 2\,{}^T\mathbf{d}^i\,|\beta|\\
\text{avec} \quad & \mathbf{a}^i_{BL} = (\nu_i, a^1_{1\to i}, a^2_{1\to i}, \cdots, a^K_{1\to i}, a^1_{2\to i}, \cdots, a^K_{M\to i})\\
\text{et} \quad & |\beta| = (|\beta_1|, |\beta_2|, \cdots, |\beta_{MK}|, |\beta_{1+MK}|)
\end{aligned}$$

2 Nouvelle méthode et résultats

La méthode mise en œuvre dans [2], pour le package *neuro-stat*, est coûteuse en temps calcul et utilisable uniquement dans le logiciel R, ce qui conduit à certaines limitations. En effet, cette méthode utilisait des calculs d'intégrales de fonctions constantes par morceaux pour définir les matrices intervenant dans le problème d'estimation. Si $Ntot$ désigne le nombre total de spikes, la complexité était alors en $O(M\,Ntot\,K^2)$ pour $\mathbf{G}$ et en $O(M\,Ntot\,K)$ pour $(\mathbf{b}, \mu_{\mathbf{A}}, \mu_{\mathbf{2}})$.

L'objectif est de considérer ici entre 1 000 et 10 000 neurones, soit un nombre supérieur à celui considéré dans [2]. La référence est le *BlueBrain project*[1] qui simule des colonnes corticales d'environ 10 000 neurones.

[1] `https://www.epfl.ch/research/domains/bluebrain/`

La nouvelle méthode présentée ici utilise, comme la précédente, des fonctions à support borné mais est moins coûteuse: on utilise la portée $A = K\delta$ de telle sorte que l'influence du spike τ_1 est nulle sur le spike τ_2 si $|\tau_1 - \tau_2| > A$. Par exemple, on obtient pour la partie intérieure de $\mathbf{G}$,

$$\mathbf{G}_{\alpha(l_1,k_1),\alpha(l_2,k_2)} = \sum_{\substack{\tau \in N^{l_1},\, \theta \in N^{l_2}, \\ (T_{\min}-A) \leqslant \tau, \theta < T^{\max}}} \int_{T_{\min}}^{T^{\max}} \mathbb{1}_{(\tau+(k_1-1)\delta, \tau+k_1\delta] \cap (\theta+(k_2-1)\delta, \theta+k_2\delta] \cap (T_{\min}, T^{\max}]}(t)\, dt$$

Ce terme peut être calculé explicitement et son calcul est décrit dans l'Algorithme 1, où l'on passe en revue chacun des spikes et l'on affecte sa contribution au bon coefficient de $\mathbf{G}$. Dans l'Algorithme 1, T est un tableau de taille $Ntot$, contenant les valeurs des spikes indépendamment des neurones dans l'ordre croissant, *neur* est un tableau de même taille que T contenant le numéro de neurone du spike correspondant.

Algorithm 1 Calcul de G

Usage: compute_G(M, K, $T_{\min}$, $T^{\max}$, T, neur, δ)

```
G ← matrix(0, nrow=(1+M*K), ncol= (1+M*K))                    # initialization of G
G[1,1] ← T^max − T_min
A ← K ∗ δ                                                      # scope
α ← get_low_index((T_min − A), T)          # lowest index in T such that T[α] > T_min − A
β ← get_low_index(T^max, T) - 1            # highest index in T such that T[β] ⩽ T^max
for (i in α:β) do
    ti ← T[i]; l_1 ← neur[i]
    for (j in (low[i]:(i-1))) do           # low[i] is the lowest index s.t T[i] − T[low[i]] ⩽ A
        tj ← T[j]; l_2 ← neur[j]
        for (k_1 in (1:K)) do
            for (k_2 in (1:K)) do
                x_1 ← min(T^max, ti + k_1 δ, tj + k_2 δ)
                x_2 ← max(T_min, ti + (k_1 − 1) δ), (tj + (k_2 − 1) δ)
                dx = x_1 − x_2
                if (dx > 0) then
                    G[(l_1-1)*K+k_1+1,(l_2-1)*K+k_2+1] += dx
                    G[(l_2-1)*K+k_2+1,(l_1-1)*K+k_1+1] += dx
    ...
```

Les Figures 1 et 2 montrent une comparaison des temps de construction de $\mathbf{G}$ entre *neuro-stat* et la nouvelle méthode, sur deux exemples, le premier à $Ntot$ fixé ($Ntot \simeq 9.0\,10^6$), simulant des processus de Poisson; le second tel que $Ntot = M\,\nu(T_{\min} - T^{\max})$ où ν est une fréquence donnée ($\nu = 20$ Hz), simulant des processus de Hawkes sur l'intervalle de temps $(0, 100]$ avec le code SPIKES (voir [3]). La nouvelle méthode est plus efficace.

D'autre part, le solveur Lasso a été modifié par rapport à [2], c'est désormais la méthode LARS (voir [4]) qui est utilisée et semble plus efficace que la précédente à partir d'un nombre suffisamment élevé de neurones. Les Figures 3 et 4 donnent une comparaison des temps de résolution du Lasso pour deux tests, la Figure 3 correspond au test 2 pour $K = 2$ et la Figure 4 correspond à un test simulant des processus de Hawkes sur $(0, 20]$ pour $K = 5$ (utilisant la fonction *Hawkesmulti* de *neuro-stat*).

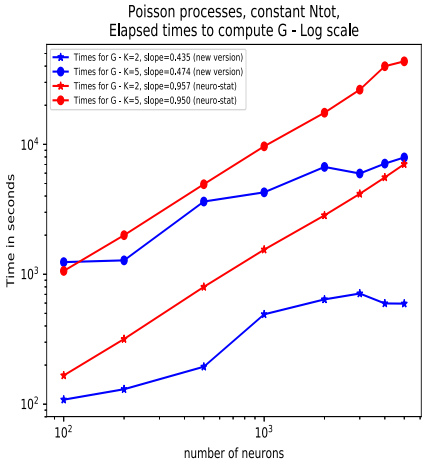


Figure 1: Comparaison des temps calcul de **G** - $K = 2$ et 5 - Test 1

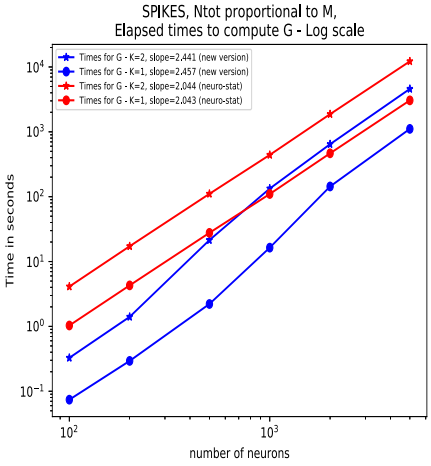


Figure 2: Comparaison des temps calcul de **G** - $K = 1$ et 2 - Test 2

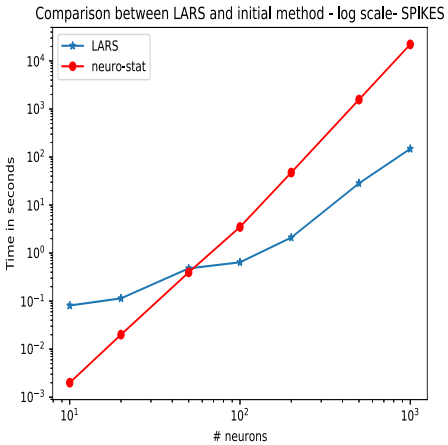


Figure 3: Temps calcul pour Lasso Test 2, K=2

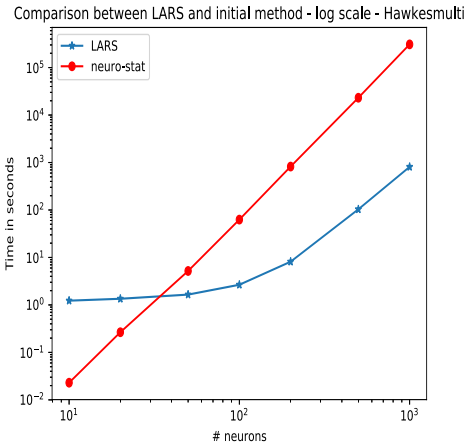


Figure 4: Temps calcul pour Lasso Test 3, K=5

3 Conclusion

On a expliqué comment améliorer le temps de calcul pour la construction de la matrice $\mathbf{G}$ par rapport à la méthode précédente. On procède de manière semblable pour $\mathbf{b}$, $\mu_{\mathbf{2}}$ et $\mu_{\mathbf{A}}$. Cependant, la nouvelle version du calcul de $\mu_{\mathbf{2}}$ reste coûteuse, il devrait être possible de l'améliorer, ce qui fait l'objet d'un travail en cours de finalisation.

Bibliographie

[1] Hansen, N-R, Reynaud-Bouret, P. and Rivoirard, V. (2015), *Lasso and probabilistic inequalities for multivariate point processes*, Bernoulli, 21(1), 83–143.

[2] Lambert, R. et al. (2018), *Reconstructing the functional connectivity of multiple spike trains using Hawkes models*, Journal of Neuroscience Methods, 297, 9–21.

[3] Mascart, C., Muzy, A. and Reynaud-Bouret, P. (2020), *Discrete event simulation of point processes: Computational complexity analysis on sparse graphs*, submitted to ACM Trans. Algor.

[4] Efron, B. and Hastie, T. and Johnstone, I. and Tibshirani, R. (2004), *Least angle regression*, The Annals of Statistics, 2-32; 407–499.

Co-clustering contraint pour le résumé de matrices document-terme

Margot Selosse [1] & Julien Jacques [2] & Christophe Biernacki [3]

[1] *5 Avenue Pierre Mendès France, 69500 Bron margot.selosse@gmail.com*
[2] *5 Avenue Pierre Mendès France, 69500 Bron julien.jacques@univ-lyon2.fr*
[3] *Inria, Université de Lille, CNRS, Laboratoire de mathématiques Painlevé 59650 Villeneuve d'Ascq Lille christophe.biernacki@inria.fr*

Résumé. Le co-clustering est une méthode de fouille de données qui produit simultanément un clustering des observations (en ligne) et un clustering des variables (en colonne). Ce travail présente un nouveau modèle de co-clustering pour résumer des données textuelles stockées sous la forme de matrice document-terme. Nous appelons bloc le croisement d'un cluster en ligne et d'un cluster en colonne. Notre modèle met en évidence des blocs homogènes, mais distingue aussi les blocs significatifs des blocs dits "de bruit". Cela est particulièrement utile pour les matrices document-terme qui sont sparses et de haute dimension. De plus, le modèle propose une organisation parmi les blocs significatifs et de bruit, ce qui rend plus facile pour l'utilisateur d'interpréter les résultats. Un algorithme Stochastic Gibbs Expectation-Maximization (SEM-Gibbs) est utilisé pour l'inférence du modèle.

Mots-clés. Modèle des blocs latents, données textuelles, interprétabilité

Abstract. Co-clustering is a data mining technique which simultaneously produces row-clusters of observations and column-clusters of features. This work presents a novel co-clustering model which easily summarizes textual data in a document-term format. In addition to highlighting homogeneous co-clusters, we also distinguish noisy co-clusters from significant co-clusters, which are particularly useful for sparse document-term matrices. Furthermore, the model proposes a structure among the significant co-clusters, thus providing improved interpretability to users. A Stochastic Expectation-Maximization algorithm is proposed to implement the model's inference as well as a model selection criterion to choose the number of co-clusters.

Keywords. Latent Block Model, textual data, interpretability

1 Introduction

Ce travail présente le modèle SOCC (Self Organised Co-Clustering). Son objectif est de résumer des matrices document-terme, dont les lignes correspondent à des documents, et dont les colonnes correspondent à des mots (ou termes). Une cellule (i, j) correspond

au nombre de fois que le j-ème terme apparaît dans le i-ème document. Le clustering, qui forme des groupes homogènes d'observations (de documents dans notre cas), est une technique non-supervisée qui a prouvé son efficacité dans plusieurs domaines. Cependant, dans des contextes sparses et de haute dimension, les techniques de clustering classiques sont moins adaptées et difficiles à interpréter. Avec de tels jeux de données, le co-clustering - qui regroupe les observations et les variables simultanément - peut se révéler plus efficace. Le co-clustering permet de résumer les jeux de données en blocs (le croisement entre un cluster en ligne et un cluster en colonne). Dans le cadre de matrices document-terme, les clusters de documents aident à trouver les documents qui parlent du même sujet, tandis que les clusters de termes aident à savoir de quoi parlent ces documents.

Nous nous intéressons ici à une approche probabiliste appelée le modèle des blocs latents [3]. Elle suppose que les données sont générées depuis un mélange de distributions de probabilités, dont chaque composant correspond à un bloc. Les paramètres des distributions correspondantes et les appartenances aux blocs sont ensuite estimés à partir des données. Cette approche modélise les éléments d'un bloc avec une distribution paramétrique: chaque bloc est interprétable à partir de ses paramètres de distribution.

Toutefois, lorsqu'il s'agit de données sparses et de haute dimension, plusieurs blocs peuvent être extrêmement sparses (composés de zéros) et causer des problèmes d'inférence. En outre, la mise en évidence de blocs homogènes n'est pas toujours suffisante pour obtenir des résultats faciles à interpréter. En effet, malgré leur homogénéité, ces blocs sparses ne sont pas significatifs du point de vue de l'interprétation, et nous avons besoin d'une nouvelle étape pour différencier les blocs significatifs des autres. En d'autres termes, il est laissé à l'utilisateur de choisir les blocs les plus utiles et de déterminer quels clusters de termes (clusters en colonnes) sont plus spécifiques à quels clusters de documents (clusters en lignes). Cette tâche n'est pas triviale, même avec un nombre raisonnable de clusters en lignes et de colonnes. Il est donc nécessaire de travailler sur une technique de co-clustering qui offre des résultats prêts à l'emploi.

2 Co-clustering et le modèle des blocs latents

2.1 Notations

Nous considérons une matrice X avec un nombre I de lignes et de J colonnes. Nous notons x_{ij} un élément de X tel que $1 \leq i \leq I$ et $1 \leq j \leq J$. Etant dans un contexte de co-clustering, nous supposons qu'il existe G clusters en ligne, H clusters en colonnes. Pour cela, nous introduisons les matrices $\boldsymbol{v}$ et $\boldsymbol{w}$, qui correspondent respectivement aux partitions des clusters en lignes et colonnes. Ainsi v_i, la ième ligne de $\boldsymbol{v}$, est un vecteur de taille G, tel que v_{ig} est égal à 1 lorsque la i-ème ligne appartient au g-ième cluster en ligne, et 0 dans le cas contraire. De la même manière, w_j, la j-ème ligne de $\boldsymbol{w}$ est un vecteur de taille H tel que w_{jh} est égal à 1 lorsque la j-ème colonne appartient h-ième cluster en colonne, et 0 autrement.

2.2 Hypohèses du modèle

Le modèle des blocs latents [3] se base sur les hypothèses suivantes:

Hypothèse 1 *Les partitions v_i, w_j sont indépendantes pour tout $\{i, j\}$.*

Cela se traduit donc par :

$$p(\boldsymbol{v}, \boldsymbol{w}) = p(\boldsymbol{v})p(\boldsymbol{w}) = \prod_i p(v_i) \prod_j p(w_j) = \prod_{ig} \alpha_g^{v_{ig}} \prod_{jh} \beta_h^{w_{jh}}, \tag{1}$$

où $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_G)$ et $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_H)$ sont les proportions de mélange des clusters, en ligne et en colonne respectivement.

Hypothèse 2 *Conditionnellement à $\boldsymbol{v}$ et $\boldsymbol{w}$, les éléments x_{ij} d'un bloc sont indépendants et identiquement distribués.*

Nous avons donc :

$$x_{ij}|v_{ig}w_{jh} = 1 \overset{iid}{\sim} f(.; \theta_{gh}) \text{ pour tout } (i, j).$$

Ici, θ_{gh} représente le paramètre de la distribution du bloc formé par les clusters en lignes g et les clusters en colonnes h. Par la suite, nous l'appellerons simplement le bloc (g, h).

Ainsi, nous obtenons :

$$p(X|\boldsymbol{v}, \boldsymbol{w}) = \prod_{ijgh} f(x_{ij}; \theta_{gh})^{v_{ig}w_{jh}}.$$

La vraisemblance du modèle peut donc être écrite :

$$p(X) = \sum_{\substack{(\boldsymbol{v},\boldsymbol{w}) \\ \in V \times W}} p(X|\boldsymbol{v}, \boldsymbol{w})p(\boldsymbol{v}, \boldsymbol{w}) = \sum_{\substack{(\boldsymbol{v},\boldsymbol{w}) \\ \in V \times W}} \prod_{ig} \alpha_g^{v_{ig}} \prod_{jh} \beta_g^{w_{jh}} \prod_{ijgh} f(x_{ij}; \theta_{gh})^{v_{ig}w_{jh}}, \tag{2}$$

où V et W sont l'ensemble des partitions possibles.

2.3 Le co-clustering avec la distribution de Poisson

Comme x_{ij} dénombre le nombre d'occurence du mot j dans le document i, la modélisation par une loi de Poisson est naturelle. Dans ce contexte, il est considéré qu'un élément x_{ij} est tiré d'une loi de Poisson de paramètre λ_{ij}, soit:

$$x_{ij} \sim \mathcal{P}(\lambda_{ij}).$$

Ainsi, nous avons:

$$f(x_{ij}; \lambda_{ij}) = e^{-\lambda_{ij}} \frac{\lambda_{ij}^{x_{ij}}}{x_{ij}!}.$$

Le paramètre λ_{ij}, lui, est considéré être une fonction d'un effet de bloc θ_{gh}, d'un effet de ligne μ_i et d'un effet de colonne ν_j:

$$\lambda_{ij} = \mu_i \nu_j \sum_{gh} v_{ig} w_{jh} \theta_{gh}.$$

Pour assurer l'identifiabilité du modèle, nous fixons μ_i et ν_j tel que suit:

$$\mu_i = \sum_j x_{ij}, \text{ et } \nu_j = \sum_i x_{ij}.$$

En fixant μ_i et ν_j, le seul paramètre à estimer concernant la distribution de Poisson est $\boldsymbol{\theta} = (\theta_{gh})$.

3 Le modèle Self-organised Co-Clustering (SOCC)

3.1 Le co-clustering contraint SOCC

Jusqu'à maintenant, nous avons décrit un modèle des blocs latents classique, avec utilisation d'une distribution de Poisson. Les paramètres θ_{gh} sont sans rapport et donc chaque bloc doit être interprété séparément des autres. Dans le modèle SOCC, cette indépendance n'est plus supposée. Ainsi, pour un bloc donné (g, h), l'effet de bloc correspondant θ_{gh} sera soit spécifique au cluster en colonnes h avec $\theta_{gh} = \theta_h$, soit non-spécifique, avec $\theta_{gh} = \theta$. Dans le cas d'un effet de bloc non-spécifique $\theta_{gh} = \theta$, le bloc (g, h) est considéré comme un bloc de bruit, et il partage le même paramètre θ avec tous les autres blocs de bruit. Dans le cas de $\theta_{gh} = \theta_h$, le bloc (g, h) est significatif, et partage le même θ_h avec tous les blocs significatifs du même cluster en colonnes h. Dans ce cas, les termes du h-ième cluster en colonne sont considérés comme spécifiques aux documents d'un ou plusieurs clusters en lignes.

Pour organiser les blocs significatifs et les blocs de bruit, plusieurs règles sont données. Tout d'abord, après avoir choisi le nombre de clusters en lignes G, le nombre de clusters en colonnes est égal à $H = G + \binom{G}{2} + 1$. De plus, les clusters en colonnes sont divisés en trois parties appelées *main*, *second* et *common*. Ces parties sont illustrées par la Figure 1 et leur objectif est expliqué ici. La partie *main* concerne les G premiers clusters en colonnes, pour $h \in \{1, ..., G\}$. Dans chaque cluster en colonnes h de cette partie, un seul bloc est significatif et paramétré par θ_h. Tous les autres blocs sont de bruit et paramétrés par θ. Par conséquent, pour chaque cluster de documents, le bloc significatif indique les termes qui sont spécifiques à ces documents. Ainsi dans la partie *main* les blocs significatifs sont en diagonale, et les autres blocs sont des blocs de bruit. La partie *second* concerne les $\binom{G}{2}$ clusters en colonnes suivants ($h \in \{G + 1, ..., G + \binom{G}{2}\}$). Dans chaque cluster en colonnes h de cette partie, deux blocs sont considérés significatifs. Chaque cluster en colonnes contient donc des termes spécifiques à deux clusters de documents. Enfin,

la partie *common* ne comprend qu'un seul cluster en colonnes et rassemble les termes communs à tous les documents.

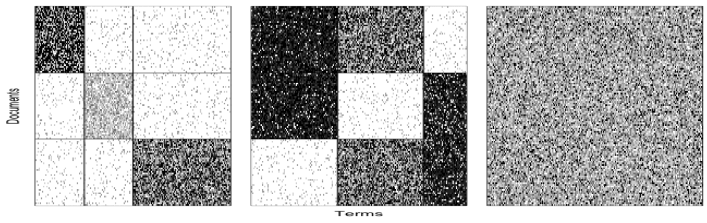


Figure 1: Illustration du co-clustering effectué avec le modèle SOCC. De gauche à droite, la partie *main*, la partie *second* et la partie *common*.

Illustration Dans la Figure 1, nous voyons clairement les blocs significatifs (les plus foncés) avec $\theta_{gh} = \theta_h$ et les blocs de bruit (les plus clairs) avec $\theta_{gh} = \theta$. Nous discernons également l'organisation entre ces blocs et les trois différentes parties *main*, *second* et *common*. Par exemple, dans la partie *main*, le premier cluster en colonnes est considéré comme spécifique au premier cluster en ligne, ainsi seul le premier bloc du cluster en colonnes a sa propre distribution spécifique avec θ_1. En revanche, les autres blocs de ce cluster en colonnes sont considérés comme de bruit et ont un paramètre d'effet de bloc θ, qui est commun à tous les blocs de bruit. Dans la partie *second*, nous observons que pour $h = 4$, les blocs $(1, 4)$ et $(2, 4)$ sont significatifs, et partagent le même effet de bloc θ_4. Cela implique que les termes de cluster en colonne 4 sont spécifiques aux documents des clusters en ligne 1 et 2. De plus, le bloc $(4, 3)$ est de bruit et a le même effet θ que les autres blocs de bruit. La partie *common* est particulière dans la mesure où elle ne contient qu'un seul cluster en colonnes, donc $h = 7$. Ce cluster en colonnes contient les termes spécifiques à tous les clusters de documents et ses blocs correspondants partagent tous le même θ_7.

En notant $\mathcal{C}_h$ les blocs significatifs du cluster en colonnes h et $\overline{\mathcal{C}}_h$ les blocs de bruit du cluster en colonnes h, la probabilité du modèle SOCC s'écrit :

$$p(X) = \sum_{\substack{(\boldsymbol{v},\boldsymbol{w}) \\ \in V\times W}} \prod_{ig} \alpha_g^{v_{ig}} \prod_{jh} \beta_h^{w_{jh}} \prod_{ijh} \prod_{g\in\mathcal{C}_h} f(x_{ij};\theta_h)^{v_{ig}w_{jh}} \prod_{g\in\overline{\mathcal{C}}_h} f(x_{ij};\theta)^{v_{ig}w_{jh}}. \tag{3}$$

3.2 Inférence du modèle et choix de G

Utilisation d'une alternative à l'algorithme EM (Expectation-Maximisation). Pour estimer les variables latentes $\boldsymbol{v}$ et $\boldsymbol{w}$ et les paramètres $\boldsymbol{\theta}$, l'algorithme Expectation-Maximisation (EM) [2] semble être un bon candidat. Cependant, dans le cadre du co-clustering, cet algorithme nécessite de calculer $p(v_{ig}w_{jh} = 1|X)$. Or, cette quantité n'est

pas facilement calculable, ce qui rend l'utilisation d'un EM classique impossible. Nous utilisons alors un algorithme appelé Stochastic-Gibbs Expectation-Maximisation (SEM-Gibbs) [4]. Cet algorithme exécute deux étapes itérativement. Le première étape consiste à utiliser un échantillonneur de Gibbs pour simuler la loi $p(\boldsymbol{v}, \boldsymbol{w}|X)$. Pour cela, nous échantillonnons $\boldsymbol{v}$ conditionnellement à $\boldsymbol{w}$ et X, puis nous échantillonnons $\boldsymbol{w}$ conditionnellement à $\boldsymbol{v}$ et X. La deuxième étape consiste à estimer les paramètres qui maximisent la vraisemblance complétée, et ce avec les partitions fixées.

Utilisation d'un critère ICL pour sélectionner le nombre de cluster. Comme le nombre de clusters en colonnes H est directement induit du nombre de clusters en ligne G (on a $H = G + \binom{G}{2} + 1$), G est le seul nombre à choisir pour utiliser le modèle SOCC. En co-clustering, le critère de sélection BIC (Bayesian Information Criterion) [5] n'est pas calculable non plus. Pour connaître le nombre G optimal, nous utilisons le critère de sélection ICL (Integrated Complete Likelihood) [1].

4 Conclusion

Le modèle SOCC est un nouvel algorithme de co-clustering, adapté pour les matrices document-terme, et basé sur le modèle des blocs latents. Nous comparerons les résultats du modèle avec d'autres techniques sur des jeux de données réels. Nous montrerons aussi les résultats obtenus avec pour données d'étude les trois premiers tomes de Harry Potter.

References

[1] C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence.*, 22(7):719–725, July 2000.

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of he Royal Statistical Society, series B*, 39(1):1–38, 1977.

[3] G. Govaert and M. Nadif. Latent block model for contingency table. *Communications in Statistics - Theory and Methods*, 39(3):416–425, 2010.

[4] Christine Keribin, Gérard Govaert, and Gilles Celeux. Estimation d'un modèle à blocs latents par l'algorithme SEM. In *42èmes Journées de Statistique*, Marseille, France, France, 2010.

[5] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.

EXPOSITION A COURT TERME A LA POLLUTION ATMOSPHERIQUE ET DYSPNEES CARDIAQUES : ETUDE DE CAS EN REGION SUD

Fanny Simões[1], Charles Bouveyron[2], Damien Piga[3], Damien Borel[4], Stéphane Descombes[5], Véronique Paquis-Flucklinger[6], Pierre Gibelin*[7], Jaques Levraut*[8], Silvia Bottini*[1]

[1] Université Côte d'Azur, Maison de la Simulation et des Interactions, France
[2]Université Côté d'Azur, Inria, CNRS, Laboratoire JA Dieudonné, Maasai research team
[3]ATMOSud, Nice, France
[4]Innovation e-Santé Sud, Hyères, France
[5]Université Côté d'Azur, Inria, CNRS, Laboratoire JA Dieudonné, Nice, France
[6]Université Côte d'Azur, CHU, Inserm U1081, CNRS UMR7284, IRCAN, Nice, France
[7]Université Côte d'Azur, Faculté de médicine, Nice, France
[8]Département Hospitalo-Universitaire de Médecine d'Urgence, Nice, France
[*] contributed as co-last authors

Résumé. La pollution de l'air, en particulier les particules fines PM et les polluants gazeux tel que le dioxyde d'azote (NO_2) et l'ozone (O_3), sont une menace maintenant avérée pour la santé mondiale. La durée d'exposition à la pollution de l'air a différents impacts sur la santé. Une exposition à court terme augmente le nombre d'admissions à l'hôpital et le taux de mortalité. Bien que la relation entre exposition à court terme à la pollution et pathologies cardiaques soit largement établi, le lien avec la dyspnée cardiaque est moins clairement établi. Dans cette étude, l'objectif est de mettre en évidence ce lien en région Sud (Provence-Alpes-Côte d'Azur). L'analyse est de 2013 à 2018 avec un total de 43 404 évènements. Les données de polluants (NO_2, PM_{10}, O_3) et climatiques (température et pression) sont journalières. L'analyse statistique s'est décomposée comme suit. Premièrement, la région Sud a été divisée en 366 zones, correspondant au niveau le plus petit de résolution compatible avec les données santés et environnementales. Afin d'identifier des zones similaires en terme de distribution de dyspnées cardiaques au cours du temps, une technique de clustering fonctionnel multivarié (fun-HDDC) a été appliquée. Puis, le distributed lag non-linear model (DLNM) a permis de mettre en évidence une relation entre l'exposition à court terme à la pollution et les dyspnées cardiaques. Enfin, le développement d'une application web nommé HEART (**H**ealth, **E**nvironment in PAC**A** **R**egion **T**ool) permet de présenter l'ensemble des résultats plus aisément. L'étude montre que chaque polluant a un effet spécifique sur les dyspnées cardiaques, le nombre de jours entre l'évènement de dyspnée cardiaque et l'exposition à un polluant diffère selon le polluant et la zone étudiée. Enfin, il est possible d'estimer le niveau de seuil de pollution qui peut augmenter le risque relatif de dyspnée cardiaque.

Mots-clés. dyspnée cardiaque, pollution, santé, environnement, DLNM, séries temporelles, effets retardés, clustering

Abstract. There is a growing body of evidence that air pollution is a significant threat to health worldwide. Air pollution is composed of particulate matter (PM) and gaseous pollutants, such as nitrogen dioxide (NO2) and ozone (O3). The time exposure to air pollution leads diverse impact on the health. Although the relationship between short-term exposure to air pollution and several cardiovascular pathologies is widely established, the link regarding cardiac dyspnea is still to explore. In this study, we aim to fill this gap using the region Sud

(Provence-Alpes-Côte d'Azur) as a model. We focused on the 2013-2018 period for which we collected 43,404 events of cardiac dyspnea. We collected pollutants (NO2, PM10 and O3) and climate (temperature and pressure) measurements on a daily basis. We set up a reproducible statistical framework to analyse these data. Briefly, we first divided the region Sud in 366 areas in order to match environmental and health data with minimal resolution. To identify areas with similar distribution of the number of dyspnea events over time, we applied a multivariate time-series clustering technique (*fun-HDDC*). Then, we employed a distributed lag non-linear model (DLNM) to show the relationship between short-term exposure to outdoor air pollution and the incidence of cardiac dyspnea events. Finally, we developed a user-friendly web application called HEART (Health, Environment in PACA Region Tool) to easily show the results of this study. This study showed that each pollutant has an effect on triggering cardiac dyspnea on different time frames between the pollution peak and the event. Overall, we established a model to allow the prediction of the pollutants threshold levels that may trigger new dyspnea events that can be generalize to other areas.

Keywords. cardiac dyspnea, pollution, health, environment, DLNM, time series, delayed effects, clustering

1. Introduction

La pollution de l'air engendre de nombreux décès et maladies à travers le monde et, selon l'OMS, elle est la cause de 7 millions de décès par an. Il existe différentes sources de pollution qui peuvent être originaires de l'activité humaine (industrie, chauffage, trafic routier) mais aussi naturelles (éruption volcanique, tempête de sable). La qualité de l'air est également influencée par les conditions climatiques (la pluie, la pression atmosphérique, le vent et la température), ce qui rend moins aisée sa modélisation. Concernant les polluants affectant la santé de l'homme, il est d'usage de considérer le PM_{10} (particules fines inférieures à 10 microns) le NO_2 (dioxyde d'azote) et l'O_3 (ozone). Il est à noter que le PM_{10} et le NO_2 sont principalement émis par les industries et le chauffage. L'ozone (O_3) quant à lui n'est pas émis directement, il se forme lorsque qu'il y a de forte chaleur et de l'ensoleillement. Il est clairement admis à présent que la pollution a des effets néfastes sur la santé et peut engendrer des maladies cardiovasculaires, des cancers ou encore des problèmes respiratoires. Les effets sur la santé sont cependant différents selon une exposition à plus ou moins long terme. Bourdrel (2017) montre qu'une exposition à long terme engendre une augmentation moyenne de 11 % de la mortalité cardiovasculaire pour une augmentation annuelle de 10 µg/m^3 en $PM_{2,5}$. De même que la mortalité cardiovasculaire augmente pour des expositions à long et court terme au NO_2. L'ozone a aussi des effets sur la santé, Raza (2019) met en exergue qu'une exposition à court terme d'O_3 est associée à un risque plus élevé d'arrêts cardiaques. Bien que certains travaux de recherche montre qu'il existe une relation entre exposition à court terme à la pollution et pathologies cardiaques, le lien avec la dyspnée cardiaque n'est pas encore clairement compris.

Cette étude a pour but de mettre en évidence ce lien et se focalise donc sur l'effet de la pollution de l'air à court terme sur les dyspnées d'origine cardiaque. Jusqu'à présent, les chercheurs ont rarement étudié de grands territoires et ce sont principalement concentrés sur des villes. Ici, l'étude porte sur la région Sud (Provence Alpes Côte d'Azur) pour analyser l'impact de la pollution sur les dyspnées cardiaques à grande échelle. Afin d'établir un lien, nous utilisons une méthode de clustering fonctionnel multivarié (FunHDDC) pour le regroupement des sites en groupes homogènes, puis, pour chaque site, le modèle Distributed Lag Non-Linear Model (DLNM) qui permet de prendre en compte l'effet retardé des polluants,

l'effet n'étant pas forcément immédiat sur la dyspnée cardiaque. Le modèle DLNM est particulièrement utilisé dans les études épidémiologiques en lien avec des données environnementales. Slama (2019) utilise un modèle DLNM afin de déterminer l'impact de la pollution de l'air sur les admissions à l'hôpital en particulier les maladies respiratoires. Il est démontré que les hospitalisations pour maladies respiratoires sont en lien avec un pic de concentration de PM_{10}, le nombre d'admissions est plus élevé dans un délais de 2 à 6 jours après un pic. De même, Xia (2019) trouve que le PM_{10} a des effets sur le court terme sur les maladies respiratoires et que le NO_2 ne semble pas avoir d'effet significatif.

2. Méthodes

2.1 Données

Notre étude s'appuie sur deux bases de données de natures différentes. D'une part, les données santé, fournies par ORUPACA (https://ies-sud.fr/), se composent de 43 404 patients âgés entre 18 et 115 ans, s'étant présentés entre 2013 à 2018 dans un des 47 services d'urgence de la région Sud et atteints de dyspnées cardiaques. L'âge et le sexe du patient sont connus, ainsi que son code postal de résidence ce qui permettra la réconciliation avec l'exposition aux polluants. D'autre part, les données environnementales, fournies par AtmoSud (https://www.atmosud.org/), sont composées de l'ensemble des mesures de pollution et de météo journalières sur la période 2013 à 2018. Ce sont des données carroyées à 4km pour l'ensemble de la région Sud (1 995 carreaux au total). Les données de pollution comportent 3 polluants par carreaux et par jour : le maximum observé de NO_2, le maximum observé d'O_3 et le niveau moyen de PM_{10}. Les données de météo sont quant à elles composées de 5 variables par carreaux et par jour (1 967 carreaux) : la température maximale, la température minimale, le niveau de pluie, la pression maximale et la pression minimale.

2.2 Prétraitement et réconciliation des bases

Il a tout d'abord été nécessaire de gérer les données manquantes de la base pollution-météo : il y avait 2 % de jours manquants pour les données pollution et 1% pour les données météo sur 2191 jours entre 2013 et 2018. Des valeurs aberrantes ont été également été détectées et ont été considérées comme des données manquantes. Pour imputer ces données manquantes, deux méthodes ont été utilisées pour remplacer les données manquantes : interpolation et random sampling. La méthode d'interpolation est utilisée pour les variables présentant une tendance ou saisonnalité et la méthode de random sampling pour les variables avec une distribution aléatoire. La méthode d'interpolation s'est faite avec la fonction *na.approx()* du package R *"zoo".* Concernant la méthode du random sampling, les valeurs qui remplacent les données manquantes sont prélevées dans une fenêtre avant et après la période de données manquantes. La taille de la fenêtre est équivalente à 4 + la longueur de la période manquante divisée par 2. La fenêtre s'adapte suivant la période de données manquantes à remplacer. Concernant la réconciliation des bases, les données de santé et environnementales ne sont malheureusement pas fournies au même niveau de résolution : les données de santé sont au niveau du code postal tandis que les données environnementales sont des données carroyées à 4 km. Un code postal peut être associé à plusieurs communes mais peut aussi être différent pour une même commune (4 codes postaux pour Nice par exemple). Afin d'opérer la réconciliation, la région Sud a donc été divisée en zones qui correspondent à des fusions de communes avec le même code postal. Une zone peut aussi contenir plusieurs codes postaux comme Nice car les zones sont basées sur le découpage

communal. Au final, on obtient 366 zones pour 385 codes postaux pour la région. Les données environnementales ont donc dû être également transformées pour être au même niveau de résolution que celle des données santé, au niveau des zones. Il s'agit là de données carroyées à 4 km ; cependant le carroyage de la pollution et de la météo n'est pas identique et le nombre de carreaux est lui aussi différent. L'association s'est donc faite en 3 temps, dans un premier temps une association données pollution – zones puis données météo – zones et enfin données pollution-météo avec les données santé au niveau des zones. Lorsqu'un carreau de 4 km recouvre la plus grande surface d'une zone il lui est associé. Plusieurs carreaux peuvent être associés à une même zone. Dans le cas où seulement un carreau est associé à une zone alors ce dernier est associé aux valeurs de météo et de pollution de l'unique carreau qu'il contient. A contrario, dans le cas où plusieurs carreaux sont associés à une même zone alors la zone est associée à une unique valeur de pollution et météo. Au terme de la réconciliation des bases, nous disposons d'une base de 34 948 lignes, correspondant aux patients associés à chaque code postal de la région, et 109 colonnes, correspondant aux variables santé, pollution et météo.

2.3 Analyse statistique

L'analyse statistique a consisté en deux principales étapes. Tout d'abord, une méthode de clustering de séries temporelles multivariées a été appliquée pour former des groupes de sites (codes postaux) ayant une relation homogène entre proportion de dyspnées et données environnementales. La méthode de clustering utilisée se nomme *funHDDC* et est basée sur un modèle de mélange adapté aux données fonctionnelles. Une fois le clustering effectué, un *distributed lag non-linear model* (DLNM) est estimé pour les sites de chaque cluster. Pour cela, des fonctions *cross-basis* sont définies au préalable pour les variables de pollution et de météo. Une fonction de temps est intégrée au modèle afin de prendre en compte la tendance et la saisonnalité des données ainsi qu'une variable correspondant au jour de la semaine pour prendre en compte l'effet du week-end. De plus, une variable de population est aussi intégrée au modèle afin de prendre en compte la taille de la zone étudiée dans l'analyse. La variable à expliquer dans ce cas est le nombre de dyspnées par zone et par date. Une régression quasi-poisson est finalement appliquée.

3. Résultats

L'analyse statistique a donc porté sur le nombre de dyspnées cardiaques par zones et par dates entre 2013 et 2018 en région Sud. L'analyse des patients atteints de dyspnées cardiaques montre que les individus les plus âgés sont les plus touchés par cette pathologie, qu'elle touche autant les hommes que les femmes et que le nombre de dyspnées cardiaques est plus élevé en période hivernale. Concernant les variables environnementales, le NO_2 et les volumes de pluie ne présentent pas de tendance particulière, tandis que l'O_3, le PM_{10}, la température et la pression ont, comme on peut l'attendre, une tendance et/ou une saisonnalité marquée.

3.1 Clustering fonctionel

La méthode FunHDDC de Bouveyron et al. (2011) a permis de déterminer 6 groupes homogènes en terme de nombre de dyspnées cardiaques. La figure 1A présente la repartition des zones dans les 6 clusters .

A B

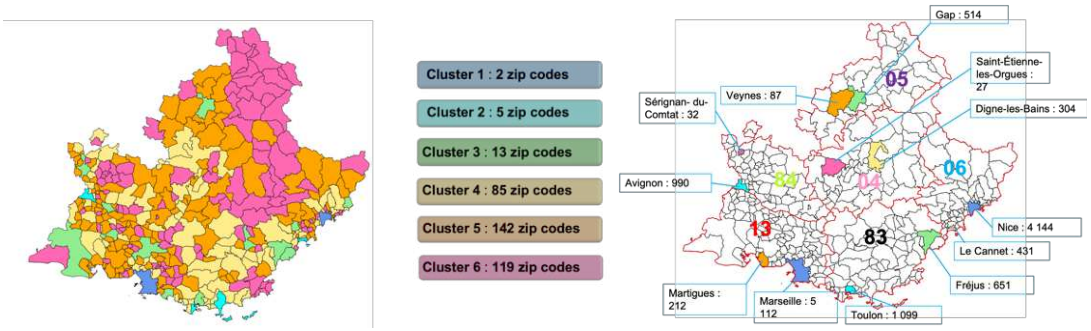


Figure 1. A) Repartition des 358 zones dans les 6 clusters trouvé avec la méthode FunHDDC. B) Carte des sites sélectionnés, le couleur représente le cluster d'appartenance, dans le carré bleu, la ville représentative de chaque zone sélectionnée et le nombre de dyspnée cardiaque sont indiqué

Sur la base de ces résultats de clustering, deux sites (codes postaux) représentatifs de chaque cluster ont été retenus : la zone géographique ayant le plus de dyspnées cardiaques en proportion et le chef-lieu d'un département. La figure 1B présente une carte des sites dont nous détaillons l'analyse avec le modèle DLNM ci-après.

3.2 Modèle DLNM

Nous détaillons à présent les résultats de l'analyse des sites géographiques retenus grâce au modèle DLNM. Le risque relatif d'avoir une dyspnée cardiaque est calculé à l'aide du modèle DLNM avec 14 jours de retards. On obtient comme résultat final l'effet global d'une variable de pollution ou météo sur la dyspnée cardiaque suivant différents seuils et l'effet propre à chaque retard selon un seuil fixé arbitrairement. Nous avons utilisé comme seuils ceux recommandés par l'OMS comme à ne pas dépasser : 200 µg/m^3 pour le NO_2, 180 µg/m^3 pour l'O_3 et 50 µg/m^3 pour le PM_{10} . La figure 2 présente les résultats obtenus pour Marseille et Nice en ce qui concerne le NO_2. On remarque que pour ces deux villes, il y a une augmentation significative du risque de dyspnée, de 1,2 et 1,5 respectivement, si le NO2 dépasse le seuil de 200 µg / m^3, trois ou deu jours avant l'événement.

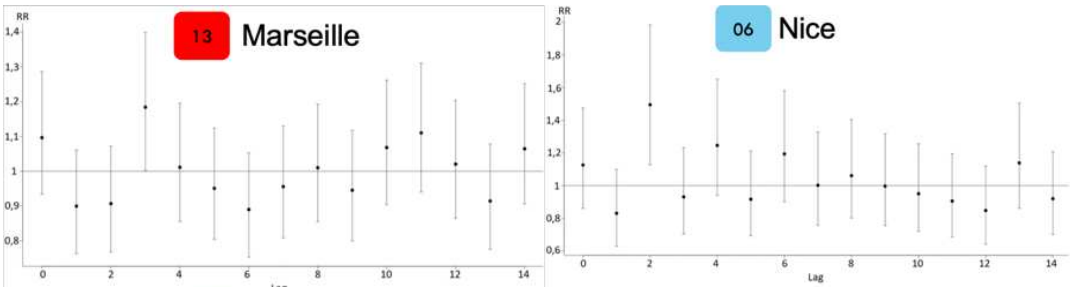


Figure 2. Résultats obtenus avec le modèle DLNM pour Marseille et Nice pour le polluant NO2.

La figure 3 présente un résumé des résultats obtenus sur l'ensemble des villes d'intérêt. Il apparaît que les 3 polluants ont des effets différents selon la zone étudiée. On peut toutefois identifier des tendances globales. Concernant l'O_3, il semble que l'ozone n'ai pas d'effet immédiat sur la dyspnée mais plutôt sur une période longue de 6 à 8 jours. Le PM_{10} à quant à lui un effet clair sur le court terme dans les grandes villes de la région. Concernant le NO_2, il n'y a aucune tendance générale qui se dégage, il est assez dépendant de la zone analysée. Les variables de météo semblent elles aussi avoir un lien avec les dyspnées cardiaques. Globalement, comme attendu, le nombre de dyspnées cardiaques étant plus élevé l'hiver que l'été, une température moyenne faible augmenterait le risque relatif de dyspnées cardiaques.

Dans les villes de 50 000 à 100 000 habitants, comme Avignon et Fréjus, nous avons constaté que l'O3 avait l'effet le plus significatif sur les événements de dyspnée cardiaque. Un pic d'O3 supérieur à 180 µg / m^3 dans ces villes, augmentera de cinq fois le nombre d'événements de dyspnée après 7 et 8 jours respectivement. Nous avons observé un effet plus modéré des PM10 sur les événements de dyspnée et principalement sur les villes du littoral, comme Marseille, Nice et Toulon, pour lesquelles le risque d'événements de dyspnée augmente significativement si le niveau de PM10 dépasse 50 µg / m^3.

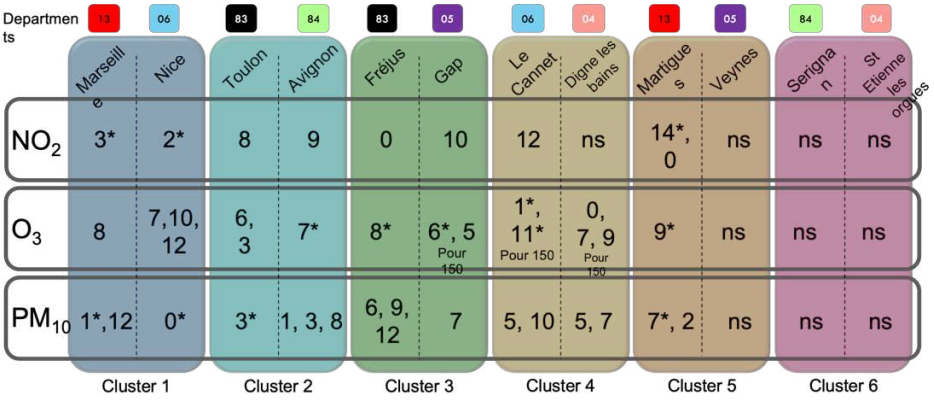


Figure 3. Pour chaque ville indiquée dans figure 1B, identification des lags dont l'impact sur la dyspnée est significatif est reporté. * indique résultats particulièrement significatifs.

4. Conclusion

Nous avons mené une étude de grande ampleur sur la région Sud afin d'établir un lien entre pollution atmosphérique et dyspnées cardiaques. Pour cela, et après avis de la CNIL, nous avons réconcilié 6 années de données des centres d'urgence de la région Sud (OruPaca) avec les données de pollution atmosphérique fournies par AtmoSud. Notre étude a démontré que chacun des polluants étudiés a un effet significatif sur le déclenchement de la dyspnée cardiaque. En particulier, nous avons constaté que les plus grandes villes de la région, avec plus de 300 000 habitants, comme Nice et Marseille ont une augmentation significative du risque de dyspnée de 1,5 et 1,2 respectivement, si le NO2 passe le seuil de 200 µg / m^3, deux ou trois jours avant l'événement. Nous avons également réalisé une application Shiny, appelée HEARTH, qui permet aux médecins et au décideurs d'explorer les résultats de notre analyse pour chacun des codes postaux de la région Sud.

Bibliographie

Bourdrel T, et al. (2017). Cardiovascular effects of air pollution, *Archives of Cardiovascular Disease,* 110, 634-642.

Bouveyron C, et al. (2011). Model-based clustering of time series in group-specific functional subspaces, *Advances in Data Analysis and Classification,* Springer Verlag, 5, 281-300.

Raza A, et al. (2019). Ozone and cardiac arrest: the role of previous hospitalizations, *Environmental Pollution*, 245, 1-8.

Slama A, et al. (2019). Impact of air pollution on hospital admissions with a focus on respiratory diseases: a time-series multi-city analysis, *Environmental Science and Pollution Research*, 26, 16998-17009.

Xia C, et al. (2019). Quantification of the Exposure-lag-response association between air pollution and respiratory disease morbidity in Chongqing city, China*, Environmental Modeling & Assessment*, 24, 331-339.

Débiaiser la descente de gradient stochastique en présence de données manquantes

Aude Sportisse [1,2] & Claire Boyer [1,3] & Aymeric Dieuleveut [4] & Julie Josse [2,4]

[1] *Laboratoire de Probabilités Statistique et Modélisation, Sorbonne Université, 4 place Jussieu, 75252 Paris, aude.sportisse,claire.boyer@upmc.fr*

[2] *Centre de Mathématiques Appliquées, Ecole Polytechnique, Bâtiment Turing, 1 Rue Honoré d'Estienne d'Orves, 91120 Palaiseau, aymeric.dieuleveut,julie.josse@polytechnique.edu*

[3] *Département de Mathématiques et applications, Ecole Normale Supérieure, Paris*

[4] *XPOP, INRIA, France*

Résumé. Les données à très grande échelle ont fait émerger un enjeu majeur: leur traitement lorsqu'elles contiennent des valeurs manquantes. Dans cet exposé, nous proposons un algorithme de gradient stochastique moyenné traitant les valeurs manquantes dans les modèles linéaires. Cette approche a l'avantage de ne nécessiter aucune modélisation de la distribution des données et de prendre en compte des proportions manquantes hétérogènes. Dans le cadre de l'apprentissage en ligne et dans celui des échantillons finis, nous prouvons que cet algorithme permet d'obtenir un taux de convergence de $\mathcal{O}(\frac{1}{n})$ à l'itération n, identique à celui obtenu sans les valeurs manquantes. Nous illustrons les propriétés de convergence et la pertinence de l'algorithme non seulement sur des données simulées mais aussi sur des données réelles, dont celles collectées dans des registres médicaux.

Mots-clés. optimisation, apprentissage supervisé, approximation stochastique, données manquantes.

Abstract. A major caveat of large scale data is their incompleteness. In this talk, we propose an averaged stochastic gradient algorithm handling missing values in linear models. This approach has the merit to be free from the need of any data distribution modeling and to account for heterogeneous missing proportion. In both streaming and finite-sample settings, we prove that this algorithm achieves convergence rate of $\mathcal{O}(\frac{1}{n})$ at the iteration n, the same as without missing values. We show the convergence behavior and the relevance of the algorithm not only on synthetic data but also on real data sets, including those collected from medical register.

Keywords. optimization, supervised learning, stochastic approximation, missing data.

1 Introduction

Les algorithmes de gradient stochastique (Robbins and Monro, 1951) jouent un rôle central dans les problèmes d'apprentissage statistique, en raison de leur faible coût de calcul et de leur mémoire par itération. Beaucoup de variantes de ces algorithmes ont été étudiées, par exemple celle qui utilise l'agrégation des itérés (Polyak and Juditsky, 1992), et des garanties théoriques fortes ont été obtenues (Moulines and Bach, 2011; Bach and Moulines, 2013; Dieuleveut et al., 2017). Plus généralement, les stratégies d'agrégation ont été utilisées pour stabiliser le comportement de l'algorithme et réduire l'impact du bruit, donnant ainsi de meilleurs taux de convergence sans s'appuyer sur l'hypothèse de forte convexité.

Le problème des valeurs manquantes est omniprésent dans l'analyse des données à grande échelle. L'un des principaux défis en présence de données manquantes est de traiter la nature semi-discrète des données qui peuvent être considérées comme un mélange de données continues (les valeurs observées) et de données catégorielles (les valeurs manquantes). En particulier pour les méthodes de gradient, la minimisation de risque avec données incomplètes devient insoluble et les résultats habituels ne peuvent pas être appliqués directement.

Contexte. Nous considérons un modèle de régression linéaire, pour $i \geq 1$,

$$y_i = X_{i:}^T \beta^\star + \epsilon_i, \tag{1}$$

paramétrisé par $\beta^\star \in \mathbb{R}^d$,où $y_i \in \mathbb{R}$, $\epsilon_i \in \mathbb{R}$ est le bruit centré et $X_{i:} \in \mathbb{R}^d$ sont les covariables pour l'observation i.

En supposant que les covariables $X_{i:}$ sont incomplètes, notre but est d'adapter les algorithmes de gradient stochastique, pour estimer les paramètres du modèle linéaire, en présence de données manquantes, et d'obtenir des garanties théoriques sur l'excès de risque.

Travaux connexes. Malgré une littérature abondante sur le traitement des valeurs manquantes (Little and Rubin, 2019), il existe encore des défis à relever, même pour les modèles de régression linéaire, en particulier lorsque les données sont de grande dimension. Une approche classique pour estimer les paramètres du modèle avec des valeurs manquantes consiste à maximiser la vraisemblance observée, en utilisant par exemple l'algorithme Espérance-Maximisation (Dempster et al., 1977). Même si cette approche peut être utilisée pour des données à grande échelle, voir par exemple (Cappé and Moulines, 2009), l'un de ses principaux inconvénients est de s'appuyer sur des hypothèses paramétriques fortes portant sur la distribution des covariables. Une autre stratégie populaire pour résoudre le problème des valeurs manquantes consiste à prédire au mieux les valeurs manquantes pour obtenir des données complètes, puis à appliquer la méthode souhaitée. De nombreuses techniques d'apprentissage machine ont été développées pour

imputer des ensembles de données à l'aide de puissants modèles prédictifs. Cependant, la complétion de la matrice est un problème très différent de l'estimation des paramètres et peut conduire à un biais non contrôlé et à une variance sous-évaluée de l'estimation (Little and Rubin, 2019). Dans le cadre de la régression, Jones (1996) a étudié le biais induit par l'imputation naïve.

Pour le sélecteur de Dantzig (Rosenbaum et al., 2010) et le lasso (Loh and Wainwright, 2011), une autre solution a été proposée, consistant à imputer naïvement la matrice incomplète et à modifier l'algorithme utilisé dans le cas complet pour tenir compte de l'erreur d'imputation. Une telle stratégie a également été étudiée par Ma and Needell (2017) pour la descente de gradient stochastique (SGD) dans le contexte de la régression linéaire avec des valeurs manquantes et avec des échantillons finis : les auteurs ont utilisé des gradients débiaisés, dans le même esprit que le débiaisage de la matrice de covariance considéré par Loh and Wainwright (2011) dans un contexte de régression linéaire parcimonieuse, ou par Koltchinskii et al. (2011) pour la complétion de la matrice. Cette version modifiée de l'algorithme SGD (Ma and Needell, 2017) converge en espérance vers l'estimateur des moindres carrés ordinaires, atteignant le taux de $\mathcal{O}(\frac{\log n}{\mu n})$ pour l'excès de risque empirique après n itérations et un pas décroissant de $(\frac{1}{\mu k})_{1\leq k\leq n}$ pour l'itération k. Les pas $(\frac{1}{\mu k})_{1\leq k\leq n}$ sont couramment utilisés lorsque la fonction objectif est supposée μ-fortement convexe. Cependant, un tel pas nécessite la connaissance de la constante μ qui est souvent inaccessible pour les données à grandes échelles.

Contributions. Nous développons un algorithme SGD agrégé débiaisé pour effectuer une régression linéaire en ligne ou avec des échantillons finis, lorsque des covariables sont manquantes. L'approche consiste à imputer les covariables par 0 et à utiliser des gradients débiaisés en conséquence. Nous supposons que le manque d'une covariable est indépendant de toute valeur de covariable (cas classique de données MCAR[1]). En outre, chaque covariable peut avoir une probabilité différente d'être manquante (cadre hétérogène).

Nous établissons des taux de convergence pour cet algorithme en termes de risque de généralisation. La procédure d'agrégation accélère la convergence, de $\log(n)/\mu n$ à $1/n$, ce qui est optimal et similaire au taux de SGD agrégé sans aucune valeur manquante.

Nous montrons la pertinence de l'approche proposée et ses propriétés de convergence sur des applications numériques ainsi que son efficacité sur des données réelles, dont le jeu de données TraumaBase®, qui permet d'aider les docteurs à prendre des décisions rapides dans la prise en charge des patients gravement traumatisés.

Dans ce document, nous présentons l'algorithme SGD agrégé débiaisé et nous énonçons notre principal résultat théorique concernant son taux de convergence.

[1]MCAR pour Missing Completely At Random en anglais.

2 Principaux résultats

Définition du problème Soit $f_i(\beta) := \left(\langle X_{i:}, \beta\rangle - y_i\right)^2/2$. Nous souhaitons estimer les paramètres du modèle linéaire

$$\beta^\star = \mathrm{argmin}_{\beta\in\mathbb{R}^d}\left\{R(\beta) := \mathbb{E}_{(X_{i:},y_i)}\left[f_i(\beta)\right]\right\}, \tag{2}$$

où $\mathbb{E}_{(X_{i:},y_i)}$ est l'espérance sous la distribution de $(X_{i:}, y_i)$ (qui est indépendente de i car les observations sont supposées i.i.d.).

Nous considérons que les covariables peuvent contenir des données manquantes, on observe alors $X_{i:}^{\mathrm{NA}} \in (\mathbb{R}\cup\{\mathtt{NA}\})^d$ telle que

$$X_{i:}^{\mathrm{NA}} = X_{i:}\odot D_{i:} + \mathtt{NA}(\mathbf{1}_d - D_{i:}),$$

où $\odot$ est le produit élément par élément, $\mathbf{1}_d \in \mathbb{R}^d$ est le vecteur de 1 et $D_{i:} \in \{0,1\}^d$ est un vecteur binaire indiquant la présence de données manquantes dans $X_{i:}$, tel que $D_{ij} = 0$ si l'entrée (i,j) est manquante dans $X_{i:}$, et $D_{ij} = 1$ sinon. On prend pour convention $\mathtt{NA}\times 0 = 0$ et $\mathtt{NA}\times 1 = \mathtt{NA}$. Nous considérons des probabilités de manque hétérogènes, c-à-d. D est une matrice de Bernoulli telle que

$$D = (\delta_{ij})_{1\le i\le n, 1\le j\le d} \quad \text{avec} \quad \delta_{ij}\sim\mathcal{B}(p_j), \tag{3}$$

où $1-p_j$ est la probabilité d'être manquante pour la covariable j.

Notre approche consiste à imputer les covariables incomplètes par 0 dans $X_{i:}^{\mathrm{NA}}$, de telle sorte que

$$\tilde{X}_{i:} = X_{i:}^{\mathrm{NA}}\odot D_{i:} = X_{i:}\odot D_{i:},$$

et en prenant en compte l'erreur d'imputation, comme détaillé dans l'algorithme ci-dessous.

Algorithme SGD agrégé débiaisé La méthode proposée est détaillée dans l'Algorithme 1. L'impact de l'imputation naïve par 0 se traduit par un biais dans le gradient. Par conséquent, à chaque itération, nous utilisons une estimation débiaisée $\tilde{g}_{i_k}$. Afin de stabiliser l'algorithme stochastique, nous considérons les itérés agrégés de Polyak-Ruppert (Polyak and Juditsky, 1992)

$$\bar{\beta}_k = \frac{1}{k+1}\sum_{i=0}^{k}\beta_i.$$

Notons que lorsque les covariables sont complètement observées, l'Algorithme 1 est équivalent à l'algorithme stochastique moyenné pour la régression des moindres carrés.

Algorithme 1 SGD agrégé pour des données manquantes hétérogènes

Entrée: données $\tilde{X}, y, \alpha$ (pas)
Initialiser $\beta_0 = 0_d$.
Soit $P = \text{diag}\left((p_j)_{j\in\{1,\dots,d\}}\right) \in \mathbb{R}^{d\times d}$.
pour $k = 1$ **à** n **calculer**

$$\tilde{g}_k(\beta_k) = P^{-1}\tilde{X}_{k:}\left(\tilde{X}_{k:}^T P^{-1}\beta_k - y_k\right) - (\mathrm{I} - P)P^{-2}\text{diag}\left(\tilde{X}_{k:}\tilde{X}_{k:}^T\right)\beta_k \tag{4}$$

$\beta_k = \tilde{X}_{k-1} - \alpha\tilde{g}_k(\beta_{k-1})$
$\bar{\beta}_k = \frac{1}{k+1}\sum_{i=0}^{k}\beta_i = \frac{k}{k+1}\bar{\beta}_{k-1} + \frac{1}{k+1}\beta_k$
end pour

Résultat théorique Le théorème suivant donne le taux de convergence de l'Algorithme 1 en terme d'excès de risque.

Théorème 1 *Supposons que pour tout i, $\|X_{i:}\| \leq \gamma$ presque-sûrement pour $\gamma > 0$. Pour tout pas constant $\alpha \leq \frac{1}{2L}$, l'Algorithme 1 garantit que pour tout $k \geq 0$:*

$$\mathbb{E}\left[R\left(\bar{\beta}_k\right) - R(\beta^\star)\right] \leq \frac{1}{2k}\left(\frac{\sqrt{c(\beta^\star)d}}{1-\sqrt{\alpha L}} + \frac{\|\beta_0 - \beta^\star\|}{\sqrt{\alpha}}\right)^2,$$

avec L la constante de Lipschitz du gradient débiaisé $\tilde{g}_k$, pour tout $k \geq 0$, défini dans l'Equation (4) et

$$c(\beta^\star) = \frac{\text{Var}(\epsilon_k)}{p_m^2} + \left(\frac{(2+5p_m)(1-p_m)}{p_m^3}\right)\gamma^2\|\beta^\star\|^2, \tag{5}$$

où $p_m = \min_{j=1,\dots d}\ p_j$.

L'excès de risque est de l'ordre de n^{-1}, sans hypothèse de forte convexité (pas de dépendance en μ) après n itérations.

Bibliographie

Bach, F. and E. Moulines
2013. Non-strongly-convex smooth stochastic approximation with convergence rate o (1/n). In *Advances in neural information processing systems*, Pp. 773–781.

Cappé, O. and E. Moulines
2009. On-line expectation–maximization algorithm for latent data models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):593–613.

Dempster, A. P., N. M. Laird, and D. B. Rubin
1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Dieuleveut, A., N. Flammarion, and F. Bach
2017. Harder, better, faster, stronger convergence rates for least-squares regression. *The Journal of Machine Learning Research*, 18(1):3520–3570.

Jones, M. P.
1996. Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American statistical association*, 91(433):222–230.

Koltchinskii, V., K. Lounici, A. B. Tsybakov, et al.
2011. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329.

Little, R. J. and D. B. Rubin
2019. *Statistical analysis with missing data*, volume 793. John Wiley & Sons.

Loh, P.-L. and M. J. Wainwright
2011. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. In *Advances in Neural Information Processing Systems*, Pp. 2726–2734.

Ma, A. and D. Needell
2017. Stochastic gradient descent for linear systems with missing data. *arXiv preprint arXiv:1702.07098*.

Moulines, E. and F. R. Bach
2011. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, Pp. 451–459.

Polyak, B. T. and A. B. Juditsky
1992. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855.

Robbins, H. and S. Monro
1951. A stochastic approximation method. *The annals of mathematical statistics*, Pp. 400–407.

Rosenbaum, M., A. B. Tsybakov, et al.
2010. Sparse recovery under matrix uncertainty. *The Annals of Statistics*, 38(5):2620–2651.

Analyse statistique des accidents routiers de la région Franche-Comté

Cécile Spychala [1], Clément Dombry [2] & Camelia Goga [3]
Laboratoire de Mathématiques de Besançon, UMR CNRS 6623,
Univ. Bourgogne Franche-Comté

[1] *cecile.spychala@univ-fcomte.fr*
[2] *clement.dombry@univ-fcomte.fr*
[3] *camelia.goga@univ-fcomte.fr*

Résumé. L'analyse et la modélisation statistique des données sur les accidents de la route sont cruciaux pour atteindre des objectifs de sécurité en aidant les autorités nationales à prendre les mesures nécessaires pour réduire la fréquence et la gravité des accidents. Ce travail vise à apporter une analyse statistique des accidents corporels routiers de la région Franche-Comté, s'intéressant particulièrement au niveau de gravité de ceux-ci (léger, grave ou mortel). Plusieurs analyses de correspondances multiples ont été menées afin d'évaluer les associations entre la gravité des accidents et de nombreux facteurs s'y relatant, tout en essayant de dresser un cadre spatio-temporel des accidents se produisant en Franche-Comté. Un modèle log-linéaire a été ajusté aux données afin d'étudier les relations et les interactions possibles entre plusieurs caractéristiques liées aux accidents. Enfin, la gravité des accidents a été modélisée par une régression logistique ordinale afin d'évaluer les effets propres de chacune des circonstances des accidents. Les données mises à disposition pour cette étude ont été extraites des fiches BAAC (Bulletin d'Analyse des Accidents Corporels), recensement national français des accidents de la route.

Mots-clés. accidentologie, analyse des correspondances multiples, modèle log-linéaire, régression logistique ordinale.

Abstract. Understanding and modeling road crash data is crucial in fulfilling safety goals by helping national authorities to undertake necessary conditions to reduce crash frequencies and severities. This work aims at conducting a statistic analysis of Franche-Comté region road crashes, with special attention on the severity (slight, serious or fatal). Multiple Correspondence Analyses have been made in order to assess relationships between road crash severity and many accident-related factors, while trying to establish a spatio-temporal framework for accidents occurring in Franche-Comté. A log-linear model was fitted in order to study associations and interactions between some accident-related factors. Then, the accident severity was modeled by an ordinal logistic regression in order to evaluate the effect of accident circumstances. Data used in this study are extracted from BAAC (*Bulletin d'Analyse des Accidents Corporels*) files, the French census of road crashes.

Keywords. road crash study, multiple correspondence analysis, log-linear model, ordinal logistic regression. ...

1 Introduction

Le nombre d'accidents routiers a considérablement diminué en France depuis de nombreuses années, mais les accidents continuent de se produire. En 2017, 61 224 accidents corporels ont été relatés contre 58 352 en 2018 (ONISR, 2019). Cette baisse de 4,7% est encourageante, cependant il est possible de diminuer d'avantage ces résultats. La mortalité, qui est l'une des conséquences de ces accidents, est de loin la motivation principale de cette étude statistique. Celle-ci est ciblée sur la région Franche-Comté, région qui démontre également une baisse de la mortalité de 2017 à 2018. En revanche à l'échelle départementale, les chiffres varient. En effet le taux de mortalité pour le département du Doubs a augmenté de 3% de 2017 à 2018, alors que pour les départements du Jura, Haute-Saône et Territoire de Belfort ce taux a diminué respectivement de 65%, 45% et 50% (ONISR, 2019).

Ce travail est à l'origine d'une collaboration avec la Gendarmerie Nationale de Besançon dont les objectifs sont de comprendre les causes des accidents pour une meilleure prévention des accidents routiers. Etablir un cadre spatio-temporel des accidents survenant en Franche-Comté leur permettrait de comprendre comment ils se produisent et d'anticiper au mieux leur arrivée. Cette étude vise à trouver et décrire les facteurs propices à la survenue d'un accident ou à la finalité de celui-ci (grave ou mortel). Ainsi des mesures pourront être proposées afin de mobiliser la conscience d'acteurs à différents niveaux (conducteurs, forces de l'ordre ou bien même municipalités).

Pour tenter de répondre à cette problématique, nous avons réalisé dans un premier temps une analyse de correspondances multiples (ACM) afin de dresser un profil temporel d'une part et spatial d'autre part. Dans un deuxième temps, un modèle log-linéaire a été ajusté afin d'étudier les dépendances entre certaines variables qualitatives comme le niveau de gravité des accidents ou les substances consommées par les conducteurs par exemple. Enfin, le niveau de gravité des accidents a été modélisé par une régression logistique ordinale afin de quantifier les effets propres des facteurs décrivant ces accidents.

2 Matériels et méthodes

2.1 Accidentologie de la région Franche-Comté

L'étude porte sur les accidents routiers de Franche-Comté s'étant produits entre le 1er Janvier 2005 et le 31 Décembre 2018. Une observation du tableau de données correspond à un accident corporel et est décrit par 15 variables qualitatives: – *type_acc*: gravité de l'accident (léger, grave ou mortel) – *substance*: substance consommée (drogue, alcool ou aucune) – *season*: saison de l'année où l'accident s'est produit (printemps, été, automne, hiver) – *week*: moment de la semaine où l'accident s'est produit (semaine ou week-end) – *daytime*: moment de la journée où l'accident s'est produit (journée ou nuit) – *time*: heure à laquelle l'accident s'est produit (entre 7 heures et 10 heures, entre 11 heures

et 15 heures, entre 16 heures et 19, entre 20 heures et 23 heures ou entre minuit et 6 heures) – *department*: département de la Franche-Comté où l'accident s'est produit (Doubs, Jura, Haute-Saône ou Territoire de Belfort) – *canton*: canton de la Franche-Comté où l'accident s'est produit (au nombre de 62 dans la région) – *weather*: conditions météorologiques de l'accident (normales ou autres) – *area*: zone où l'accident s'est produit (agglomération ou hors agglomération) – *intersection*: témoigne de si l'accident s'est produit à une intersection ou non – *obstacle*: le type d'obstacle heurté lors de l'accident (véhicule, piéton, autre ou aucun) – *collision*: le type de collision lors de l'accident (usuelle, autre ou aucune) – *shape_road*: la forme de la route sur laquelle l'accident s'est produit (rectiligne ou courbe) – *type_road*: le type de route sur laquelle l'accident s'est produit (communale, départementale, nationale, autoroute ou autre).

2.2 Méthodes statistiques et résultats

Analyse des Correspondances Multiples

L'Analyse des Correspondances Multiples (ACM) est une méthode efficace pour analyser des tableaux croisant des individus en lignes et des variables qualitatives en colonnes. Cette méthode d'analyse multidimensionelle vise à évaluer les relations existantes entre les variables et à examiner les associations entre les différentes modalités de ces variables. Des plans factoriels (représentations graphiques) sont construits sur lesquels on peut visuellement observer les proximités entre les modalités des variables qualitatives et les observations, synthétisant ainsi l'information. Pour plus de détails, lire Escofier et Pagés (2008) (chapitre 4) ou Husson et al. (2016) (chapitre 3).

L'ACM a permis d'établir des "profils" d'accidents routier au niveau temporel et spatial. Concernant le cadre temporel, plusieurs contrastes ont été mis en évidence: – les accidents se produisant en semaine s'opposent à ceux se produisant durant le week-end – les accidents survenant en été s'opposent à ceux survenant en hiver.

Du point de vue du cadre spatial, plusieurs cantons ont été liés à différentes caractéristiques: le canton de Baume-les-Dames est associé au caractère mortel – le canton de Besançon s'associe également au caractère mortel lorsque le temps est mauvais – ou bien encore, le canton de Dole est associé à la consommation de drogue.

Modèlisation log-linéaire

Lorsque les données se présentent sous la forme d'une table de contingence obtenue par le croisement de plusieurs variables qualitatives, le modèle log-linéaire se révèle être une méthode efficace pour modéliser les effectifs contenus dans les cellules de cette table. Il s'agit d'un modèle linéaire généralisé (GLM) décrivant les associations et les interactions entre plusieurs variables qualitatives. Il permet d'établir plusieurs types d'associations entre les variables. Pour plus de détails, voir Agresti (2002) (chapitre 8).

Un modèle log-linéaire a été ajusté avec les variables suivantes: *type_acc*, *substance* et *department*. Il ne subsiste aucune interaction de troisième ordre, les distributions des deux autres ne diffèrent pas quelques soient les niveaux de la troisième variable. Ce modèle nous permet de conclure, d'autre part, sur différents types de dépendance entre ces trois variables.

Régression logistique ordinale

La régression logistique est une méthode permettant d'analyser les dépendances entre une variable de réponse (binaire ou à plusieurs modalités) et des variables explicatives (quantitatives et/ou qualitatives). On parle de régression ordinale lorsqu'il existe un ordre hiérarchique naturel entre les modalités de la variable de réponse. L'ajustement de ce modèle permet de dire si une variable explicative particulière influe sur la variable de réponse ou non. Nous sommes également capables de quantifier l'effet d'une variable sur la variable de réponse. Pour plus de détails, lire McCullagh and Nelder (1989, chapitre 5), Agresti (2002, chapitre 7) ou Hothorn and Everitt (2014, chapitre 7).

Une régression ordinale a été ajustée sur les données de la Franche-Comté. La variable de réponse est *type_acc* dont les niveaux dans leur ordre hiérarchique sont "slight_safe", "serious", et "fatal" (respectivement léger, grave et mortel). Après une sélection de variables, le modèle final informe que les variables *time*, *substance*, *department*, *collision* et *area* influent sur la variable de réponse. Les odds ratio ont permis de quantifier les effets propres de chacune de ces variables. Ce modèle conclut donc que lorsqu'une substance est consommée (alcool, drogue ou les deux), le risque que l'accident en question soit plus grave augmente fortement. D'autres caractéristiques augmentent ce risque comme le département, le type de collision ou la zone.

3 Conclusion

L'objectif principal de cette étude était de pouvoir dresser un profil temporel et spatial des accidents se produisant en Franche-Comté. Cet aspect permettrait à la Gendarmerie Nationale de Besançon de comprendre au mieux comment les accidents se produisent et ainsi d'anticiper leur survenue. D'un point de vue temporel, un plus grand nombre d'accidents se qualifiant de grave ou mortel se produisent durant le week-end, entre 20 heures et 6 heures. D'un point de vue spatial, plusieurs cantons ont été identifiés et associés à certaines caractéristiques comme la consommation de substance (alcool et/ou drogue). De plus, la consommation de substance s'est révélée être le facteur le plus influent sur la finalité d'un accident corporel. Les mesures qui peuvent être entreprises, pour prévenir au mieux les accidents ou leurs conséquences, seraient de concentrer d'avantage les interventions aux périodes mises en évidence dans cette étude et de mobiliser la conscience des conducteurs au regard de la consommation de substance dans les cantons qui s'y associent le plus.

L'étude se poursuit actuellement avec l'ajout de facteurs supplémentaires comme les coordonnées gps des accidents qui permettraient d'être plus précis que l'échelle du canton pour l'analyse spatiale. L'exploration d'autres méthodes statistiques est également envisagée pour s'adapter aux nouvelles données mises à disposition.

Bibliographie

Agresti, A. (2002), *Categorical Data Analysis*. Wiley, second edition.

Escofier, B. et Pagès, J. (2008). *Analyses factorielles simples et multiples*. Dunod, 4ème édition.

Hothorn, T. and Everitt, B. S. (2014). *A Handbook of Statistical Analyses Using R*. CRC Press, third edition.

Husson, F., Lê, S. et Pagès, J. (2016). *Analyses des données avec R.*. Presses Universitaires de Rennes, 2ème édition.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman & Hall/CRC, second edition.

ONISR (2019). *La sécurité routière en France – Bilan de l'accidentalité de l'année 2018*.

Sur la construction et le pouvoir prédictif du classement Elo

Paul Steffen [1,2] & Léo Gerville-Réache [2]

[1] *Betclic 117 Quai de Bacalan, 33300 Bordeaux, p.steffen@betclicgroup.com*
[2] *Univ. Bordeaux, CNRS, IMS, UMR 5218, 33405 Talence, leo.gerville-reache@u-bordeaux.fr*

Résumé. Dans de nombreuses disciplines sportives, les classements sont utilisés comme un outil de comparaison entre les participants à une compétition. Le classement Elo, initialement utilisé dans les échecs, a prouvé par ses nombreuses applications qu'il parvenait à donner une image fidèle du niveau de jeu des compétiteurs. Dans cette communication, nous utilisons les adaptations de ce système de classement, pour prédire l'issue des rencontres de football, de tennis et de basketball.

Mots-clés. Elo, classement, notation, prédiction statistique des résultats sportifs

Abstract. In many sports, rankings are used as a comparison tool between participants in a competition. The Elo rating, originally used in chess, has proven by its many applications that it gives a true picture of the competitors' strength. In this communication, we use adaptations of this rating system to predict the outcome of football, tennis and basketball matches.

Keywords. Elo, ranking, rating, sports forecasting

1 Introduction

En 1960, la Fédération américaine des échecs adopte la procédure de classement des joueurs proposée par Arpad Elo. Cette méthode de classement rend possible une évaluation quantitative du niveau de jeu des participants à une compétition. Utilisé depuis pour le jeu de go ou encore le e-sport, le classement Elo a été aussi employé pour classer les photos de profil des membres de la 1ère version de The Facebook.
En sport, chaque fédération utilise une méthode de classement spécifique. Les méthodes évoluent régulièrement et sont souvent éloignées de la proposition d'Arpad Elo.

Dans cette communication, nous revenons sur les principes du classement Elo et questionnons son emploi dans le cadre de l'estimation des probabilités de victoire, de défaite et de nul d'une rencontre sportive. Aussi, après avoir rappelé les mécanismes du classement Elo et ses adaptations aux spécificités sportives, nous appliquons cette procédure au tennis, basket et football et discutons de la pertinence d'un tel classement pour estimer les probabilités des différentes issues possibles d'une rencontre sportive.

2 Le classement, un indicateur de niveau de jeu

Dans le monde du sport, les classements peuvent avoir de nombreux objectifs. De manière évidente, ces derniers permettent d'obtenir une indication sur le niveau actuel et la progression des compétiteurs. Ils permettent également aux fédérations de motiver leurs licenciés à participer à des compétitions. Le tennis, par exemple, utilise son propre système de classement, dans les 2 associations professionnelles que sont l'ATP[1] et le WTA[2].

2.1 Le classement Elo

Basé sur les performances antérieures des compétiteurs, ce système de classement assigne à chaque participant un nombre de points, tel que 2 compétiteurs ayant le même niveau de jeu ont le même nombre de points. Plus un compétiteur a un niveau élevé, plus il aura de points. Mis à part des ajustements, spécifiques aux cas d'application, ce classement est un jeu à somme nulle où plus les adversaires d'une rencontre ont une différence de points importante, plus le nombre de points échangés en cas de victoire du favori sera faible et inversement si l'outsider remporte la partie. Ainsi, le score Elo moyen d'un ensemble de joueurs en compétition est égal à la valeur initial du système de classement.
Dans ce classement, la performance d'un compétiteur est modélisée par une variable aléatoire centrée sur la notation de ce dernier. Initialement, la distribution de cette variable suivait une loi normale, avant de suivre une loi logistique permettant des queues de distribution plus lourdes.
Les points des participants à une rencontre sont mis à jour à l'issue de leur opposition de la manière suivante:

$$elo'_i = elo_i + K \cdot G(O - E) \tag{1}$$

où l'issue espérée de la rencontre est E et le résultat observé est $O \in \{0, 0.5, 1\}$, correspondant respectivement à la défaite, au match nul et à la victoire.
Le paramètre K est un paramètre d'échelle concernant l'impact des résultats les plus récents sur le Elo. Ce paramètre déterminant la volatilité du classement, peut être augmenté pour les joueurs ayant disputé peu de rencontres, afin d'entraîner des changements plus importants et une progression vers le niveau réel plus rapide.
Le paramètre G est aussi un paramètre d'échelle. Ce dernier n'apparaît pas dans la forme initiale de la formule d'actualisation des points Elo. Il a été ajouté afin de s'adapter aux spécificités des différents cas d'application de ce système de classement. Il peut permettre d'identifier l'importance d'une rencontre, ou une marge de victoire importante afin d'augmenter le nombre de points échangés lors de la rencontre.
Le paramètre E, d'issue espérée, s'exprime à l'aide de la relation suivante:

$$E = \frac{1}{1 + 10^{\frac{-dr_A}{400}}} \tag{2}$$

[1] Association of Tennis Professionals
[2] Women's Tennis Association

où $dr_A = elo_A - elo_B$ est la différence de points entre les 2 compétiteurs. Plus ce résultat sera proche de 1, plus l'équipe A sera favorite, et inversement.

2.2 Les spécificités associées aux sports étudiés

Créé par Arpad Elo, ce classement fût imaginé pour les échecs et son efficacité a encouragé son utilisation dans de tout autre domaines.
Ainsi, les aplications concernant des sports comme le football ou le basket ont intégré l'avantage induit par le fait de jouer à domicile, ou l'augmentation de points échangés pour une victoire avec une marge importante. Concernant le tennis, des ajustements ont été effectués pour l'importance du match, et le type de surface sur lequel le match est joué.
On différenciera alors les spécificités influençant l'issue de la rencontre (2), de celles n'intervenant que lors de l'évolution des points Elo suite à une rencontre (1).
L'avantage à domicile, présent dans le football et le basketball, sera incorporé par l'ajout d'une constante dans la différence des points des opposants, en faveur de l'équipe à domicile, de la manière suivante: $dr_A = elo_A - elo_B + HomeAdvantage$.
Pour le tennis, l'ajustement concernant la surface sur laquelle se dispute le match, se manifeste par une moyenne pondérée entre les points Elo spécifiques à la surface, et les points Elo de toutes les surfaces: $elo = \lambda \cdot elo_{global} + (1 - \lambda) \cdot elo_{surface}$.
Enfin, l'amplification de l'évolution des points selon la différence de score au football s'effectuera de la manière suivante: si $\Delta_{goals} <= 1$, alors $G = 1$, si $\Delta_{goals} = 2$, alors $G = 1,5$, sinon $G = \frac{11+\Delta_{goals}}{8}$. Pour le basketball, Silver, N. et Fischer-Baum R. (2015) proposent $G = \frac{(\Delta_{score}+3)^{0.8}}{7.5+0.006 \cdot dr_A}$.
Concernant le tennis, G sera influencé par l'importance du match à l'aide d'un coefficient, et selon le classement du joueur à l'aide de la formule suivante proposée par Cekovic, M. (2015): $G = 1 + \frac{18}{1+2^{(rating-1500)/63}}$

3 Prévision de l'issue d'une rencontre

Un système de classement comme celui du Elo représente alors un double intérêt pour une modélisation statistique de l'issue d'une rencontre: il peut être utilisé directement pour modéliser la probabilité d'issue de la rencontre, ou pour générer une variable, représentant le niveau de performance du compétiteur, utile à une modélisation plus complète, utilisant d'autres variables.

3.1 Du classement Elo à la probabilité de résultat

Bien que la formule d'actualisation des points Elo (1) intègre la possibilité d'un match nul ($0 = 0,5$), l'éventualité que cet évènement survienne n'apparaît pas dans la relation

(2).
Afin de construire la probabilité de match nul, possible lors d'un match de football, la densité de probabilités empirique de l'issue des rencontres des 5 championnats majeurs européens de 1993 à 2020 a été utilisée. Ces matchs ont été ordonnés selon la différence de cotes du bookmaker Bet365 entre la victoire de l'équipe à domicile et celle de l'équipe à l'extérieur dans l'histogramme groupé de la Figure 1.
La modélisation de ces résultats empiriques, visibles sur les courbes de la Figure 1, a été effectuée à l'aide d'un classement par régression logistique ordinale inspiré de celui proposé par Lasek, J. (2016) où les probabilités d'occurence des évènements sont déterminées de la manière suivante:

$$P(A) = \frac{1}{1+10^{c-\frac{dr_A}{400}}} \qquad P(B) = 1 - \frac{1}{1+10^{-c-\frac{dr_A}{400}}}$$
$$P(draw) = \frac{1}{1+10^{-c-\frac{dr_A}{400}}} - \frac{1}{1+10^{c-\frac{dr_A}{400}}} \tag{3}$$

où $c \in R_+$ est le paramètre déterminant la part des matchs nuls, évalué par une minimisation de l'entropie croisée, tout comme le $HomeAdvantage$ incorporé à la différence de points Elo.

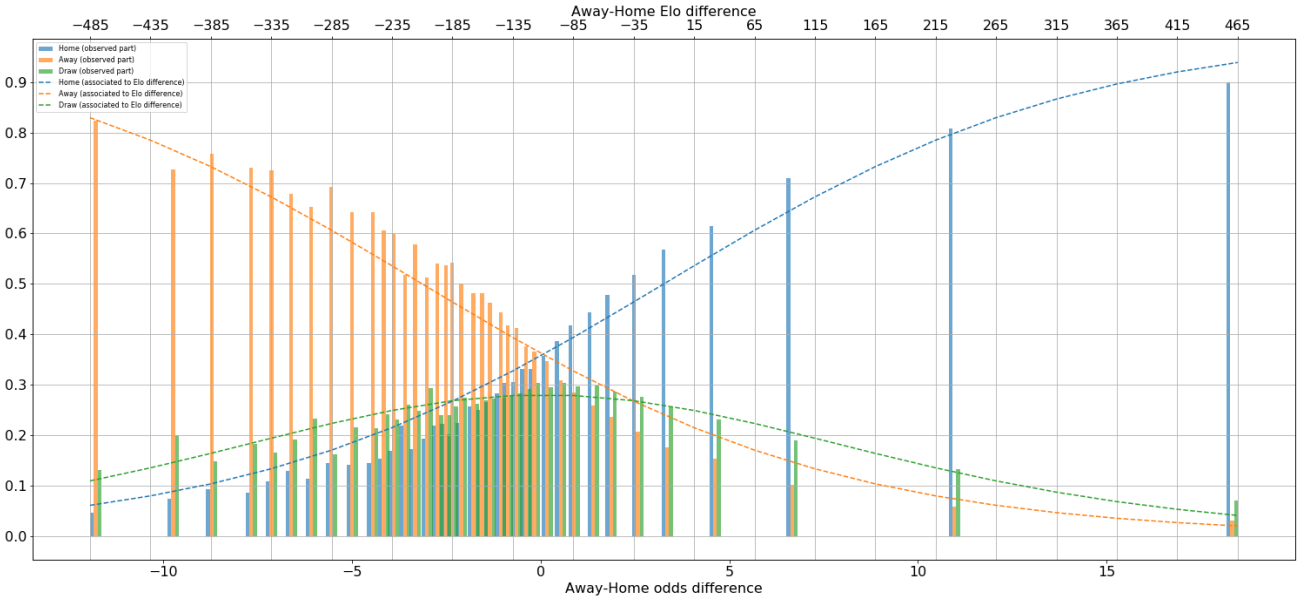


Figure 1: Part des issues des rencontres de football des 5 championnats majeurs européens de 1993 à 2020 et probabilités d'issue associées à une différence de points Elo

4 Application aux principaux championnats de chaque discipline

A l'aide des résultats des rencontres des 5 championnats majeurs européens de football depuis la saison 2009/2010, des matchs du championnat de basketball NBA depuis la saison 2000/2001, et des résultats du circuit de tennis ATP depuis 1991, 3 systèmes de classement Elo ont été evalués.

4.1 Méthode d'évaluation

Dans un objectif d'estimer les probabilités des différentes issues possibles d'une rencontre sportive, 2 métriques d'évaluation sont utilisées. La principale est la perte logistique, permettant de prendre en compte l'incertitude d'une prédiction à l'aide de sa différence avec le label observé:

$$L_{\log}(Y, P) = -\log \Pr(Y|P) = -\frac{1}{N} \sum_{i=0}^{N-1} \sum_{k=0}^{K-1} y_{i,k} \log p_{i,k} \tag{4}$$

où N est le nombre de rencontres, K appartient à l'ensemble des labels d'issue de la rencontre $\{Home, Away, Draw\}$, $p_{i,k}$ est la probabilité d'issue de la rencontre, calculée à l'aide des relations (3.1) et $y_{i,k} = 1$ si le résultat observé a le label k et 0 sinon.
Enfin, l'exactitude, définie comme le nombre de prédictions dont la valeur prévue est égale à la valeur réelle, permet une meilleure interprétabilité des résultats.
Ces métriques ont été calculées sur l'ensemble des rencontres de la saison 2017/2018, pour les 3 sports.

4.2 Résultats

Afin de mettre en perspective les résultats d'une prédiction à l'aide de la différence de points Elo de 2 opposants lors d'une rencontre sportive, il a été choisi d'utiliser les prédictions issues d'une classification naïve bayésienne, nécessitant comme le Elo, peu de données relatives à la rencontre (équipe à domicile pour le football et le basketball, et joueur le mieux classé au classement ATP pour le tennis).

Model	Log-loss	Accuracy
Soccer Elo Ranking	0.98	0.53
Soccer Naive Bayesian	1.07	0.45
Tennis Elo Ranking	0.63	0.65
Tennis Naive Bayesian	0.66	0.64
Basket Elo Ranking	0.62	0.65
Basket Naive Bayesian	0.68	0.59

Table 1: Métriques d'évaluation calculées sur la saison 2017/2018

4.3 Discussion

A la vue de nos résultats, on remarque un faible gain de performance pour le tennis, s'expliquant par un bon niveau d'information contenu dans le classement ATP afin de déterminer l'issue d'une rencontre.
Dans notre objectif de proposer une prédiction statistique de l'issue d'une rencontre sportive, le Elo va plus loin qu'une simple comparaison des rangs entre les opposants au sein d'un classement. Par l'utilisation de la différence de points Elo, la prédiction est basée sur une comparaison relative de la force des opposants.
Cependant, de nombreux paramètres déterminant cette force (composition initiale d'une équipe, conditions climatiques, style de jeu etc...) ne sont pas utilisés. Et, bien que le système Elo permette une amélioration allant jusqu'à 0.1 point d'exactitude dans la prédiction de l'issue d'une rencontre sportive, par rapport à un modèle naïf, l'utilisation de points Elo semble plus pertinente au sein d'un modèle d'apprentissage statistique plus élaboré.

Bibliographie

Lasek, J. (2016). EURO 2016 Predictions Using Team Rating Systems. MLSA@PKDD/ECML.

Lasek, J. et Szlávik, Z. et Bhulai, S. (2013). The predictive power of ranking systems in association football. Int. J. of Applied Pattern Recognition. 1. ied Pattern Recognition. 10.1504/IJAPR.2013.052339.

EloRatings.net (2012). The World Football Elo Rating System.
http://www.eloratings.net/about

Lacy, S. (2018). Implementing an Elo rating system for European football.
https://stuartlacy.co.uk/2017/08/31/implementing-an-elo-rating-system-for-european-football/

Silver, N. et Fischer-Baum R. (2015). How we calculate NBA Elo Ratings.
https://fivethirtyeight.com/features/how-we-calculate-nba-elo-ratings/

FIFA.com (2018). Revision of the FIFA / Coca-Cola World Ranking.
http://www.fifa.com/worldranking/procedureandschedule/menprocedure/index.html

Cekovic, M. (2015). Tennis Crystal Ball.
https://github.com/mcekovic/tennis-crystal-ball

FootballDatabase. FootballDatabase methodology. https://footballdatabase.com/methodology.php

Quantification Robuste de l'Incertitude d'une Mesure de Risque Issue d'un Code de Calcul

Jérôme Stenger [1,2] & Fabrice Gamboa [2,3] & Merlin Keller [1] & Bertrand Iooss [1,2]

[1] *EDF R&D, 6 quai Watier 78401 Chatou, France. jerome.stenger@edf.fr.*
[2] *Université Paul Sabatier, 118 route de Narbonne 31400 Toulouse, France.*
[3] *Artificial and Natural Intelligence Toulouse Institute (ANITI), France*

Résumé. La quantification d'incertitudes a pour but d'évaluer l'impact d'un manque de connaissance des paramètres d'entrées (considérés aléatoires) sur les résultats d'une expérience numérique. Dans ce travail, nous prenons en compte un second niveau d'incertitude qui affecte le choix du modèle probabiliste des paramètres d'entrées. Nous évaluons les bornes d'une quantité d'intérêt sur l'ensemble des mesures de probabilités uniquement définies par leur bornes et certains de leurs moments. Du fait du grand nombre de contraintes l'optimisation numérique est complexe. Nous montrons que le problème d'optimisation peut se paramétriser sur les points extrémaux de cet espace de mesures de probabilité contraintes. De plus nous proposons une nouvelle paramétrisation libre de contraintes basées sur les moments canoniques.

Mots-clés. quantification d'incertitude, analyse de robustesse, moment canonique ...

Abstract. In uncertainty quantification studies, a major topic of interest lies in assessing the uncertainties tainting the results of a computer simulation. In this work we seek to gain robustness on the quantification of a risk measurement by accounting for all sources of uncertainties tainting the inputs of a computer code modeled as random variables. To that end, we evaluate the maximum of a quantity of interest over a class of bounded distributions satisfying moments constraint. Two options are available when dealing with such complex optimization problems : one can either optimize under constraints ; or preferably, one should reformulate the objective function. We identify a well-suited parameterization to compute the maximal quantile based on the theory of canonical moments. It allows an effective, free of constraints, optimization. This methodology is applied to an industrial computer code related to nuclear safety.

Keywords. uncertainty quantification, robustness analysis, canonical moments ...

1 Introduction

La quantification des incertitudes en contexte industriel est réalisée en modélisant les paramètres d'entrées d'un système physique par des variables aléatoires $\mathbf{X} = (X_1, \ldots, X_d) \sim \mu$. Afin de propager les incertitudes affectant les entrées, un modèle de simulation numérique

G est appelé avec différentes combinaisons de paramètres d'entrées générés suivant leur loi de probabilité [1]. Il est alors possible d'étudier la variabilité de la sortie du code $Y = G(X_1, \ldots, X_d)$, ou d'estimer certaines quantités d'intérêt spécifique. Une quantité d'intérêt est une grandeur statistique sur la sortie Y, comme par exemple une moyenne, une probabilité de dépassement de seuil, ou bien encore un quantile [2]. Le code étant considéré comme une boîte noire déterministe, la quantité d'intérêt dépend uniquement du choix de la loi de probabilité des entrées $\mathbf{X} \sim \mu$. Elle peut donc être vue comme une fonction scalaire définie sur l'ensemble des mesures de probabilité.

La distribution de probabilité μ qui modélise l'incertitude des paramètres d'entrées est elle-même incertaine. En général, cette mesure de probabilité est choisie grâce aux avis d'experts, qui sont subjectifs et parfois contradictoires, mais aussi grâce à des données expérimentales souvent en nombre insuffisant et entâchées d'erreurs. De ce fait, le choix du modèle probabiliste pour les variables d'entrée est lui-même incertain. Cette variabilité dans le choix de la distribution, qui peut être vue comme une incertitude de deuxième niveau, se propage jusqu'à la quantité d'intérêt. Ainsi, deux choix de lois de probabilités différentes en entrée donnent lieu à deux valeurs différentes de la quantité d'intérêt. Notre objectif est donc d'être robuste vis à vis de la variabilité du choix de la loi de probabilité des entrées [3].

Nous présenterons dans la Section 2 la méthodologie mise en place afin de prendre en compte l'ensemble des sources d'incertitudes sur la mesure de probabilité d'entrée. Dans la Section 3 nous mettons en avant les difficultés numériques et nous introduisons une nouvelle paramétrisation qui relaxe les contraintes du problème d'optimization. Enfin dans la Section 4 nous présentons une application numérique de nos résultats sur un cas industriel réel.

2 Analyse de Robustesse en Quantification d'Incertitudes

Afin de prendre en compte cette incertitude de deuxième niveau, nous proposons d'évaluer des bornes sur la quantité d'intérêt. De ce fait on ne considère plus une distribution fixée μ, mais un ensemble de mesures de probabilité $\mathcal{A}$ sur lequel la quantité d'intérêt est optimisée. Il s'agit donc bien d'une optimisation sur un espace de mesure de probabilité *a priori* de dimension infinie et non paramétrique. Fort heureusement, la solution de ce problème d'optimisation est donnée grâce à un théorème de réduction [4]. Nous proposons ici une généralisation du théorème de réduction issue des travaux de [3] qui introduit cette méthode sous le nom de d'Optimal Uncertainty Quantification (OUQ). Sous l'hypothèse où la quantité d'intérêt est une fonction quasi-convexe et semicontinue inférieurement de la mesure de probabilité $\mu \in \mathcal{A}$, nous avons montré que le maximum de la quantité d'intérêt est atteint sur les points extrémaux de l'espace de mesure de probabilité $\mathcal{A}$ lorsque celui-ci est convexe et compact (l'hypothèse de compacité peut être

relaxée).

Le choix de l'espace de mesure $\mathcal{A}$ est donc important, en particulier parce qu'il nous faut connaître ses points extrémaux pour pouvoir appliquer le théorème de réduction. Dans ce travail nous nous intéressons à l'ensemble de toutes les mesures de probabilité (unimodales) sous contraintes de moments. Cet espace s'appelle la classe de moments (respectivement la classe de moments unimodales). Plus précisément, nous nous intéressons à l'ensemble des distributions de probabilités (unimodales) qui vérifient certains moments, la classe des moments s'écrit :

$$\mathcal{A} = \left\{ \mu = \mu_1 \times \cdots \times \mu_d \in \prod_{i=1}^{d} \mathcal{P}([l_i, u_i]) \mid \mathbb{E}_{\mu_i}[X^j] = c_i^{(j)} \ , \text{ pour } 1 \leq j \leq N_i \text{ et } 1 \leq i \leq d \right\} ,$$

où $\mathcal{P}([l_i, u_i])$ dénote l'ensemble des mesures de probabilité sur l'intervalle $[l_i, u_i]$. Un élément μ dans la classe de moment $\mathcal{A}$ est simplement une mesure d-dimensionnelle qui modélise le vecteur aléatoire d'entrée $\mathbf{X} = (X_1, \ldots, X_d)$ du code G. Le produit des mesures marginales $\mu_i \sim X_i$ signifie que chaque paramètre d'entrée est indépendant des autres, et chaque distribution μ_i est bornée entre $[l_i, u_i]$ avec ses N_i premiers moments fixés. Il est plutôt habituel d'imposer des bornes sur les variables aléatoires puisqu'elles représentent des paramètres physiques dont les valeurs ne peuvent pas tendre vers l'infini. Le choix d'imposer certains moments sur les distributions de chaque variable est relativement adapté au contexte industriel. En effet, les mesures de probabilité sont en pratique souvent choisies avec pour unique connaissance une valeur moyenne et une variance. Nous n'avons pas décrit la classe des moments unimodales mais il suffit simplement d'ajouter l'unimodalité des distributions marginales μ_i.

La classe des moments et la classe des moments unimodales ont une structure topologique très intéressante puisqu'il est possible de caractériser leurs points extrémaux. Dans le cas de la classe des moments les points extrémaux sont les mesures de probabilité qui s'écrivent comme une combinaison convexe de masses de Dirac. Plus spécifiquement, si N_i contraintes de moments sont imposées à la mesure μ_i, alors les points extrémaux sont les mesures discrètes supportées par au plus $N_i + 1$ points. L'ensemble des points extrémaux de la classe des moments $\mathcal{A}$ s'écrit ainsi

$$\Delta = \left\{ \mu \in \mathcal{A} \mid \mu_i = \sum_{k=1}^{N_i+1} \omega_i^{(k)} \delta_{x_i^{(k)}} \ , \text{ où } x_i^{(k)} \in [l_i, u_i] \text{ pour } 1 \leq i \leq d \right\} .$$

Ces propriétés ont été étudiés par Winkler [5] et s'inspirent de résultats de la théorie de Choquet [6]. L'intérêt de réduire l'optimisation de la quantité d'intérêt aux points extrémaux est que cela permet de reparamétriser le problème qui ne dépend maintenant uniquement que des poids $\omega_i^{(j)}$ et des points du support $x_i^{(j)}$ des mesures discrètes. Dans le cas de la classe de moments unimodales nous avons montré que les points extrémaux sont très similaires, mais à la place de masses de Dirac, ils sont combinaisons convexes de lois uniformes dont l'une des bornes se confond avec le mode de la distribution [7].

3 Approche Numérique par les Moments Canoniques

L'un des principaux problèmes reste la complexité numérique de l'optimisation de la quantité d'intérêt, même lorsque le problème est réduit sur les points extrémaux. En effet, une optimisation globale est réalisée afin de trouver l'optimum, mais cela nécessite d'explorer efficacement l'espace d'optimisation. Or il n'est pas évident de générer des mesures discrètes $\mu_i = \sum_{k=1}^{N_i+1} \omega_i^{(k)} \delta_{x_i^{(k)}}$ (c'est-à-dire générer des poids et des positions) qui vérifient le système de contraintes de moments suivants :

$$\left\{ \begin{array}{lcccll} \omega_i^{(1)} & + & \dots & + & \omega_i^{(N_i+1)} & = 1 \\ \omega_i^{(1)} x_i^{(1)} & + & \dots & + & \omega_i^{(N_i+1)} x_i^{(N_i+1)} & = c_i^{(1)} \\ \vdots & & & & \vdots & \vdots \\ \omega_i^{(1)} {x_i^{(1)}}^{N_i} & + & \dots & + & \omega_i^{(N_i+1)} {x_i^{(N_i+1)}}^{N_i} & = c_i^{(N_i)} \end{array} \right. \tag{1}$$

Des algorithmes de programmation semi-définie [8] ont été proposé par Lasserre [9] mais le solveur déterministe atteint rapidement ses limites lorsque la dimension du problème augmente. D'autres approches basées sur des algorithmes d'optimisation stochastique ont été implémentées dans une toolbox appelée *Mystic* Framework [10] qui intègre entièrement le principe de l'OUQ. Toutefois, la toolbox est très générale et l'optimisation sur la classe des moments (unimodales) peut être améliorée. C'est ce que nous proposons dans une nouvelle approche basée sur les moments canoniques [11]. Les moments canoniques peuvent se voir comme la version normalisée de la suite de moments d'une mesure. Pour cela, on définit les valeurs maximale et minimale du $(n+1)$ème moment quand les n premiers moments $(c_1, \dots, c_n)$ sont connus :

$$\begin{aligned} c_{n+1}^{+} &= \max \{ c \in \mathbb{R} : (c_1, \dots, c_n, c) \text{ est la suite de moment d'une mesure de probabilité } \} \ , \\ c_{n+1}^{-} &= \min \{ c \in \mathbb{R} : (c_1, \dots, c_n, c) \text{ est la suite de moment d'une mesure de probabilité } \} \ . \end{aligned}$$

Les moments canoniques s'écrivent alors

$$p_n = \frac{c_n - c_n^-}{c_n^+ - c_n^-} \ .$$

Cela définit une bijection entre la suite des moments et la suite des moments canoniques. Toutefois, les moments canoniques sont fondamentalement liés à la distribution, et possèdent des propriétés intéressantes puisqu'ils sont tous compris dans l'intervalle $[0, 1]$ et sont invariants par transformation affine du support de la mesure. Nous proposons ici de reparamétriser le problème d'optimisation avec les moments canoniques. Cette paramétrisation permet d'explorer efficacement l'essemble des points extrémaux de la classe des moments (unimodales). En effet, il existe un lien direct entre les moments canoniques et les positions du support d'une mesure discrète. Cela permet de générer les positions

$x_i^{(k)}$ d'une mesure μ_i à partir d'une séquence de moments canoniques (c'est-à-dire un vecteur de $[0,1]^{N_i+1}$). Finalement, l'espace d'optimisation, c'est-à-dire l'ensemble des mesures discrètes qui satisfont le système de contraintes (Equation (1)) est exploré facilement. Autrement dit, l'optimisation de la quantité d'intérêt sous cette nouvelle paramétrisation relaxe l'ensemble des contraintes du problème. Cela a pour effet d'augmenter significativement les performances des algorithmes d'optimisations stochastiques globaux utilisés par exemple dans la toolbox *Mystic*. Ces résultats ont pu être validés sur des modèles jouets ainsi que sur des cas industriels réels.

4 Applications

Nous considérons ici un modèle numérique qui reproduit la réponse thermohydraulique à transitoire de température dans le conduit primaire d'un réacteur nucléaire avec une brèche en branche froide. Le modèle retourne en sortie le pic de température de gaine lors du transitoire. Le code numérique prend un grand nombre d'entrée, les paramètres incertains considérés comme les plus influents sont au nombre de 9. Ils représentent des transferts thermiques ou des frottement interfaciaux dans plusieurs élements du circuit primaire. Les lois associées sont explicitées dans la dernière colonne du tableaux ci dessous, les variables étant toutes supposées indépendantes.

Table 1 – Moments et distributions des 9 variables plus influentes du code CATHARE [12].

Variable	Bornes	Moyenne	Deuxième moment	Distribution initiale (tronquée aux bornes)
$n°1$	$[0.1, 10]$	1.33	3.02	$LogNormal(0, 0.76)$
$n°2$	$[0, 12.8]$	6.4	45.39	$Normal(6.4, 4.27)$
$n°3$	$[11.1, 16.57]$	13.83	192.22	$Normal(13.79$
$n°4$	$[-44.9, 63.5]$	9.3	1065	$Uniform(-44.9, 63.5)$
$n°5$	$[0.1, 10]$	1.33	3.02	$LogNormal(0, 0.76)$
$n°6$	$[0.1, 10]$	1.33	3.02	$LogNormal(0, 0.76)$
$n°7$	$[0.235, 3.45]$	0.99	1.19	$LogNormal(-0.1, 0.45)$
$n°8$	$[0.1, 3]$	0.64	0.55	$LogNormal(-0.6, 0.57)$
$n°9$	$[0.1, 10]$	1.33	3.02	$LogNormal(0, 0.76)$

Nous considérons que les distributions de la table 1 sont incertaines. Nous supposons à la place que seuls les moments d'ordre un et deux (moyenne et variance) de chaque variable d'entrée sont connus. Le quantile de sortie du code est alors maximisé sur l'ensemble des lois de probabilité vérifiant ces moments et ces bornes, ce qui est équivalent à calculer l'enveloppe inférieure des fonctions de répartition $\inf_{\mu \in \mathcal{A}} F_\mu(T)$ [4] (voir Figure 1). Par

exemple le quantile à 95% est de 760°C en prennant pour loi d'entrée les lois de probabilité de la Table 1. Tandis que la prise en compte de l'incertitude sur la distribution fournie une borne supérieure sur le quantile égale à 788°C en paramétrisant le problème avec les moments canoniques. Le solveur Mystic qui devrait fournir les même résultats que notre méthode n'atteint pas la borne et fourni une valeur maximale de 770°C pour des paramètres de solveur identiques. Afin de converger vers le même optimum que le notre, il serait nécessaire d'utiliser beaucoup plus de puissance de calcul en raison de la mauvaise exploration de l'espace d'optimisation.

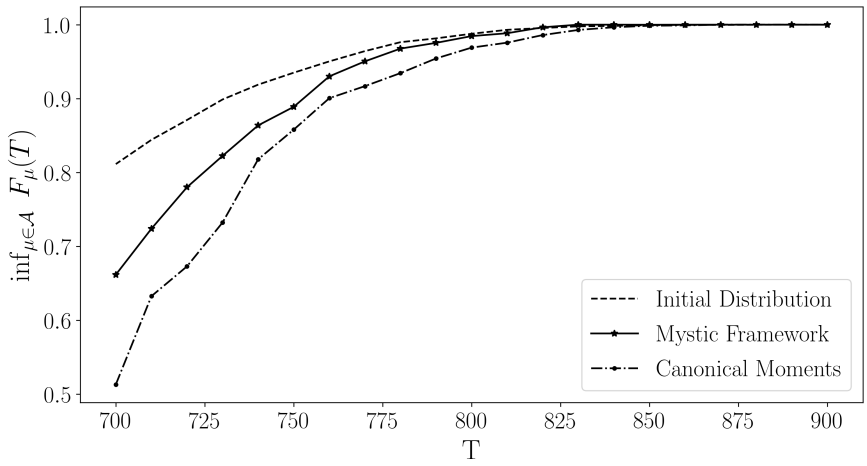


FIGURE 1 – L'optimisation fournie une borne maximale sur le quantile. Ainsi, quelque soit la loi de probabilité de la variable d'entrée vérifiant les contraintes de moment de la Table 1, le quantile de sortie est garanti d'être en dessous de cette borne. Nous comparons les résultats de notre paramétrisation avec celle de la toolbox Mystic en utilisant le même optimiseur global. La fonction de répartition initiale est calculée à partir de méthode Monte-Carlo, la méthode Mystic et moment canonique correspondent à l'optimisation robuste du quantile réalisée avec le même solveur à Evolution Différentielle. Il est clair que la paramétrisation de la toolbox Mystic ne permet pas d'atteindre efficacement l'optimum.

Références

[1] E. de Rocquigny. *Modelling Under Risk and Uncertainty : An Introduction to Statistical, Phenomenological and Computational Methods*. Wiley, 2012.

[2] G.B. Wallis. Uncertainties and probabilities in nuclear reactor regulation. *Nuclear Engineering and Design*, 237 :1586–1592, 2004.

[3] Houman Owhadi, Clint Scovel, Timothy John Sullivan, Mike McKerns, and Michael

Ortiz. Optimal Uncertainty Quantification. *SIAM Review*, 55(2) :271–345, January 2013. arXiv : 1009.0679.

[4] Jerome Stenger, Fabrice Gamboa, Merlin Kerller, and Bertrand Iooss. Optimal Uncertainty Quantification of a risk measurement from a thermal-hydraulic code using Canonical Moments. *International Journal for Uncertainty Quantification*, 2020.

[5] Gerhard Winkler. Extreme Points of Moment Sets. *Math. Oper. Res.*, 13(4) :581–587, November 1988.

[6] Gustave Choquet, Jerrold Marsden, and Stephen Gelbart. Lectures on analysis / Gustave Choquet. *SERBIULA (sistema Librum 2.0)*, July 2018.

[7] Jérôme Stenger, Fabrice Gamboa, and Merlin Keller. Optimization Of Quasi-convex Function Over Product Measure Sets. preprint, 2019.

[8] Didier Henrion, Jean-Bernard Lasserre, and Johan Löfberg. GloptiPoly 3 : moments, optimization and semidefinite programming. *Optimization Methods and Software*, 24(4-5) :761–779, October 2009.

[9] Jean-Bernard Lasserre. *Moments, positive polynomials and their applications.* Number v. 1 in Imperial College Press optimization series. Imperial College Press ; Distributed by World Scientific Publishing Co, London : Signapore ; Hackensack, NJ, 2010. OCLC : ocn503631126.

[10] M. McKerns, H. Owhadi, C. Scovel, T. J. Sullivan, and M. Ortiz. The optimal uncertainty algorithm in the mystic framework. *CoRR*, abs/1202.1055, 2012.

[11] Holger Dette and William J. Studden. *The Theory of Canonical Moments with Applications in Statistics, Probability, and Analysis.* Wiley-Blackwell, New York, September 1997.

[12] Bertrand Iooss and Amandine Marrel. Advanced methodology for uncertainty propagation in computer experiments with large number of inputs. *hal : 01907198*, October 2018.

Bayesian estimation of multivariate Hawkes processes: a graph perspective and a general approach to interaction modelling

Déborah Sulem [1] & Vincent Rivoirard [2] & Judith Rousseau [3]

[1] *Department of Statistics, University of Oxford, deborah.sulem@stats.ox.ac.uk*
[2] *CEREMADE, University Paris Dauphine, vincent.rivoirard@dauphine.fr*
[3] *Department of Statistics, University of Oxford & University Paris Dauphine, judith.rousseau@stats.ox.ac.uk*

Résumé. Les processus ponctuels temporels apparaissent dans de nombreuses applications pour modéliser des données ponctuelles, comme les événements sociaux, environnementaux ou neuronaux. Les processus multivariés de Hawkes sont particulièrement intéressants pour modéliser une dépendance par rapport au passé et des interactions entre entités complexes. Dans ce travail, nous nous intéressons, dans un contexte bayésien non-paramétrique, à deux approches émergentes dans le cadre fréquentiste: l'estimation du graphe d'interaction entre les dimensions du processus, et l'estimation de l'ensemble des paramètres dans un modèle général d'interactions par renforcement et inhibition. Nous prouverons d'abord, pour le modèle avec renforcement, la consistance de la distribution a posteriori pour le paramètre du graphe. Puis nous étudierons le comportement d'un estimateur minimisant une fonction de risque. Enfin, dans le modèle général, nous déterminerons une borne supérieure de la vitesse de concentration de la distribution a posteriori, pour des fonctions d'interactions de type histogrammes à nombre d'intervalles fini.

Mots-clés. Processus de Hawkes multivariés, estimation bayésienne non-paramétrique, graphe d'interaction, inhibition

Abstract. Temporal point processes appear in a large variety of applications such as earthquakes' modelling, social media interactions and neuronal activity. In this context, multivariate Hawkes processes are particularly interesting to model complex history dependencies and type of interactions. In this work, we study in the non-parametric Bayesian framework, two novel approaches that have gained attention in the frequentist literature: the estimation of the graph of interaction between dimensions of the process and the general model of mutual-excitation and inhibition. Firstly, in the original model of mutual-excitation, we will prove the consistency of the posterior distribution on the graph parameter. Secondly, we will study the consistency of a graph estimator minimizing a risk function. Finally, we will prove an upper-bound on the posterior concentration rate in the general model of interaction, for histogram interaction functions with a finite number of bins.

Keywords. Multivariate Hawkes processes, non-parametric Bayesian estimation, interaction graph, inhibition

1 Problem setting

We consider a probability space $(\mathcal{X}, \mathcal{G}, \mathbb{P})$ and multivariate point process $N = (N_t)_t = (N_t^1, \dots, N_t^K)$ in $K \in \mathbb{N}^\star$. Each component N_t^k records the number of events that have occurred on the k-th component until time t. The linear (mutually-exciting) Hawkes model with finite memory is defined as follows.

Let $f = ((\nu_k)_{k=1}^K, (h_{lk})_{k,l=1}^K)$ such that for all k, l, $\nu_k > 0$ and $h_{kl} : [0, A] \to \mathbb{R}^+$ integrable and $A > 0$. Let $(\mathcal{G}_t)_t$ and $\mathcal{G}$ such that $\mathcal{G}_t = \mathcal{G}_0 \vee \sigma(N_s, s \leq t)$ with $\mathcal{G}_t \subset \mathcal{G}$ and $\mathcal{G}_0 \subset \mathcal{G}$. $(N_t)_t$ is a linear Hawkes process with parameter f adapted to $(\mathcal{G}_t)_t$ if

i. almost surely, for all k, l, $(N_t^k)_t$ and $(N_t^l)_t$ never jump simultaneously;

ii. for all k the intensity process $(\lambda_t^k(f))_t$ of $(N_t^k)_t$ is given by

$$\lambda_t^k(f) = \nu_k + \sum_{l=1}^K \int_{t-A}^{t^-} h_{lk}(t-s) dN_s^l. \tag{1}$$

Under the conditions above, almost surely there exists a unique non-explosive pathwise process. If the matrix $\rho = (\rho_{lk})_{l,k=1}^K \in \mathbb{R}_+^{K \times K}$ with $\rho_{lk} = \|h_{lk}\|_1$, has a spectral radius strictly smaller than 1, there exists a unique stationary distribution with finite average intensity. The parameter spaces are defined as follows

$$\mathcal{H} = \{(h_{lk})_{l,k=1}^K; h_{lk} \text{ is integrable}, h_{lk} \geq 0, \|h_{lk}\|_\infty < \infty, \text{support}(h_{lk}) \subset [0, A], \forall k, l \leq K\},$$
$$\mathcal{F} = \{f = ((\nu_k)_{k=1}^K, (h_{lk})_{k,l=1}^K); 0 < \nu_k < \infty, \forall k \leq K, (h_{lk})_{lk} \in \mathcal{H}\}.$$

In the context of graph estimation, we reparametrize the process by introducing a graph parameter $\delta = (\delta_{lk})_{1 \leq l,k \leq K} \in \{0, 1\}^{K \times K}$ and for any k, l:

$$h_{lk}(t) = \delta_{lk} \tilde{h}_{lk}(t), \tag{2}$$

with $\tilde{h}_{lk} : [0, A] \to \mathbb{R}^+$ and $\tilde{h}_{lk} = 0$ if $\delta_{lk} = 0$.

The general Hawkes model does not constraint the interaction functions to be non-negative and is defined as follows:

$$\lambda_t^k(f) = \phi\left(\nu_k + \sum_{l=1}^K \int_{t-A}^{t^-} h_{lk}(t-s) dN_s^l\right)_+, \tag{3}$$

with $\forall l, k$ $h_{kl} : [0, A] \to \mathbb{R}$ and $\phi : \mathbb{R} \to \mathbb{R}^+$ a non-linear function. Typically, ϕ is Lipschitz and a classical choice is $\phi(x) = \max(x, 0)$.

We now introduce the Bayesian estimation framework. We assume that we observe a Hawkes process with true parameter $f_0 = ((\nu_k^0)_{k=1}^K, (h_{lk}^0)_{k,l=1}^K) \in \mathcal{F}$ on the time window $[0, T]$, $T > 0$, with spectral norm $\|\rho^0\| < 1$ and $\sigma(N_s, s < 0) \subset \mathcal{G}_0$. We denote δ^0 the matrix

of binary variables s.t. $\forall l,k \in [K], \delta^0_{lk} = 1 \iff h^0_{lk} \neq 0$. We denote $\mathbb{P}_0$ the stationary distribution of N and $\mathbb{P}_0(.|\mathcal{G}_0)$ its conditional distribution. $\mathbb{E}_0$ is the expectation associated to $\mathbb{P}_0$. For $f \in \mathcal{F}$, the log-likelihood has the following expression:

$$L_T(f) := \sum_{k=1}^{K} \left[\int_0^T \log(\lambda_t^k(f)) dN_t^k - \int_0^T \lambda_t^k(f) dt \right]. \tag{4}$$

The conditional probability distribution with parameter f is defined as

$$d\mathbb{P}_f(.|\mathcal{G}_0) = e^{L_T(f) - L_T(f_0)} \mathbb{P}_0(.|\mathcal{G}_0).$$

We define the L_1 distance and the stochastic pseudo-distance d_{1T} on the parameter space. For any $f, f' \in \mathcal{F}$,

$$\|f - f'\|_1 = \sum_{k=1}^{K} |\nu_k - \nu'_k| + \sum_{k=1}^{K} \sum_{l=1}^{K} \|h_{lk} - h'_{lk}\|_1,$$

$$d_{1T}(f, f') = \frac{1}{T} \sum_{k=1}^{K} \int_0^T |\lambda_t^k(f) - \lambda_t^k(f')| dt.$$

Finally, we consider a prior distribution $\Pi(f)$ on $\mathcal{F}$ and the pseudo-posterior distribution:

$$\Pi(B|N) = \frac{\int_B \exp(L_T(f)) d\Pi(f)}{\int_{\mathcal{F}} \exp(L_T(f)) d\Pi(f)}, \qquad B \subset \mathcal{F}.$$

2 Main results

2.1 Assumptions

Let $\mathcal{N}(u, \mathcal{H}, d)$ be the covering number of a set $\mathcal{H}$ w.r.t a metric d by balls with radius u. We assume that N is observed on $[-A, T]$ and let ϵ_T be a positive sequence such that $\epsilon_T = o(1)$ and $\log^3 T = o(T\epsilon_T^2)$. For $B > 0$, we consider:

$$B_\infty(\epsilon_T, B) = \left\{ f \in \mathcal{F}, \max_k |\nu_k - \nu_k^0| \leq \epsilon_T, \max_{l,k} ||h_{lk} - h_{lk}^0||_\infty \leq \epsilon_T \right\}. \tag{5}$$

We assume that the following conditions are satisfied for T large enough:

(A0) There exist $c_1 > 0$ and $B > 0$ such that

$$\Pi(B(\epsilon_T, B)) \geq \exp^{-c_1 T \epsilon_T^2}.$$

(A1) There exist subsets $\mathcal{H}_T \subset \mathcal{H}$, $\kappa > 0$, $\xi_0 > 0$ and $x_0 > 0$ such that

$$\frac{\Pi(\mathcal{H}_T^c)}{\Pi(B(\epsilon_T, B))} \leq \exp^{-(2\kappa+3) T \epsilon_T^2},$$

and

$$\log \mathcal{N}\left(\xi_0 \epsilon_T, \mathcal{H}_T, ||.||_1\right) \leq x_0 T \epsilon_T^2.$$

(A2) For all $u_0 > 0$ and T large enough,

$$\Pi\left(||\rho|| > 1 - u_0 (\log T)^{1/6} \epsilon_T | N\right) \leq e^{2c_1 T \epsilon_T^2}.$$

(A3) For all $1 \leq l, k \leq K$ and for all $C > 0$,

$$\Pi\left(\rho_{lk} \leq C \epsilon_T\right) = o\left(e^{-c_1 T \epsilon_T^2}\right).$$

Assumptions (A0)-(A1) come from the general method of Ghosal et al. [1] to derive posterior concentration rates. (A0) allows to derive the Kullback-Leibler condition - which guarantees that the prior distribution puts enough mass on the Kullback-Leibler neighborhood of the true parameter. The first part of (A1) imposes the existence of an increasing sequence of subsets of (the non-parametric part of) the parameter space - called sieves - on which it will possible to construct adequate statistical tests. The latter are designed to separate the true parameter from parameter that are far away, and need to have an exponential decay in type I and II errors. The second part of (A1) restricts the number of these tests by setting an upper bound on the complexity of the space of interaction functions. (A2) is specific to Hawkes point processes, which can become explosive when the spectral radius of ρ is equal to the critical value 1. This assumption imposes to stay relatively far away from this limit and allows to derive useful bounds on the number of events. Finally, (A3) is peculiar to the graph estimation problem and is sufficient to distinguish between null and non-null interaction functions.

2.2 Consistency of the posterior distribution on the graph

Proposition 2.1 *Under Assumptions (A0)-(A3), the posterior distribution is consistent on the graph parameter* $\delta \in \{0,1\}^{K \times K}$*:*

$$\mathbb{E}_0\left[\Pi\left(\{\delta \neq \delta^0\} | N\right)\right] \xrightarrow[T \to \infty]{} 0.$$

This result can be simplified in the following cases.

Proposition 2.2 *If the interaction functions are all the same i.e.* $\forall k, l, h_{lk}(t) = h(t)$ *or are only receiver-dependent i.e.* $\forall k, l, h_{lk}(t) = h_k(t)$*, Assumption (A4) can be replaced by the following condition:*

$$\textit{(A4)} \quad \forall C > 0, \qquad \Pi\left(\rho_{lk} \leq C \sqrt{\frac{\log T}{T}}\right) = o\left(\frac{1}{(\log T)^{K/2}}\right).$$

Under Assumptions (A0)-(A2), the conclusion of Proposition 2.1 holds.

By comparison with the original article by Donnet et al. [2], (A4) or (A5) is the only additional assumption. However, we can show that this condition only restricts the prior on the parametric matrix ρ, and is satisfied with mild assumptions on classical families of priors.

2.3 Consistency of a graph estimator

Assumption (A4) is not needed to prove the consistency of a graph estimator designed as follows. We define the loss function on the graph parameter: for a true value $f_\star$,

$$L(\delta, f_*) = \sum_{kl} \mathbf{1}_{\delta^\star_{lk}=0}\mathbf{1}_{\delta^*_{lk}=1} + \mathbf{1}_{\delta_{lk}=1}\{\mathbf{1}_{\delta^*_{lk}=0} + \mathbf{1}_{\delta^*_{lk}=1}F(\rho^*_{lk})\}.$$

with $F : \mathbb{R}^+ \to \mathbb{R}^+$ non-increasing - for instance, a threshold function $F(x) = \mathbf{1}_{x\leq\epsilon}$ with $\epsilon > 0$. $L(\delta, f_*)$ is designed to have a "level of detection" - such as the parameter ϵ in the previous example. The corresponding averaged risk after an observation of N writes:

$$\begin{aligned} r(\delta, \Pi(.|N)) &= \int_{\mathcal{F}} L(\delta, f_*) d\Pi(f_*|N) \\ &= \sum_{kl} \mathbf{1}_{\delta_{lk}=0}\Pi(\delta^*_{lk} = 1|N) + \mathbf{1}_{\delta_{lk}=1}\left\{\Pi(\delta^*_{lk} = 0|N)\right. \\ &\quad \left. + \Pi(\delta^*_{lk} = 1|N)\mathbb{E}^\Pi[F(\rho^*_{lk})|N, \delta^*_{lk} = 1]\right\}. \end{aligned}$$

Let $\hat{\delta}^\Pi(N) = \arg\min_\delta r(\delta, \Pi(.|N))$ be the estimator minimizing the risk.

Proposition 2.3 *Under Assumptions (A0)-(A2),*

$$\mathbb{E}_0\left[\{\hat{\delta}^\Pi \neq \delta^0\}\right] \xrightarrow[T\to\infty]{} 0.$$

3 Posterior concentration rate in the general model

We extend the results by Donnet et al. [2] on the posterior concentration rate of the linear Hawkes model to the non-linear model defined in the first section. The proof makes use of the new results from the renewal framework by Costa et al. in [3].

For any interaction function h, we define the positive and negative parts as follows:

$$\forall t \in [0, A], \quad h(t) = \max(h(t), 0) - \max(-h(t), 0) := h^+(t) - h^-(t).$$

At the present time, we are able to determine an upper bound of the posterior concentration rate for parametric interaction functions in the space of histograms with finite

number of bins and rational coefficients. We define the spaces $\mathcal{H}_J$ and $\mathcal{F}_J$ as follows: for any $h = (h_{kl})_{1\leq k,l\leq K} \in \mathcal{H}_J$,

$$\forall k, l \quad h_{kl}(t) = \sum_{j=1}^{J} \omega_{kl}^{j} \mathbb{1}_{I_j}(t),$$

with $J < \infty$, $\{I_j\}_{j=1}^{J}$ a finite partition of $[0, A]$.

$$\mathcal{F}_J = \{f = ((\nu_k)_{k=1}^{K}, (h_{lk})_{k,l=1}^{K}); \nu_k > 0, \forall k, (h_{lk})_{lk} \in \mathcal{H}_J \quad \& \quad \forall k, l, \sup_t h_{lk}^{-}(t) < \nu_k\}.$$

Proposition 3.1 *Suppose that $f_0 \in \mathcal{F}_J$ and for all $j = 1, \ldots, J$, ω_{kl}^{0j} are rational numbers, i.e. $\omega_{kl}^{0j} = \frac{p_{kl}^{j}}{q_{kl}^{j}}$, $p_{kl}^{j} \in \mathbb{Z}$ and $q_{kl}^{j} \in \mathbb{N}$. Under Assumptions (A0-A3), the posterior distribution concentrates at least at the rate ϵ_T in L_1 and d_{1T} metrics, i.e. for any sequence $w_T \to \infty$,*

$$\mathbb{E}_0 \left[\Pi(d_{1T}(f, f_0) > w_T \epsilon_T | N)\right] = o(1),$$
$$\mathbb{E}_0 \left[\Pi(\|f - f_0\|_1 > w_T \epsilon_T | N)\right] = o(1).$$

References

[1] Subashis Ghosal and Aad van der Vaart. Convergence rates of posterior distributions for noniid observations. *Annals of Statistics*, 35(1):192–223, 2007.

[2] Sophie Donnet, Vincent Rivoirard, and Judith Rousseau. Nonparametric bayesian estimation of multivariate hawkes processes. *arXiv:1802.05975v2*, 2018.

[3] Manon Manon Costa, Carl Graham, Laurence Marsalle, and Viet Chi Tran. Renewal in hawkes processes with self-excitation and inhibition. *arXiv:1801.04645v1*, 2018.

A probabilistic model for the Rand Index

Martina Sundqvist [1,2], Julien Chiquet [1] & Guillem Rigaill [3,4,5]

[1] *UMR 518 MIA Paris, Université Paris-Saclay, AgroParisTech, INRAE, martina.sundqvist@agroparistech.fr*

[2] *Institut Curie - PSL Research University, Translational Research Department, Plateforme RPPA, 26 rue d'Ulm, 75005 Paris*

[3] *Institute of Plant Sciences Paris Saclay IPS2, CNRS, INRAE, Université Paris-Sud, Université Evry, Université Paris Saclay, Batiment 630, 91405 Orsay, France*

[4] *Institute of Plant Sciences Paris-Saclay IPS2, Paris Diderot, Sorbonne Paris-Cité, Bâtiment 630, 91405, Orsay, France*

[5] *Laboratoire de Mathématiques et Modélisation d'Evry, UMR CNRS 8071, Université d'Evry Val d'Essonne, 23 boulevard de France, 91037 Evry, France*

Résumé.

L'indice de Rand (en anglais "Rand Index" – RI) mesure la similitude de deux clusterings en estimant la probabilité d'avoir une paire cohérente. Le RI dépend du nombre de groups et est donc difficile à interpréter. Le RI ajusté (en bref ARI) contourne ce problème en corrigeant ce qui se passerait si les deux clusterings étaient indépendants (hypothèse nulle $\mathcal{H}_0$).

Le RI et l'ARI sont rarement présentés comme des estimateurs de la probabilité d'avoir des paires cohérentes. Nous proposons ici une analyse statistique rigoureuse à l'aide d'un modèle génératif multinomial sous $\mathcal{H}_0$ ou sour l'hypothèse alternative $\mathcal{H}_1$ de dépendance entre les clusterings. En fait, sous $\mathcal{H}_1$, il n'est pas évident d'estimer ce qui se passerait sous le $\mathcal{H}_0$ nul. En particulier, nous montrons que l'ARI est biaisé.

Keywords. Classification non-supervisée, Indice de Rand, Indice de Rand Adjusté

Abstract.

The Rand Index (RI) measures the similarity between two clusterings and estimates the probability of having a coherent pair. The RI depends on the number of groups and is therefore difficult to interpret. To overcome this issue the Adjusted RI (in short ARI) corrects for what would happen if the two clustering were independent (referred to as null hypothesis $\mathcal{H}_0$ hereafter).

The RI and ARI are rarely presented as estimators of the probability of having coherent pairs, under the null or under the alternative (dependence between clustering, denoted by $\mathcal{H}_1$). Here we propose a rigorous statistical analysis using a generative multinomial model. In fact, under the alternative $\mathcal{H}_1$, it is not trivial to estimate what would happen under the null $\mathcal{H}_0$. In particular, we show that the ARI is biased.

Keywords. Clustering, Rand Index, Adjusted Rand Index

1 Introduction

Cluster comparison is a common issue in statistics. Many cluster comparison measures have been proposed (Vinh et al., 2010), whereof one of the most popular ones is the Rand index (RI) (Rand, 1971) and its adjusted variant (ARI) (Hubert and Arabie, 1985; Morey and Agresti, 1984).

The RI tries to estimate the probability p_S of having a coherent pair by counting the number of coherent pairs, *i.e.*, pairs of observations that are either in the same group in both clusterings or in different groups in both clusterings.

The RI depends on the number of groups and is therefore difficult to interpret. To overcome this issue the Adjusted RI (in short ARI) corrects for what would happen if the two clustering were independent (referred to as null hypothesis $\mathcal{H}_0$ hereafter). This is done by substracting an estimator of $p_{S_{\mathcal{H}_0}}$, *i.e.*, the value of p_S when the two clusterings are independent.

The RI and ARI are rarely presented as estimators of p_S and $p_S - p_{S_{\mathcal{H}_0}}$ and has pretty much been derived as ad-hoc quantities. Here we propose a rigorous statistical analysis of RI and ARI using a generative multinomial model to derive their bias regarding the estimation of $p_S - p_{S_{\mathcal{H}_0}}$.

2 Statistical model

Notations and Settings. Let us consider two clustering $C^1 = \{c_1^1, c_2^1, \dots c_K^1\}$ and $C^2 = \{c_1^2, c_2^2, \dots c_L^2\}$ composed by K and L groups of the same $i = 1, \dots, n$ observations. The information is usually summarized in the contingency Table 1 representing the number of observations $n_{k\ell}$ in group k in C^1 and group ℓ in C^2.

A natural probabilistic model to generate the $n_{k\ell}$ is the multinomial distribution with probability $\pi_{k\ell}$ for each couple (k, ℓ) such as $\sum_{k,\ell}^{K,L} \pi_{k\ell} = 1$. We also define the marginal probabilities of being in group k in C^1 : $\sum_\ell^L \pi_{k\ell} = \pi_{k.}$ and in group ℓ in C^2 : $\sum_k^K \pi_{k\ell} = \pi_{.\ell}$. The corresponding probability distributions are represented in Table 2.

$C^1 \backslash C^2$	c_1^2	c_2^2	$\cdots$	c_L^2	$Sums$
c_1^1	n_{11}	n_{12}	$\cdots$	n_{1L}	$n_{1.}$
c_2^1	n_{21}	n_{22}	$\cdots$	n_{2L}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
c_K^1	n_{1K}	n_{2K}	$\cdots$	n_{KL}	$n_{K.}$
$Sums$	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.L}$	$\sum_1^{K,L} n_{k\ell} = n$

Table 1: Contingency Table between clusterings C^1 and C^2; each entry $n_{k\ell}$ corresponds to the number of observations in group k in C^1 and group ℓ in C^2.

$C^1 \backslash C^2$	c_1^2	c_2^2	$\cdots$	c_L^2	$Sums$
c_1^1	π_{11}	π_{12}	$\cdots$	π_{1L}	$n_{1.}$
c_2^1	π_{21}	π_{22}	$\cdots$	π_{2L}	$\pi_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
c_K^1	π_{1L}	π_{2L}	$\cdots$	$\pi_{k\ell}$	$\pi_{K.}$
$Sums$	$\pi_{.1}$	$\pi_{.2}$	$\cdots$	$\pi_{.L}$	$\sum_1^{K,L} \pi_{k\ell} = 1$

Table 2: Probabilistic distribution $\pi_{k\ell} = \mathbb{P}(i \in c_k^1 \cap c_\ell^2)$

People studying the (A)RI most often considered an hypergeometric model for $n_{k\ell}$ under the null $\mathcal{H}_0$. Extending this to the alternative (here after noted $\mathcal{H}_1$) is not obvious.

Definitions of the (A)RI. Let S_{ij} be a Bernoulli variable indicating whether (i, j) is a coherent pair or not:

$$S_{ij} = \begin{cases} 1 \text{ if } (i,j) \text{ are in the same group for } C^1 \text{ and } C^2 \\ 1 \text{ if } (i,j) \text{ are not in the same group for } C^1 \text{ and } C^2 \\ 0 \text{ otherwise} \end{cases}$$

The probability of having a consistent pair is set to $\mathbb{P}(S_{ij} = 1) = p_S$ and is obtained from the $\pi_{k\ell}$ contingency table by

$$p_S = 1 + 2\sum_{k,\ell}^{K,L} \pi_{k\ell}^2 - \sum_k^K \pi_{k.}^2 - \sum_\ell^L \pi_{.\ell}^2.$$

Note that under the null $\mathcal{H}_0$, this simplifies to

$$p_{S_{\mathcal{H}_0}} = 1 + 2\sum_{k,\ell}^{K,L} \pi_{k.}^2 \pi_{.\ell}^2 - \sum_k^K \pi_{k.}^2 - \sum_\ell^L \pi_{.\ell}^2$$

.

To estimate p_S, the RI was initially introduced as the sum of coherent pairs among the total number of pairs:

$$RI(C^1, C^2) = \sum_{i<j}^n S_{ij} \Big/ \binom{n}{2}.$$

For its computational simplicity, the RI is commonly presented as

$$RI(C^1, C^2) = 1 + \left[2\sum_{k,\ell}^{K,L} \binom{n_{k\ell}}{2} - \sum_k^K \binom{n_{k.}}{2} - \sum_\ell^L \binom{n_{.\ell}}{2}\right] \Big/ \binom{n}{2}.$$

The Adjusted Rand Index (ARI) is usually defined as (see Brennan and Light, 1974; Hubert and Arabie, 1985)

$$ARI = \sum_{k,\ell} \binom{n_{k\ell}}{2} \Big/ \binom{n}{2} - \sum_{k} \binom{n_{k.}}{2} \sum_{\ell} \binom{n_{.\ell}}{2} \Big/ \binom{n}{2}^2,$$

and can be viewed as an estimator of $p_S - p_{S_{\mathcal{H}_0}}$.

Under the multinomial model, we study the properties of the RI and the ARI as estimators of p_S, respectively $p_S - p_{S_{\mathcal{H}_0}}$ under $\mathcal{H}_0$ or $\mathcal{H}_1$.

3 Statistical Inference

The natural estimator of p_S is the RI. Indeed, by construction, the RI is unbiased, under $\mathcal{H}_0$ or $\mathcal{H}_1$:

$$\mathbb{E}(RI) = \mathbb{E}\Big(\sum_{i<j}^{n} S_{ij} \Big/ \binom{n}{2} \Big) = \binom{n}{2} \mathbb{E}(S_{ij}) \Big/ \binom{n}{2} = \mathbb{P}(S_{ij} = 1) = p_S.$$

In order to compute the ARI, we need an estimator of $\widehat{p}_{S_{\mathcal{H}_0}}$ on top of the RI. Estimating $p_{S_{\mathcal{H}_0}}$ is not trivial as in practise we do not observe a $n_{k\ell}$ contingency table under the null $\mathcal{H}_0$ but rather under the alternative $\mathcal{H}_1$.

Here we consider the usual estimator $p_{S_{\mathcal{H}_0}}$ proposed by (Brennan and Light, 1974; Hubert and Arabie, 1985), which is based on a hypergeometric distribution, assuming the marginals of the $n_{k\ell}$ contingency table fixed:

$$\widehat{p}_{S_{\mathcal{H}_0}} = 1 + \frac{2}{\binom{n}{2}^2} \sum_{k\ell}^{KL} \binom{n_{k.}}{2} \binom{n_{.\ell}}{2} - \frac{1}{\binom{n}{2}} \Big[\sum_{k}^{K} \binom{n_{k.}}{2} + \sum_{\ell}^{\ell} \binom{n_{.\ell}}{2} \Big]. \tag{1}$$

We studied the statistical properties (bias and variance) of the RI and $\widehat{p}_{S_{\mathcal{H}_0}}$. In particular we get the following proposition that gives the expected value of $\widehat{p}_{S_{\mathcal{H}_0}}$.

Proposition 3.1. *Suppose that the two clusterings C^1 and C^2 are drawn under the alternative $\mathcal{H}_1$ (i.e. C^1 and C^2 are dependent) with the multinomial model. Then, we can show that the expectation of estimator* (1) *is*

$$\mathbb{E}_{\mathcal{H}_1}(\widehat{p}_{S_{\mathcal{H}_0}}) = \frac{(n+1)(n-2)+8}{(n-1)^2} + \frac{n(2n-3)-1}{n(n-1)^2} \Big[\sum_{k} \pi_{k.}^2 + \sum_{l} \pi_{.\ell}^2 \Big] + \frac{8}{n^2(n-1)^2} \Big[\binom{n}{2} \sum_{kl} \pi_{k\ell}^2 + 6\binom{n}{3} \sum_{kl} \pi_{k\ell} \pi_{.\ell} \pi_{k.} + 6\binom{n}{4} \sum_{kl} \pi_{k.}^2 \pi_{.\ell}^2 \Big].$$

Hence, $\mathbb{E}_{\mathcal{H}_1}(\widehat{p}_{S_{\mathcal{H}_0}}) \neq p_{S_{\mathcal{H}_0}}$ and $\widehat{p}_{S_{\mathcal{H}_0}}$ is biased.

To derive those results we re-formulated $\widehat{p}_{S_{\mathcal{H}_0}}$ as a sum of Bernoulli variables.

Interestingly, we see that the expectation of $\widehat{p}_{S_{\mathcal{H}_0}}$ is different from $p_{S_{\mathcal{H}_0}}$. Thus the usual estimator of the ARI is biased under a multinomial model.

In the rest of the presentation we will present numerical simulations illustrating this bias under different cases of clustering dependence.

References

Robert L Brennan and Richard J Light. Measuring agreement when two observers classify people into categories not defined in advance. *British Journal of Mathematical and Statistical Psychology*, 27(2):154–163, 1974.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2 (1):193–218, 1985.

Leslie C Morey and Alan Agresti. The measurement of classification agreement: An adjustment to the rand statistic for chance agreement. *Educational and Psychological Measurement*, 44(1):33–37, 1984.

W.M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(Oct):2837–2854, 2010.

Fair Adversarial Network and explainability

Simon Grah [1] & Ilyes Mahammed Chikouche [1]& Vincent Thouvenot [1]

[1] *Thales SIX GTS France, 1 Avenue Augustin Fresnel, Palaiseau*
prenom.nom@thalesgroup.com

Résumé. Les modèles de Machine Learning ont tendance à reproduire et à amplifier les biais. Certaines techniques, s'appuyant sur la Fairness, permettent de lutter contre ces biais. La Fairness est l'un des sous-thèmes du "Trustable Machine Learning". Un autre sous-thème est l'explicabilité. En Machine Learning, nous utilisons certaines techniques souvent très fortement non linéaires. Comprendre le comportement des modèles est une tâche compliquée. Dans cet exposé, nous utilisons un réseau advseraire issu de la littérature pour limiter l'influence de certains paramètres sur un classifieur afin de respecter les contraintes de Fairness. Nous mesurons le gain en Fairness par rapport à un classifieur classique avec une métrique nommée la p-rule, avec des métriques basées sur la distance entre deux distributions et avec la Differential Fairness. Pour avoir une explication locale de la différence entre le classifieur classique et le classifieur adversaire, nous calculons les valeurs de Shapley, une technique d'explication classique issue de la théorie des jeux, dont la définition dépendra d'une population de référence, et non de la prédiction moyenne comme il est souvent fait. Enfin, nous montrons l'apport du réseau adversaire en termes de protection des données. Les travaux ont été réalisé dans le cadre du projet H2020 SPARTA.

Mots-clés. Explainabilité, Fairness, Réseau adversaire, Valeur de Shapley

Abstract. Machine Learning models tend to reproduce and amplify biases. Some techniques, relying on Fairness, allow to fight against these bias. Fairness is one of the sub-topic of Trustable Machine Learning. Another sub-topic is explainability. In Machine Learning, we use some techniques often very strongly non-linear and understanding models behavior is a hard task. In this talk, we will use an adversarial network coming from literature to limit influence of some parameters on a classifier to achieve Fairness. We will measure the gain of Fairness compared to a classifier with the base rate metric named p-rule, with metrics based on the distance between two distributions and with the Differential Fairness. To have local explanation of the difference between the classical and the fair classifier, we will compute the Shapley Value, a classical explanation technique coming from Game Theory, whose definition will depend on a reference population, and not the average prediction as often seen. Finally, we will show the contribution in terms of data protection of the fair classifier. This work is funded under the SPARTA H2020 project.

Keywords. Adversarial Network, Explainability, Fairness, Shapley Value

1 Introduction

1.1 Fairness

Machine Learning models tend to reproduce and amplify biases. These biases can come from the data: there are known biases such as selection bias when sampling is poor, historical biases, for example when a population is disadvantaged, etc. There are also biases that can come from algorithms: some recommendation algorithms lock people in bubbles instead of offering them new possibilities, when data are unbalanced, etc. There are several possible definitions of Fairness. First, we can wish that the model outputs would be similar for two subpopulations. It is to be hoped that the errors in the model will be similar for the different subpopulations. Finally, we can wish that two similar individuals except the sub-population to which they belong receive similar treatment.

1.2 Explainability

In Machine Learning (ML), and more generally in artificial intelligence, we mainly focus on performance. Highly non-linear models such as Random Forests, Gradient Boosting, Deep Learning, etc. are often used. Generally, performance are satisfactory. However, we are unable to explain how the model builds its decisions. This can be a problem, especially when dealing with critical systems. There are several levels of explanation in ML models. The first level is global: we want to understand the model general behavior and the features impact on the model's outputs. At local level, we explain why a prediction was made for an observation. The explanation position has to be defined. Do we learn a model that is interpretable by nature, like e.g. additive model ? Do we learn the model to generate an explanation at the same time as we make a prediction (see e.g. Baratt, 2017)? Does the explanation require an additional component, independent of the type of model used (see e.g. Ribeiro *et al*, 2016)? Some objectives of explainability are to verify ML functionality and increase confidence of users. Since these systems are used as decision support, the task is to design an approach to optimize the relation between performance boost provided by these systems and explainability of its decisions. This explainability can be directed to the engineer developing these systems, to improve it, and the final users of these systems who needs to understand the decisions made, for being more confident in the results and enriching the by its business knowledge.

2 Fair Adversarial Network

2.1 Fairness measure

In this subsection, we give some classical measure of Fairness on a binary classification task (see e.g. Friedler *et al.*, 2018, Zafar *et al.*, 2015). Note a binary class prediction

$\hat{Y} \in \{0,1\}$ and a binary sensitive attribute $Z \in \{0,1\}$. A base rate metrics measures the change of models' outputs according to a sensitive attribute. For instance, p%-rule is given by $\min\left(\frac{P(\hat{Y}=1|Z=1)}{P(\hat{Y}=1|Z=0)}, \frac{P(\hat{Y}=1|Z=0)}{P(\hat{Y}=1|Z=1)}\right) \geq \frac{p}{100} \in [0,1]$. Group-conditioned accuracy metrics measure the model's errors according a sensitive attribute. For instance, the Z-accuracy is given by $P(\hat{Y} = y|Z = z, Y = y) \in [0,1]$. The group-conditioned calibration measure the label repartition knowing model prediction according to an attribute. For instance, Z-calibration + is given by $P(Y = 1|Z = z, \hat{Y} = 1) \in [0,1]$. Individual-level discrimination measure how the model handle one individual comparing the most similar individuals. The consistency is given by $1 - \frac{1}{n}\sum_{i=1}^{n}\sum_{j \in knn(i)} |\hat{Y}_i - \hat{Y}_j| \in [0,1]$, where $knn(i)$ are the K-Nearest Neighbors of i.

Foulds and Pan (2018) proposes the Differential Fairness which states that a mechanism $M(x)$ is ε-differentially fair in a framework (A, Θ), where A is the ensemble of attributes to protect, if for all $\theta \in \Theta$ with $x \sim \theta$ and $y \in Range(M)$,

$$\exp(-\varepsilon) \leq \frac{P_{M,\theta}(M(x) = y|\mathbf{s_i}, \theta)}{P_{M,\theta}(M(x) = y|\mathbf{s_j}, \theta)} \leq \exp(\varepsilon),$$

for all $(\mathbf{s_i}, \mathbf{s_j}) \in A \times A$, where $P(\mathbf{s_i}|\theta) > 0$, $P(\mathbf{s_j}|\theta) > 0$.

2.2 Adversarial Network

Several techniques have been proposed to achieve Fairness. Some are based on regularization methods: the direct and indirect information which rely to the sensitive attribute(s) are penalized (see e.g. Raff *et al.*, 2017). Some others techniques will use new representations that hide attribute, e.g. by Deep Learning (see e.g. Louizos *et al.*, 2015). Some others techniques introduce fairness constraint by modifying the inputs or the outputs of the algorithms (see e.g. Kamiran *et al.*, 2018).

We consider the adversarial network proposed by Louppe *et al.*, 2017. They use this architecture in context where they want to introduce independence between the outputs of a classifier and some nuisance parameters. The overview of the architecture is given by Figure 1. Our objective will be to use this architecture to make independent the outputs of a classifier of constituent elements to a sub-population to be protected or to elements which we wish that they cannot be revealed by the model's outputs.

Note $\hat{\mathbf{Y}}$ classifier prediction based on input X, $\mathbf{Y}$ the true value and A the sensitive attributes, $\theta_{\mathbf{clf}}$ and $\theta_{\mathbf{ad}}$ parameters of respectively the classifier and the adversarial networks, $Loss_Y(\theta_{\mathbf{clf}})$ and $Loss_A(\theta_{\mathbf{clf}}, \theta_{\mathbf{ad}})$ the loss of a pre-trained classifier and adversarial networks.

During the iteration, the objective function given by Equation (1) is considered:

$$\theta_{\mathbf{clf}}, \theta_{\mathbf{ad}} = \arg\min_{\theta_{\mathbf{clf}}} \max_{\theta_{\mathbf{ad}}} \left(Loss_Y(\theta_{\mathbf{clf}}) - \lambda Loss_A(\theta_{\mathbf{clf}}, \theta_{\mathbf{ad}})\right). \quad (1)$$

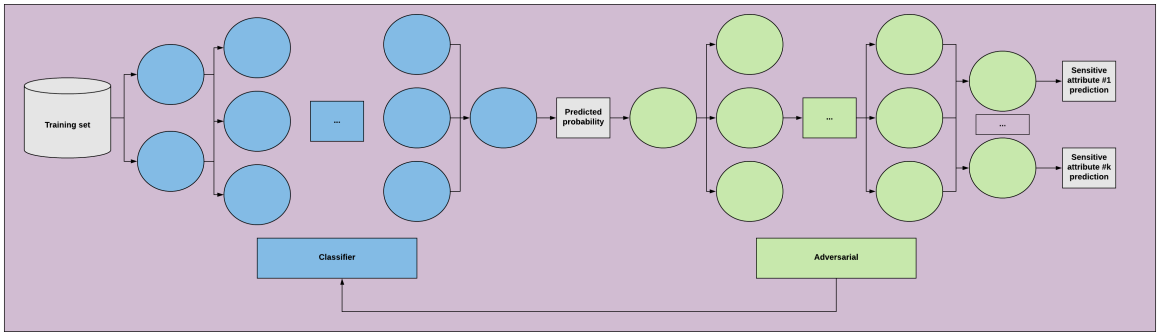


Figure 1: Overview of adversarial network architecture of Louppe *et al.* (2017)

3 Shapley Value

3.1 Definition

Note $v : 2^N \to \Re$ such as $v(\emptyset) = 0$ and N a set of players. If $S \subset N$, $v(S)$ is the amount of wealth produced by S when they cooperate. The Shapley Value (see Shapley, 1953) is a fair share of the global wealth $v(N)$ produced by all players together, among themselves.

$$\Phi_i(N, v) = \sum_{S \subset N \setminus i} \frac{(card(N) - card(S) - 1)!card(S)!}{card(N)!} \left(v(S \cup i) - v(S)\right).$$

The Shapley value is the only indicator that respects the four following properties:

- Additivity: $\Phi_i(N, v + w) = \Phi_i(N, v) + \Phi_i(N, w)$ for all i;
- Null player: if $v(S \cup i) = v(S)$ for all $S \subset N \setminus i$ then $\Phi_i(N, v) = 0$;
- Symmetry: $\Phi_{\pi i}(\pi N, \pi v) = \Phi_i(N, v)$ for every permutation π on N;
- Efficiency: $\sum_{i \in N} \Phi_i(N, v) = v(N)$.

In Machine Learning, Shapley Value can be expressed according the Equation (2).

$$\Phi_i = \sum_{S \subset F \setminus i} \frac{(card(F) - card(S) - 1)!card(S)!}{card(F)!} \left(f_{S \cup i}(\mathbf{x_{S \cup i}}) - f_S(\mathbf{x_S})\right), \tag{2}$$

where $f_S(x_S) = \int \hat{f}(x_1, \ldots, x_p) dP_{x \notin S} - E_X\left(\hat{f}(X)\right)$, F the features set, S a subset of features such as $S \subset F$, $v(S) = f_S(x_S) = E(\hat{f}(\mathbf{x}|\mathbf{x_S}) = \mathbf{x_S^*}$, with $\mathbf{x_S^*}$ real value of the features associate at the instance explained on features subset S.

3.2 Approximation

The exact computation of the Shapley Value involves $O(2^p)$ calculations, which is very quickly untractable to do. Some authors rewrite the Shapley Value problem as a weighted least square optimization problem (see e.g. Lundberg and Lee, 2017; Aas *et al.*, 2019).

Some authors use Monte Carlo techniques to approximate the Shapley Value (see e.g. Strumbelj and Kononenko, 2014; Maleki *et al.*, 2014). Maleki *et al.*, 2014 proves the good performance in terms of errors of Monte Carlo estimation of Shapley Value when the variance or the range of the players' marginal contributions is known.

3.3 Adaptation to a reference population

In classical application of Shapley Value, the reference population is the average prediction and we measure the contribution of each feature to the difference between this prediction and the prediction made for the instance. Merrick and Taly, 2019 propose a generalization of this definition. In the game proposed, the amount of wealth produced by S when they cooperate is written by:

$$v_{\mathbf{x},D^{ref}}(S) = E_{\mathbf{R}\sim D^{ref}}\left(f(\mathbf{z}(\mathbf{x},\mathbf{R},S))\right) - E_{\mathbf{R}\sim D^{ref}}\left(f(\mathbf{R})\right),$$

where $\mathbf{z} : (\mathbf{x},\mathbf{R},S) \mapsto (z_1,\ldots,z_p)$ with $z_i = x_i \times \mathbb{1}_{\mathbb{i}\in\mathbb{S}} + r_i \times \mathbb{1}_{\mathbb{i}\notin\mathbb{S}}$ for all $i \in (1,\ldots,p)$, D^{ref} a sampling distribution (e.g. uniform on the range of the features or a sampling from features marginal distribution).

4 Application

During this talk, after introducing the fair adversarial network and the Shapley Value, we will propose an illustration of the use of the Fair Adversarial Network on one use case which could be on COMPAS dataset, a dataset of justice predictive where it exists some known bias. We will compare the performance both in term of accuracy and in term of Fairness. For Fairness evaluation, in addition to the p-rule and the Differential Fairness, we will propose base rate metric based on the difference between two distributions which that will solve the threshold problem of the p-rule. Assume we consider a case of binary classification with $\hat{Y} = 1$ or 0 where a protected attribute A takes two modality: a_1, the discriminated modality, and a_2. We can compute kernel estimator of the densities of $P(\hat{Y} = 1|A = a_1)$ and $P(\hat{Y} = 1|A = a_2)$, and then compute the Kullback Leibler (KL) divergence between the two estimated distributions. We compute the max between of the two KL divergence when $A = a_1$ and $A = a_2$ are respectively the reference population. Another metric consists to compute the Kolmogorov-Smirnov statistic between the two empirical distributions. Last metric we propose is to compute the Dynamic Time Warping distance between $(p_{(1)},\ldots,p_{(n_1)})$ and $(q_{(1)},\ldots,q_{(n_2)})$, where for all $i \in \{1,\ldots,n_1\}$, $p_{(i)}$ (resp. for all $i \in \{1,\ldots,n_2\}$, $q_{(i)}$) is the statistic of order i of the probability predicted

by the classifier for the instance such as $A = a_1$ (resp. the probability predicted by the classifier for the instance such as $A = a_2$), with n_1 (resp. n_2) the cardinal of the instances which have a_1 modality (resp. a_2 modality). Thanks to the Shapley Value, we will explain the difference between a classical classifier, which has no Fairness constraint, and the Fair classifier. Moreover we will illustrate the contribution in terms of data protection made possible by the Fair Adversarial Network.

5 Acknowledgement

This work is funded under the SPARTA project, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 830892.

Reference

Aas, Jullum, and Loland (2019), Explaining individual predictions when features are dependent: more accurate approximations to shapley values
Baratt (2017), InterpNET: Neural Introspection for Interpretable Deep Learning
Foulds and Pan (2018), An Intersectional Definition of Fairness
Friedler, Scheidegger, Venkatasubramanian, Choudhary, Hamilton, Roth (2018), A comparative study of fairness-enhancing interventions in machine learning, Proceedings of the Conference on Fairness, Accountability, and Transparency
Kamiran, Mansha, Karim and Zhang (2018), Exploiting reject option in classification for social discrimination control, Information Sciences
Louizos, Swersky, Li, Welling, Zemel (2015), Variational Fair Autoencoder, ICLR 2016
Louppe, Kagan, and Cranmer (2017), Learning to Pivot with Adversarial Networks, Advances in Neural Information Processing Systems 30
Lundberg and Lee (2017), A Unified Approach to Interpreting Model, NIPS 2017
Maleki, Tran-Thanh, Hines, Rahwan, and Rogers (2014). Bounding the Estimation Error of Sampling-based Shapley Value Approximation
Merrick and Taly (2019), The explanation Game: Explaining Machine Learning Models with cooperative Game Theory
Raff, Sylvester, Mills (2017), Fair Forests: Regularized Tree Induction to Minimize Model Bias, AAAI / ACM conference on AIES 2018
Ribeiro, Singh, Guestrin (2016), Local Interpretable Model-Agnostic Explanations
Shapley (1953), A Value for n-person Games. Contributions to the Theory of Games.
Strumbelj and Kononenko (2014), Explaining prediction models and individual predictions with feature contributions, Knowledge and information systems
Zafar, Valera, Gomez Rodriguez, Gummadi (2015), Fairness Constraints: Mechanisms for Fair Classification, Proceedings of the 20th IAISTATS

Robust estimators for PDMP

Charles Tillier [1] & Patrice Bertail [2] & Gabriela Ciolek [3]

[1] *University of Versailles, 45 avenue des états-unis, 78000 Versailles, France*
[1] *charles.tillier@gmail.com*
[2] *University of Nanterre, 200 avenue de la république, 92000 Nanterre, France*
[2] *patrice.bertail@gmail.com*
[2] *University of Aarhus, Ny Munkegade 118, 8000 Aarhus C, Denmark*
[2] *gabriela.ciolek@math.au.dk*

Résumé. Dans ce document, on propose une méthode pour construire des estimateurs robustes pour des processus déterministes par morceaux markoviens (PDMP) qui repose sur la théorie de renouvellement des chaînes de Markov et la technique de bootstrap par blocks approximativement régénératifs introduite dans [1] et [2], qui consiste à éliminer les blocks ayant soit une contribution, soit une longueur trop importante sur la statistique d'intérêt. Dès lors, on peut construire des estimateurs robustes pour la chaîne de Markov immergée, ce qui, en explicitant un lien entre le processus à temps continu et cette dernière, donne lieu à des estimateurs robustes pour le PDMP. On met en avant l'applicabilité de cette méthode en proposant des estimateurs robustes de la mesure stationnaire de deux processus en théorie du risque : le modèle de Sparre Andersen avec une barrière et le modèle de pharmacocinétique KDEM étudié dans [3].

Mots-clés. PDMP, estimateurs robustes, chaînes de Markov, théorie du risque, renouvellement

Abstract. In this document, we propose a method to build robust estimators for PDMPs, which is of particular interest when the underlying process is contaminated by outliers. The aforementioned method relies on a renewal theory for Markov chains and further developments of the *approximate regenerative block bootstrap method* introduced in [1] and [2], i.e. we eliminate blocks having either too much contribution to the statistics of interest or having a too large length and we build efficient robust estimators for the embedded Markov chain associated to the PDMP. Relating the properties of the underlying process and its embedded chain, this leads to robust estimators for the PDMP. To highlight the applicability of the method, we consider robust estimators of the stationary measure of PDMPs in risk theory: the Cramér-Lundberg model with a dividend

barrier and the Kinetic Dietary Exposure Model used in modeling pharmacokinetics of contaminants studied in [3].

Keywords. PDMP, robust estimators, Markov chains, risk theory, renewewal properties

1 The framework

Throughout the document, all the random variables are defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Moreover, we assume that

(H1): $(W_i)_{i\in\mathbb{N}}$ are i.i.d. nonnegative r.v's with a common mean γ and the c.d.f. F_W.

(H2): $(\Delta T_i)_{i\in\mathbb{N}^*}$ is an i.i.d. sequence of a.s. positive r.v.'s with the c.d.f. H independent of the sequence $(W_i)_{i\in\mathbb{N}}$. We assume $\lambda = \mathbb{E}[\Delta T_1] < \infty$ and $Var[\Delta T_1] < \infty$.

(H3): $(T_i)_{i\in\mathbb{N}}$, defined for all $i \geq 1$ by $T_i = \sum_{k=1}^{i} \Delta T_k$ forms an increasing sequence of r.v's. By convention, we set $T_0 = 0$.

(H4): The counting process $\{(N(t)\}_{t\geq 0}$ defined by $N(t) := \#\{i \in \mathbb{N}^* : T_i \leq t\}$ for $t \geq 0$ is a renewal process and $A(t) = t - T_{N(t)}$ is the backward recurrence time.

We start by defining two models in risk theory for which we will propose robust estimators for the stationary measure later on.

Example 1: The Sparre Andersen model with a dividend barrier $d > 0$ (S.A.D.B) models the reserve of an insurance company through time when dividends are paid out whenever the surplus level attains the threshold d. The *claims* W_i's arise at $T_0, T_1, \ldots$ and the ΔT_i's are the periods between claims, see [4]. Denoting $X(t)$ (resp. $X_{T_{N(t)}} = X(T_{N(t)})$) the reserve of an insurance company at time $t \geq 0$ (resp. at the latest claim before t), it is defined by

$$X(t) = \min(d, X_{T_{N(t)}} + cA(t)), \quad t \geq 0. \tag{1.1}$$

The process $X \equiv X(t)$ is a PDMP and the analysis of its long-term behavior boils down to investigating the properties of the embedded Markov chain $\tilde{X} = (X_n)_{n\geq 1}$, which corresponds to the PDMP X evaluated on the claim instants $X_n = X(T_n)$ for all $n \geq 1$ and is defined as

$$X_{n+1} = (X_n + Z_{n+1})\mathbb{I}_{(X_n + c\Delta T_{n+1} < d)} + (d - W_{n+1})\mathbb{I}_{(X_n + c\Delta T_{n+1} \geq d)},$$

where $(Z_i)_{i\in\mathbb{N}}$ is defined for $i \geq 0$ by $Z_i = c\Delta T_i - W_i$ and for which we assume $\mathbb{E}[Z_i] > 0$. The limiting behavior of X is represented by a stationary probability measure μ that describes the equilibrium state to which the process settles as time goes to infinity. μ is hardly tractable in general but one solution is to find a link between μ and $\tilde{\mu}$, the stationary measure of the embedded chain $\tilde{X} \equiv X_n$ which is easier to handle. Under (H1)-(H4), one may show that $X(t)$ has an absolutely continuous limiting probability distribution μ given by

$$\mu([-\infty, v[) = \lambda^{-1} \int_{-\infty}^{v} \int_{0}^{\infty} \max\left(t, \frac{v-x}{c}\right) \tilde{\mu}(dx)H(dt), \quad \infty < v \leq d. \tag{1.2}$$

Example 2. The KDEM for Kinetic Dietary Exposure Model studied in [3] models the evolution of a contaminant in the human body through time. Here the W_i's correspond to contaminated intakes occuring at T_i's and ΔT_i's are understood as the durations between the $(i-1)$-th and the i-th intake. Keeping on the same notation except that now $X(t)$ is the total body burden of a chemical at the instant $t \geq 0$, it is defined as

$$X(t) = X_{T_{N(t)}} \times e^{-\omega A(t)}.$$

The bivariate process $\{(X(t), A(t))\}_{t\geq 0}$ is a PDMP and looks similar to the S.A.D.B defined in (1.1) except that it is reversed, there is a natural barrier at 0 and the deterministic motion is not anymore linear but exponential. The embedded chain of X is defined by the stochastic recurrence equation

$$X_{n+1} = X_n \times e^{-\omega \Delta T_{n+1}} + W_{n+1}, \quad n \geq 0.$$

Under (H1)-(H4) and the additional assumption $\mathbb{E}[\log(\max(1, W_1))] < \infty$, [3] have related the continuous-time process X with the embedded chain $\tilde{X}$ by means of their stationary distributions as follows

$$\mu([u, \infty[) = \lambda^{-1} \int_{u}^{\infty} \int_{0}^{\infty} \left(t \wedge \omega^{-1} \log(x/u)\right) \tilde{\mu}(dx)H(dt), \quad u > 0.$$

2 Robustness for risk measure of PDMP

In most of the risk models, μ is itself a bounded functional of $\tilde{\mu}$ so one may construct a robust estimator of μ by just plugging the expression of a robust estimator of $\tilde{\mu}$ for which

we can most of the time easily get an explicit formulae. Indeed, from the Kac's theorem, $\tilde{\mu}$ can be written as a functional of the distribution of the blocks:

$$F_{\tilde{\mu}}(y) = \frac{\mathbb{E}_A\left(\sum_{i=1}^{\tau_A} I_{\{X_i \leq y\}}\right)}{\mathbb{E}_A[\tau_A]}$$

where A (resp. τ_A) is the atom (resp. the regenerative time) of the Markov chain. Let $M > 0$. Consider the robustified version of this c.d.f., which is simply obtained by eliminating too large blocks and given by

$$\tilde{F}_{\mathcal{L},M}(y) = \frac{\mathbb{E}_A\left[\left(\sum_{i=1}^{\tau_A} I_{\{X_i \leq y\}}\right) I_{\{\tau_A \leq M\}}\right]}{\mathbb{E}_A[\tau_A I_{\{\tau_A \leq M\}}]}.$$

A straightforward computation of the influence function of $\tilde{F}_{\mathcal{L},M}(y)$ leads to the expression

$$\tilde{F}_M^{(1)}(b,\ y, \mathcal{L}) = \frac{\sum_{i=1}^{L(b)} \left(I_{\{b_i \leq y\}} - F_\mu(y)\right) I_{\{L(b) \leq M\}}}{\mathbb{E}_A[\tau_A I_{\{\tau_A \leq M\}}]}, \quad \forall b \in \mathbb{T} = \cup_{n \geq 1} \mathbb{R}^n.$$

From this expression, we deduce that

$$\|\tilde{F}_{\mathcal{L},M} - F_\mu\|_\infty = \sup_y |F_{\mathcal{L},M}(y) - F_\mu(y)| \to 0$$

a.s. when $M \to \infty$. When $\|\cdot\| = \|\cdot\|_\infty$, its gross-error sensitivity is bounded by $M/\mathbb{E}_A[\tau_A I_{\{\tau_A \leq M\}}]$. It follows that the plug-in estimator of this quantity is given by

$$\tilde{F}_{\mathcal{L},M,n}(y) = \frac{\sum_i^{l_n - 1} \sum_{j=\tau_A(i)+1}^{\tau_A(i+1)} I_{\{X_j \leq y\}} I_{\{\tau_A(i+1) - \tau_A(i) \leq M\}}}{\sum_{i=1}^{l_n - 1} \left(\tau_A(i+1) - \tau_A(i)\right) I_{\{\tau_A(i+1) - \tau_A(i) \leq M\}}}.$$

A robust estimator of μ is obtained by plugging $\tilde{F}_{\mathcal{L},M,n}$ as it is illustrated in the following two examples.

Example 1. The stationary distribution μ of the S.A.D.B. defined in (1.2) may be also rewritten as a functional of the blocks :

$$\begin{aligned}
\mu([-\infty, v[) &= \lambda^{-1} \int_{-\infty}^{v} \int_0^{\infty} \left(t \wedge \frac{v - x}{c}\right) \frac{\mathbb{E}_A\left[\sum_{i=1}^{\tau_A} \delta_{X_i}(dx)\right]}{\mathbb{E}_A[\tau_A]} H(dt) \\
&= \frac{1}{\lambda \mathbb{E}_A[\tau_A]} E_A\left(\sum_{i=1}^{\tau_A} \int_{-\infty}^{\infty} I_{\{x \leq v\}} \int_0^{\infty} \left(t \wedge \frac{v - x}{c}\right) \delta_{X_i}(dx) H(dt)\right)
\end{aligned}$$

$$= \frac{1}{\lambda \mathbb{E}_A[\tau_A]} E_A \left(\sum_{i=1}^{\tau_A} I_{\{X_i \leq v\}} \int_0^\infty \left(t \wedge \frac{v - X_i}{c} \right) H(dt) \right).$$

Its robustified version is given by

$$\begin{aligned}
\mu([-\infty, v[) &= \lambda^{-1} \int_{-\infty}^{v} \int_0^\infty \left(t \wedge \frac{v-x}{c} \right) \frac{\mathbb{E}_A \left[\left(\sum_{i=1}^{\tau_A} \delta_{X_i}(dx) \right) I_{\{\tau_A \leq M\}} \right]}{\mathbb{E}_A[\tau_A I_{\{\tau_A \leq M\}}]} H(dt) \\
&= \frac{1}{\lambda \mathbb{E}_A[\tau_A I_{\{\tau_A \leq M\}}]} \mathbb{E}_A \left(\sum_{i=1}^{\tau_A} I_{\{X_i \leq v\}} \int_0^\infty \left(t \wedge \frac{v - X_i}{c} \right) I_{\{\tau_A \leq M\}} H(dt) \right)
\end{aligned}$$

and can be estimated by the robust plug-in estimator

$$\begin{aligned}
&\widehat{\mu}_n([-\infty, v[) \\
&= \lambda^{-1} \int_{-\infty}^{v} \int_0^\infty \left(t \wedge \frac{v-x}{c} \right) \tilde{F}_{\mathcal{L},M,n}(dx) H(dt) \\
&= \lambda^{-1} \int_{-\infty}^{v} \int_0^\infty \left(t \wedge \frac{v-x}{c} \right) \frac{\sum_i^{l_n - 1} \sum_{j=\tau_A(i)+1}^{\tau_A(i+1)} \delta_{X_j}(dx) I_{\{\tau_A(i+1) - \tau_A(i) \leq M\}}}{\sum_{i=1}^{l_n-1} (\tau_A(i+1) - \tau_A(i)) I_{\{\tau_A(i+1) - \tau_A(i) \leq M\}}} H(dt), \\
&= \frac{\sum_i^{l_n-1} \sum_{j=\tau_A(i)+1}^{\tau_A(i+1)} I_{\{\tau_A(i+1) - \tau_A(i) \leq M\}} I_{\{X_i \leq v\}} \int_0^\infty \left(t \wedge \frac{v - X_j}{c} \right) H(dt)}{\lambda \sum_{i=1}^{l_n-1} (\tau_A(i+1) - \tau_A(i)) I_{\{\tau_A(i+1) - \tau_A(i) \leq M\}}}.
\end{aligned}$$

In particular, if the ΔT_i's are exponential we have

$$\int_0^\infty \left(t \wedge \frac{v - X_j}{c} \right) H(dt) = \lambda \left(1 - \exp\left(- \frac{(v - X_j)}{\lambda c} \right) \right).$$

It follows that the estimator is essentially a mean of $\lambda \left(1 - \exp\left(- \frac{(v - X_j)}{\lambda c} \right) \right)$ over the $X_i's$ lower than ν which belongs to blocks with length smaller than M that is

$$\widehat{\mu}_n([-\infty, v[) = \frac{\sum_i^{l_n-1} \sum_{j=\tau_A(i)+1}^{\tau_A(i+1)} I_{\{\tau_A(i+1) - \tau_A(i) \leq M\}} I_{\{X_i \leq v\}} \left(1 - \exp\left(- \frac{(v - X_j)}{\lambda c} \right) \right)}{\sum_{i=1}^{l_n-1} (\tau_A(i+1) - \tau_A(i)) I_{\{\tau_A(i+1) - \tau_A(i) \leq M\}}}.$$

Note that the plug-in (non robust) estimator of μ in that case is simply

$$\mu_n([-\infty, v[) = 1 - n^{-1} \sum_{i=1}^{n} \exp\left(- \frac{(v - X_j)}{\lambda c} \right) I_{\{X_i \leq v\}}.$$

It is clear that this estimator is not robust to the presence of a large contaminated block.

Example 2. Similarly, in the KDEM model, using Kac's representation, we have the expression of the stationary measure of the continuous process given by

$$\mu([u,\infty[) = \lambda^{-1}\int_u^\infty \int_0^\infty \left(t \wedge \omega^{-1}\log(x/u)\right) \frac{\mathbb{E}_A\left(\sum_{i=1}^{\tau_A}\delta_{X_i}(dx)\right)}{\mathbb{E}_A[\tau_A]} H(dt), \quad u>0,$$

and the robust estimator is thus given by

$$\widehat{\mu}_n([u,\infty[) = \lambda^{-1}\int_u^\infty \int_0^\infty \left(t \wedge \omega^{-1}\log(x/u)\right) \tilde{F}_{\mathcal{L},M,n}(dx)H(dt)$$

$$= \frac{\sum_{i=1}^{l_n-1}\sum_{j=\tau_A(i)+1}^{\tau_A(i+1)} I_{\{\tau_A(i+1)-\tau_A(i)\leq M\}}\int_u^\infty\int_0^\infty \left(t\wedge\omega^{-1}\log(x/u)\right)\delta_{X_j}(dx)H(dt)}{\lambda\sum_{i=1}^{l_n-1}\left(\tau_A(i+1)-\tau_A(i)\right)I_{\{\tau_A(i+1)-\tau_A(i)\leq M\}}}$$

$$= \frac{\sum_{i}^{l_n-1}\sum_{j=\tau_A(i)+1}^{\tau_A(i+1)} I_{\{\tau_A(i+1)-\tau_A(i)\leq M\}}I_{\{X_j\geq u\}}\int_0^\infty \left(t\wedge\omega^{-1}\log(X_j/u)\right)H(dt)}{\lambda\sum_{i=1}^{l_n-1}\left(\tau_A(i+1)-\tau_A(i)\right)I_{\{\tau_A(i+1)-\tau_A(i)\leq M\}}}.$$

Again, in the exponential inter-arrival case, we have the expression

$$\int_0^\infty \left(t\wedge\omega^{-1}\log(X_j/u)\right)H(dt) = \lambda\left(1-(X_j/u)^{-1/(\omega\lambda)}\right).$$

Notice that in that case, the (non robust) plug-in estimator of $\mu([u,\infty[)$ is of the form

$$\mu_n([u,\infty[) = \frac{1}{n}\sum_{i=1}^n I_{\{X_j\geq u\}}\left(1-(X_j/u)^{-1/(\omega\lambda)}\right).$$

The robust estimator is simply the version of its mean only over the $X_i's$ which do not belong to large blocks, i.e.

$$\widehat{\mu}_n([u,\infty[) = \frac{\sum_{i=1}^{l_n-1}\sum_{j=\tau_A(i)+1}^{\tau_A(i+1)} I_{\{\tau_A(i+1)-\tau_A(i)\leq M\}}I_{\{X_j\geq u\}}\left(1-(X_j/u)^{-1/(\omega\lambda)}\right)}{\sum_{i=1}^{l_n-1}\left(\tau_A(i+1)-\tau_A(i)\right)I_{\{\tau_A(i+1)-\tau_A(i)\leq M\}}}.$$

Bibliographie

[1] Bertail P. and Clémençon, S. Regenerative block bootstrap for Markov chains, *Bernoulli*, vol. 12, p. 689-712, 2006.

[2] Bertail P. and Clémençon, S. Approximate regenerative block bootstrap for Markov chains, *Computational Statistics and Data Analysis*, vol. 52, p. 2739-2756, 2008.

[3] Bertail P., Clémençon S and Tressou, J. A storage model with random release rate for modelling exposure to food contaminants, *Math. Biosc. and Eng.*, vol. 5, p. 35-60, 2008.

[4] Mikosch, T. *Non life Insurance Mathematics*, Springer-Verlag, Berlin, 2010.

Processus ponctuels déterminantaux pour les coresets

Nicolas Tremblay & Simon Barthelmé & Pierre-Olivier Amblard

CNRS, GIPSA-lab, Grenoble-INP, Univ. Grenoble-Alpes, Grenoble
E-mail: prenom.nom@gipsa-lab.fr

Résumé. Face à une grande masse de données dont on veut apprendre certains paramètres d'un modèle sous-jacent, une solution possible pour accélérer l'apprentissage est l'échantillonnage: garder uniquement une partie des données et estimer la solution du problème initial en apprenant uniquement sur ce petit sous-ensemble. Les *coresets* sont de tels sous-ensembles qui, étant donnée une certaine tâche d'apprentissage, garantissent que l'estimation obtenue sur ce sous-ensemble est une bonne approximation de ce qu'on aurait obtenu sur le grand jeu initial de données, à une erreur relative près. Une des directions de l'état de l'art pour générer de tels *coresets* est d'échantillonner aléatoirement des éléments du jeu initial de manière iid, via une densité de probabilité particulièrement bien choisie (proportionnelle à une métrique appelée "sensitivité"). Les sous-ensembles obtenus par échantillonnage indépendent souffrent néanmoins de redondance: nous explorons ici comment les processus ponctuels déterminantaux, grâce entre autres à leur capacité à conserver la diversité d'un jeu de données, peuvent apporter des améliorations aux théorèmes iid existants.

Mots-clés. Processus ponctuels déterminantaux, coresets, apprentissage

Abstract. When faced with a data set too large to be processed all at once, an obvious solution is to retain only part of it. In practice this takes a wide variety of different forms, and among them "coresets" are especially appealing. A coreset is a (small) weighted sample of the original data that comes with the following guarantee: a cost function can be evaluated on the smaller set instead of the larger one, with low relative error. For some classes of problems, and via a careful choice of sampling distribution (based on the so-called "sensitivity" metric), iid random sampling has turned to be one of the most successful methods for building coresets efficiently. However, independent samples are sometimes overly redundant, and one could hope that enforcing diversity would lead to better performance. The difficulty lies in proving coreset properties in non-iid samples. We show that the coreset property holds for samples formed with determinantal point processes (DPP). DPPs are interesting because they are a rare example of repulsive point processes with tractable theoretical properties, enabling us to prove general coreset theorems.

Keywords. Determinantal point processes, coresets, learning

1 Introduction

Considérons une tâche d'apprentissage à résoudre sur une grande masse de données. Une des solutions possibles pour accélérer le processus d'apprentissage quand les données sont trop volumineuses est l'échantillonnage: garder uniquement une partie de l'information, jeter le reste, et estimer la solution du problème initial en apprenant seulement à partir de l'information conservée. Pour fixer les idées, supposons que $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$ est un grand ensemble de n points dans $\mathbb{R}^d$: c'est le corpus initial à échantillonner. Supposons que nous cherchons à résoudre un problème d'apprentissage sur $\mathcal{X}$ qui consiste à trouver le paramètre (ou jeu de paramètres) θ qui minimise une fonction de coût définie sous la forme particulière suivante:

$$L(\mathcal{X}, \theta) = \sum_{x \in \mathcal{X}} f(x, \theta)$$

où f est une fonction à valeurs dans $\mathbb{R}^+$. De nombreux problèmes classiques d'apprentissage s'écrivent sous cette forme, notamment k-means, la régression linéaire ou logistique, les machines à vecteurs de support, l'approximation en rang bas de matrices, etc. À titre d'exemple, pour k-means, le jeu de paramètres θ est un ensemble de k centroïdes en dimension d: $\theta = \{c_1, \ldots, c_k\}$ et la fonction de coût s'écrit:

$$L(\mathcal{X}, \theta) = \sum_{x \in \mathcal{X}} \min_{c \in \theta} ||x - c||^2.$$

Un objectif usuel en apprentissage est de calculer $\theta^{\text{opt}} = \text{argmin}_\theta \, L(\mathcal{X}, \theta)$. Ce problème de minimisation (ou même les heuristiques habituellement utilisées pour le résoudre approximativement) peut s'avérer trop lourd à calculer quand la taille n du jeu de données grandit. Un coreset est un sous-ensemble $\mathcal{S}$ des points de $\mathcal{X}$ pour lequel il est garanti que la solution du problème de minimisation sur $\mathcal{S}$ est une bonne estimation de θ^{opt}, à une erreur relative près. La litérature existante traîte de différentes techniques pour obtenir de tels coresets: de manière aléatoire ou déterministe, en cherchant à minimiser la taille du coreset ou bien en mettant plus l'accent sur l'efficacité de l'échantillonnage, etc. (voir par exemple la revue récente des techniques coresets Munteanu (2018))

Parmi les différentes directions explorées par la communauté, l'échantillonnage iid avec remise a le double avantage de produire des coresets qui sont de petite taille et rapides à échantillonner (une fois que l'on connaît la bonne distribution de probabilité à utiliser, qui elle peut s'avérer difficile à calculer). Les sous-ensembles obtenus par échantillonnage indépendent souffrent néanmoins de redondance. En effet, une fois que l'on a échantillonné un élément x de $\mathcal{X}$, rien ne nous empêche d'échantillonner un voisin y arbitrairement proche de x, qui n'apporterait pas d'information nouvelle à $\mathcal{S}$.

Nous explorons ici comment les processus ponctuels déterminantaux, grâce entre autres à leur capacité à conserver la diversité d'un jeu de données, peuvent apporter des améliorations aux théorèmes iid existants; et conséquemment inspirer de meilleurs algorithmes pour produire des coresets plus performants.

2 Coresets: définitions et état de l'art

Tout d'abord, précisons certaines notations. Notons $\mathcal{S} = \{x_{s_1}, \ldots, x_{s_m}\}$ un sous-ensemble de $\mathcal{X}$ de taille m (possiblement avec des répétitions). A chaque élément x_s de $\mathcal{S}$ on associe un poids non-négatif $\omega(x_s) \in \mathbb{R}^+$. On associe à un tel sous-ensemble pondéré $\mathcal{S}$ une fonction de coût estimée $\hat{L}$:

$$\hat{L}(\mathcal{S}, \theta) = \sum_{x_s \in \mathcal{S}} \omega(x_s) f(x_s, \theta).$$

Définition *Soit* $\epsilon \in]0, 1[$. *Un sous-ensemble pondéré* $\mathcal{S}$ *est un* ϵ-coreset *pour la fonction de coût* L *si, pour tout paramètre* θ, *le coût estimé est égal au vrai coût à une erreur relative près:*

$$\forall \theta \quad \left| \frac{\hat{L}}{L} - 1 \right| \leqslant \epsilon.$$

Cette définition est très contraignante[1] dans la mesure où l'erreur doit être contrôlée *pour tout* θ. Son intérêt vient des propriétés suivantes. Notons $\hat{\theta}^{\text{opt}} = \text{argmin}_\theta \hat{L}(\mathcal{S}, \theta)$ le paramètre optimal de la fonction de coût estimée $\hat{L}$. Si $\mathcal{S}$ est un ϵ-coreset pour L, on a:

$$(1-\epsilon)L(\mathcal{X}, \theta^{\text{opt}}) \leqslant (1-\epsilon)L(\mathcal{X}, \hat{\theta}^{\text{opt}}) \leqslant \hat{L}(\mathcal{S}, \hat{\theta}^{\text{opt}}) \leqslant \hat{L}(\mathcal{S}, \theta^{\text{opt}}) \leqslant (1+\epsilon)L(\mathcal{X}, \theta^{\text{opt}})$$

i.e., le coût estimé sur le coreset en $\hat{\theta}^{\text{opt}}$ est une approximation contrôlée (à une erreur relative ϵ près) du coût qu'on aurait obtenu sur $\mathcal{X}$ en θ^{opt}. Si bien que si L a un minimum suffisamment marqué autour de θ^{opt}, alors $\hat{\theta}^{\text{opt}}$ s'avèrera être une très bonne approximation de θ^{opt}. En d'autres termes, lancer un algorithme d'optimisation sur le coreset $\mathcal{S}$ donnera un résultat qui est une approximation contrôlée du vrai résultat (qu'on aurait obtenu en lançant l'algorithme sur l'ensemble des données). Une fois un coreset identifié, et pour autant que le coreset soit de petite taille, les gains en temps de calcul (pour résoudre le problème d'optimisation) peuvent être immenses! Avant d'évoquer l'état de l'art puis notre contribution dans ce domaine, définissons le concept de sensitivité.

Définition *La* sensitivité *d'un élément* x_i *de* $\mathcal{X}$ *pour la fonction de coût* L *est:*

$$\sigma_i = \max_\theta \frac{f(x_i, \theta)}{L(\mathcal{X}, \theta)} \quad \in [0, 1].$$

La somme de toutes les sensitivités est notée $\mathfrak{S} = \sum_i \sigma_i$.

L'état de l'art pour trouver des coresets est (entre autres) l'échantillonnage iid selon une densité de probabilité bien particulière. En effet, on a le théorème suivant:

[1] Il existe une version plus faible de cette définition, que nous ne considérerons pas, où il suffit que l'erreur soit contrôlée autour de $\theta = \theta^{\text{opt}}$

Théorème [Coresets avec échantillonnage iid, voir Langberg (2010) ou Bachem (2017)] *Soit* $\boldsymbol{p} \in [0,1]^n$ *une distribution de probabilité définie sur l'ensemble des points de* $\mathcal{X}$ *avec* p_i *la probabilité d'échantillonner* $\boldsymbol{x}_i$ *et* $\sum_i p_i = 1$. *Tirer aléatoirement* m *échantillons avec remise suivant* $\boldsymbol{p}$. *Associer à chaque échantillon* $\boldsymbol{x}_s$ *un poids*[2] $\omega(\boldsymbol{x}_s) = 1/mp_s$. *Le sous-ensemble pondéré obtenu est un* ϵ*-coreset avec probabilité supérieure à* $1-\delta$ *si* $m \geqslant m^*$ *avec:*

$$m^* = \mathcal{O}\left(\frac{1}{\epsilon^2}\left(\max_i \frac{\sigma_i}{p_i}\right)^2 (d' + \log(1/\delta))\right),$$

où d' *est la pseudo-dimension (une généralisation de la dimension de Vapnik-Chervonenkis) de* Θ, *l'espace des paramètres. La distribution de probabilité optimale qui minimise* m^* *est* $p_i = \sigma_i/\mathfrak{S}$. *Dans ce cas, la propriété coreset est vérifiée pour* $m \geqslant \mathcal{O}\left(\frac{\mathfrak{S}^2}{\epsilon^2}(d' + \log(1/\delta))\right)$.

À titre d'exemple, dans le contexte de k-means, $d' = dk\log k$ et $\mathfrak{S} = O(k)$, si bien qu'il suffit de $m = \mathcal{O}\left(\frac{dk^3 \log k}{\epsilon^2}\right)$ échantillons pour avoir un coreset avec grande probabilité: la taille du coreset minimal est indépendante de n! Cependant, ce remarquable résultat est rendu possible en relègant toute la difficulté du problème dans le calcul de la sensitivité, qui s'avère difficile[3]. En pratique, la communauté a développé des techniques d'encadrement des σ_i pour passer outre leur calcul exact, au prix de theorèmes –et donc d'algorithmes– moins puissants (voir par exemple Bachem (2017)).

3 Processus ponctuels déterminantaux: définitions et application aux coresets

L'idée générale de nos travaux est de passer d'une stratégie d'échantillonnage aléatoire iid à une stratégie avec dépendance pour éviter la redondance typique des sous-ensembles obtenus par tirage indépendant, et donc espérer générer des coresets de plus petite taille. Intuitivement, pour le problème coreset, nous avons envie d'un échantillonnage répulsif: si un élément x est échantillonné nous n'avons pas envie d'échantillonner un autre élément de $\mathcal{X}$ qui soit trop proche de x. Les stratégies d'échantillonnage répulsif sont très nombreuses. En revanche, il en existe très peu qui soient répulsives et tractables analytiquement (par exemple de constante de normalisation connue, de probabilités marginales à tous les ordres connues, etc.): les processus ponctuels déterminantaux (on utilisera le sigle anglais DPP) combinent ces deux propriétés, et c'est pourquoi nous les considérons.

[2] le lecteur reconnaîtra les poids usuels de l'échantillonnage par importance, qui assurent $\mathbb{E}(\hat{L}) = L$.

[3] avant nos travaux, la sensitivité n'avait même pas de forme analytique connue dans aucune des applications classiques mentionnées en introduction. Nous n'en parlerons pas ici mais les lemmes 23 et 25 de notre article Tremblay (2019) donnent la forme analytique de la sensitivité pour le problème 1-means ainsi que pour la régression linéaire.

L'objet central des DPPs est appelé L-ensemble, et n'est rien d'autre qu'une matrice semi-définie positive $\mathsf{L} \in \mathbb{R}^{n \times n}$ (à ne pas confondre avec la fonction coût L), que l'on supposera symmétrique ici. On écrit ses valeurs propres $0 \leqslant \lambda_1 \leqslant \lambda_2 \leqslant \ldots \leqslant \lambda_n$. Dans la suite, nous notons $2^{[n]}$ l'ensemble de tous les sous-ensembles des n premiers entiers.

Définition *Considérons un processus ponctuel, i.e., un processus qui tire aléatoirement un élément $\mathcal{S} \in 2^{[n]}$. Il est* déterminantal *avec L-ensemble L si*

$$\mathbb{P}(\mathcal{S}) = \frac{\det(\mathsf{L}_\mathcal{S})}{\det(\mathsf{I} + \mathsf{L})},$$

où $\mathsf{L}_\mathcal{S}$ est la restriction de L aux lignes et aux colonnes indexées par $\mathcal{S}$.

Les quelques propriétés suivantes sont bien connues, voir Kulesza (2012) pour des détails. Tout d'abord, la normalisation est bien correcte: $\sum_\mathcal{S} \det(\mathsf{L}_\mathcal{S}) = \det(\mathsf{I} + \mathsf{L})$. Aussi, toutes les probabilités d'inclusion, à tous les ordres, sont explicites:

$$\forall \mathcal{A} \in 2^{[n]} \qquad \mathbb{P}(\mathcal{A} \subseteq \mathcal{S}) = \det(\mathsf{K}_\mathcal{A})$$

où $\mathsf{K} = \mathsf{L}(\mathsf{I} + \mathsf{L})^{-1} \in \mathbb{R}^{n \times n}$ est appelé *noyau marginal*. En particulier, la probabilité d'échantillonner i, que l'on note génériquement π_i, vaut simplement K_{ii}. En outre, le caractère répulsif des DPPs est visible par exemple en considérant la probabilité jointe de tirer deux éléments i et j: $\mathbb{P}(\{i, j\} \subseteq \mathcal{S}) = \det(\mathsf{K}_{\{i,j\}}) = \pi_i \pi_j - \mathsf{K}_{ij}^2 \leqslant \pi_i \pi_j$ la probabilité jointe du processus de Poisson (indépendant) associé.

On peut également mentionner que le nombre d'éléments d'un DPP est lui-même aléatoire et distribué selon une somme indépendante de n lois de Bernoulli de paramètres $\{\frac{\lambda_i}{1+\lambda_i}\}$. Dans de nombreux cas pratiques, l'utilisateur préfère spécifier de manière déterministe le nombre d'échantillons, ce qui a ammené à la définition des m-DPPs: des DPPs conditionnés à échantillonner m éléments. Les m-DPPs sont plus utiles en pratique. En revanche, ils sont moins tractables. En particulier, il n'existe plus en général de noyau marginal pour calculer les probabilités d'inclusion. Dans d'autres travaux, voir Barthelmé (2019), nous avons développé des techniques approchées pour les calculer efficacement.

Nous terminons cette suite de définitions avec les DPPs de projection:

Définition *Un* DPP de projection *est un m-DPP dont le L-ensemble est une projection de rang m: $\mathsf{L} = \mathsf{U}\mathsf{U}^\top$, où $\mathsf{U} \in \mathbb{R}^{n \times m}$ vérifie $\mathsf{U}^\top \mathsf{U} = \mathsf{I}_m$.*

Lemme *Un DPP de projection de L-ensemble L est aussi un DPP, de noyau marginal L.*

Ce lemme, issu de Barthelmé (2019), explique pourquoi les DPPs de projection sont des objets très utiles: ils ont à la fois le côté pratique des m-DPPs (un nombre d'échantillons fixe) et la simplicité analytique des DPPs (par exemple π_i est simplement L_{ii}, *i.e.*, la somme des carrés de la i-ème ligne de U).

Dans Tremblay (2019), nous détaillons un ensemble de résultats sur l'utilisation des DPPs au problème des coresets, dont nous reportons ci-dessous un seul morceau

choisi. Nous notons $\hat{L}_{iid}$ l'estimateur de L associé au sous-ensemble $\mathcal{S}$, obtenu par échantillonnage iid de m éléments selon une densité de probabilité $\boldsymbol{p}$ et pondéré par les poids d'échantillonnage d'importance, comme expliqué dans le théorème de la section précédente. Nous comparons cet estimateur avec l'estimateur $\hat{L}$ associé au sous-ensemble $\mathcal{S}$ obtenu par DPP de projection de L-ensemble $\mathsf{L} = \mathsf{U}\mathsf{U}^\top$ où $\mathsf{U} \in \mathbb{R}^{n\times m}$ est tel que i/ $\mathsf{U}^\top\mathsf{U} = \mathsf{I}_m$ (pour que le DPP associé soit bien un DPP de projection), ii/ la probabilité marginale d'échantillonner i, $\pi_i = \mathsf{L}_{ii} = \sum_j \mathsf{U}_{ij}^2$, est fixée[4] à mp_i (cela permet de se comparer équitablement au cas iid). On montre dans Tremblay (2019) qu'un tel U existe nécessairement, et qu'il en existe en général de nombreux. On a:

Théorème *La variance de l'estimateur $\hat{L}$ est nécessairement inférieure à celle de l'estimateur iid équivalent. En particulier:*

$$\forall\theta \in \Theta \qquad \mathrm{Var}(\hat{L}) = \mathrm{Var}(\hat{L}_{\mathrm{iid}}) - \frac{m-1}{m}\left\|\sum_i \frac{f(\boldsymbol{x}_i,\theta)}{mp_i}\tilde{\boldsymbol{v}}_i\right\|^2,$$

où $\tilde{\boldsymbol{v}}_i \in \mathbb{R}^{m^2-m}$ est le vecteur diagramme (cf. Copenhaver (2014)) de la i-ème ligne de U.

En d'autres termes, il est toujours préférable d'échantillonner des coresets via un DPP de projection que de manière iid. Ce résultat se décline de différentes manières, par exemple via des théorèmes coresets généraux de la forme de celui présenté plus haut. En contrepartie d'une meilleure performance des DPPs (vérifiée expérimentalement également), échantillonner un DPP est plus coûteux algorithmiquement qu'échantillonner de manière iid. En effet, dans le cas d'un DPP de projection de rang m, le coût d'échantillonnage coûte $\mathcal{O}(nm^2)$. Nous pointons le lecteur intéressé à notre article Tremblay (2019) pour de plus amples détails.

Bibliographie

O. Bachem, M. Lucic, A. Krause. Practical Coreset Constructions for Machine Learning. *arXiv:1703.06476 [stat]*, 2017.

S. Barthelmé, P.-O. Amblard, N. Tremblay. Asymptotic equivalence of fixed-size and varying-size determinantal point processes. *Bernoulli*, 25(4B):3555–3589, 2019.

M. Copenhaver, Y. Kim, C. Logan, K. Mayfield, S. Narayan, M. Petro, J. Sheperd. Diagram vectors and tight frame scaling in finite dimensions. *Operators and Matrices*, 8(1):73–88, 2014.

A. Kulesza, B. Taskar. Determinantal Point Processes for Machine Learning. *Foundations and Trends in Machine Learning*, 5(23):123–286, 2012.

M. Langberg, L. Schulman. Universal ϵ-approximators for integrals. In *Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms*, pp. 598–607, 2010.

A. Munteanu, C. Schwiegelshohn. Coresets-Methods and History: A Theoreticians Design Pattern for Approximation and Streaming Algorithms. *Künstliche Intelligenz*, 32, pp.37–53, 2018.

N. Tremblay, S. Barthelmé, P.-O. Amblard. Determinantal point processes for coresets. *Journal of Machine Learning Research*, 20(168):1–70, 2019.

[4] pour que ce soit soit bien défini, on suppose que toute entrée de $\boldsymbol{p}$ est inférieure à $1/m$

On the Distribution of the Weighted Sum of Chi-Squared Variables

Ayşe Ünsal & Raymond Knopp

EURECOM, Sophia Antipolis, France
firstname.lastname@eurecom.fr

Résumé. Cet article prsente la densité (p.d.f.) et la fonction de répartition (c.d.f.) dune somme de variables alatoires chi-carré indépendantes centrales pondérées par des poids non-nuls. La méthode utilise les fonctions génératrices de moments. Pour obtenir le p.d.f. et c.d.f. d'une telle somme, nous dérivons d'abord la fonction génératrice de moments de cette somme pondérée en utilisant le développement de fraction partielle et la méthode des résidus. Les résultats couvrent les sommes de deux et trois variables aléatoires chi-carré pondérées, et se généralise facilement pour des cas plus généraux.

Mots-clés. Fonction génératrice des moments, chi-carré, décomposition en fractions partielles

Abstract. This paper presents the probability distribution function (p.d.f.) and cumulative distribution function (c.d.f.) of the weighted sum of central independent chi-squared random variables with non-zero weighs based on a method using moment generating functions. To obtain the p.d.f. and c.d.f. of such a function, we first derive the moment generating function of this weighted sum using the partial fractions decomposition or the residue method. The results cover the sum of two and three weighted chi-squared random variables, which can easily be adapted to more general cases.

Keywords. moment generating function, chi-squares, partial fractions, residue method

1 Introduction

Statistical distributions come up in various application areas of the probability theory. Information theory, which studies and analyzes the fundamental limits of communication systems through probabilistic tools, is one of these areas where mathematical statistics plays an important role. For instance, in *network information theory*, where the communication system is composed of multiple transmitters, multiple receivers or both, following normal distributions with various parameters, information theorists encounter a distribution in the form of a finite linear combination of independent central chi-square random variables in their analyses.

In this work, we focus on the derivation of the p.d.f. and c.d.f. for such a linear function defined as follows

$$f(x) = \sum_j \lambda_j \mu^2_{(n_j)}, \text{ for } j = 2, 3, \cdots \tag{1}$$

where $\mu^2_{(n_j)}$'s denote independent chi-squared random variables with n degrees of freedom and λ_j's are non-zero and real factors. In this paper, we derive the *exact* distribution of (1) for $j = 2$ and $j = 3$, through translating this expression into a sum of partial fractions. This method can easily be generalized to any value of j.

1.1 Related Work

In [3], the authors present the cumulative distribution function of a linear combination of independent central chi-square random variables through translation of its moment generating function into an infinite gamma series. The main difference between [3] and the current work is that in (1) the weighs represented by the terms λ_j's are not limited to be positive. A less recent work on the same problem in [1], considers the case where $\sum \lambda_j = 1$ in addition to the condition of positive factors. [2] however, focuses on the distribution of the quadratic forms defined as $\sum_{i=1}^{n} \lambda_j (\mu_{j,i} + a_{j,i})^2$ where $a_{j,i} > 0$. In [2], the authors propose significance points for several values of j and present two new approximations using the first three and four moments of the considered quadratic form. [4] and [5] are focused on series representations of quadratic forms for the vector consisting multivariate normal variables in different scenarios. In a quite related and rather recent study [6], the authors characterize the distribution of positive quadratic forms of normal random variables deriving Laguerre expansions for the probability and cumulative distribution functions which is based on the inverse Laplace transforms.

2 The Partial Fractions of the Moment Generating Function

The random variables denoted $\mu^2_{(n_j)}$ in (1) has the following moment generating function of the chi-squared distribution with n degrees of freedom

$$M(t) = (1 - 2t)^{-n/2}. \tag{2}$$

The sum of chi-squared variables that are weighted by the eigenvalues of the quadratic forms as given by (1) may occur in various scenarios one of which is the derivation of the mutual information function for finite n. When the weighs (i.e., λ_j's) are equal to 1, sum of the chi-squared variables correspond to a gamma distribution with $n/2$ degrees of freedom. Here we have a special case of the gamma distribution since some of the factors

are allowed to be negative and the gamma distribution is defined positive. Let us consider the simplest scenario for $j = 2$, the corresponding moment generating function of the sum of two chi-squared distributed variables with weighs denoted by λ_1 and λ_2 is given as

$$M_g(t) = [1 - 2\lambda_1 t]^{-n/2}[1 - 2\lambda_2 t]^{-n/2} \tag{3}$$

One could imagine $M_g(t)$ in a form of the sum of partial fractions as follows

$$M_g(t) = \sum_{i=1}^{n} \left(A_i[1 - 2\lambda_1 t]^{-n/2+i-1} + B_i[1 - 2\lambda_2 t]^{-n/2+i-1} \right) \tag{4}$$

Multiplying both sides of (4) by $[1 - 2\lambda_1 t]^{n/2}$, we have

$$M_g(t)[1 - 2\lambda_1 t]^{n/2} = \sum_{i=1}^{n} \left(A_i[1 - 2\lambda_1 t]^{i-1} + B_i[1 - 2\lambda_1 t]^{n/2}[1 - 2\lambda_2 t]^{-n/2+i-1} \right) \tag{5}$$

Taking the derivatives of both sides of (5) upto $i - 1$, the partial fraction coefficients A_i and B_i are respectively derived as follows.

$$\begin{aligned} A_i &= \frac{(-2\lambda_1)^{1-i}}{(i-1)!} \frac{d^{i-1}}{dt^{i-1}}[1 - 2\lambda_1 t]^{n/2} M_g(t)\Big|_{t=1/2\lambda_1} \\ &= \frac{(\lambda_1/\lambda_2)^{1-i}}{(i-1)!} \left(\prod_{j=1}^{i-1} -\frac{n}{2} - (j-1) \right) (1 - \lambda_2/\lambda_1)^{-n/2-i+1} \end{aligned} \tag{6}$$

$$\begin{aligned} B_i &= \frac{(-2\lambda_2)^{1-i}}{(i-1)!} \frac{d^{i-1}}{dt^{i-1}}[1 - 2\lambda_2 t]^{n/2} M_g(t)\Big|_{t=1/2\lambda_2} \\ &= \frac{(\lambda_2/\lambda_1)^{1-i}}{(i-1)!} \left(\prod_{j=1}^{i-1} -\frac{n}{2} - (j-1) \right) (1 - \lambda_1/\lambda_2)^{-n/2-i+1} \end{aligned} \tag{7}$$

The resulting probability distribution and cumulative distribution functions for $j = 2$ and $x \geq 0$ are

$$f_2(x) = \sum_{i=1}^{n} \left(\frac{A_i}{\Gamma\left(\frac{n}{2} - i + 1\right) (2\lambda_1)^{\frac{n}{2}-i+1}} \mathrm{e}^{-\frac{x}{2\lambda_1 i}} + \frac{B_i}{\Gamma\left(\frac{n}{2} - i + 1\right) (2\lambda_2)^{\frac{n}{2}-i+1}} \mathrm{e}^{-\frac{x}{2\lambda_2 i}} \right) x^{\frac{n}{2}-i}, \tag{8}$$

$$F_2(x) = \sum_{i=1}^{n} \left(\frac{A_i}{\Gamma\left(\frac{n}{2} - i + 1\right)} \gamma\left(\frac{n}{2} - i + 1, \frac{x}{2\lambda_1}\right) + \frac{B_i}{\Gamma\left(\frac{n}{2} - i + 1\right)} \gamma\left(\frac{n}{2} - i + 1, \frac{x}{2\lambda_2}\right) \right), \tag{9}$$

respectively. A_i and B_i are respectively given by (6) and (7), where the gamma function $\Gamma(.)$ is defined as

$$\Gamma(a) = \int_0^\infty t^{a-1} \mathrm{e}^{-t} dt \tag{10}$$

and $\gamma(.)$ denoting the incomplete function that is $\Gamma_x(a) = \int_0^x t^{a-1} \mathrm{e}^{-t} dt$.

2.1 Three terms case

Imagine the case where $j = 3$ in (1). The corresponding moment generating function in this scenario denoted $M_{g_3}(t)$ becomes

$$M_{g_3}(t) = [1 - 2\lambda_1 t]^{-n/2}[1 - 2\lambda_2 t]^{-n/2}[1 - 2\lambda_3 t]^{-n/2} \tag{11}$$

(11) can be rewritten through its partial fractions as follows

$$M_{g_3}(t) = \sum_{i=1}^{n/2} \left\{ A_i[1 - 2\lambda_1 t]^{-n/2+i-1} + B_i[1 - 2\lambda_2 t]^{-n/2+i-1} + C_i[1 - 2\lambda_3 t]^{-n/2+i-1} \right\} \tag{12}$$

where the terms A_i, B_i and C_i are respectively given by

$$\begin{aligned}
A_i &= \frac{(-2\lambda_1)^{1-i}}{(i-1)!} \frac{d^{i-1}}{dt^{i-1}}[1 - 2\lambda_1 t]^{\frac{n}{2}} M_{g_3}(t)\Big|_{t=1/2\lambda_1} \\
&\overset{(a)}{=} \frac{(-2\lambda_1)^{1-i}}{(i-1)!} \sum_{k=0}^{i-1} \binom{i-1}{k} \frac{d^{i-1-k}}{dt^{i-1-k}}[1 - 2\lambda_2 t]^{i-1-k} \frac{d^k}{dt^k}[1 - 2\lambda_3 t]^k|_{t=1/2\lambda_1} && (13) \\
&= \frac{(\lambda_1/\lambda_2)^{1-i}}{(i-1)!} \sum_{k=0}^{i-1} \left[\binom{i-1}{k} \left(\prod_{j=1}^{i-1-k} -\frac{n}{2} - j + 1 \right) \left(1 - \frac{\lambda_2}{\lambda_1}\right)^{-\frac{n}{2}-(i-1-k)} \right. && (14) \\
&\qquad \left. \left(\prod_{j=1}^{k} -\frac{n}{2} - j + 1 \right) \left(1 - \frac{\lambda_3}{\lambda_1}\right)^{-\frac{n}{2}-k} \left(\frac{\lambda_3}{\lambda_2}\right)^k \right] && (15)
\end{aligned}$$

$$\begin{aligned}
B_i &= \frac{(-2\lambda_2)^{1-i}}{(i-1)!} \frac{d^{i-1}}{dt^{i-1}}[1 - 2\lambda_2 t]^{n/2} M_{g_3}(t)\Big|_{t=1/2\lambda_2} \\
&\overset{(b)}{=} \frac{(-2\lambda_2)^{1-i}}{(i-1)!} \sum_{k=0}^{i-1} \binom{i-1}{k} \frac{d^{i-1-k}}{dt^{i-1-k}}[1 - 2\lambda_1 t]^{i-1-k} \frac{d^k}{dt^k}[1 - 2\lambda_3 t]^k|_{t=1/2\lambda_2} && (16) \\
&= \frac{(\lambda_2/\lambda_1)^{1-i}}{(i-1)!} \sum_{k=0}^{i-1} \left[\binom{i-1}{k} \left(\prod_{j=1}^{i-1-k} -\frac{n}{2} - j + 1 \right) \left(1 - \frac{\lambda_1}{\lambda_2}\right)^{-\frac{n}{2}-(i-1-k)} \right. && (17) \\
&\qquad \left. \left(\prod_{j=1}^{k} -\frac{n}{2} - j + 1 \right) \left(1 - \frac{\lambda_3}{\lambda_2}\right)^{-\frac{n}{2}-k} \left(\frac{\lambda_3}{\lambda_1}\right)^k \right] && (18)
\end{aligned}$$

$$C_i = \frac{(-2\lambda_3)^{1-i}}{(i-1)!}\frac{d^{i-1}}{dt^{i-1}}[1-2\lambda_3 t]^{n/2}M_{g_3}(t)\Big|_{t=1/2\lambda_3}$$

$$\overset{(c)}{=} \frac{(-2\lambda_3)^{1-i}}{(i-1)!}\sum_{k=0}^{i-1}\binom{i-1}{k}\frac{d^{i-1-k}}{dt^{i-1-k}}[1-2\lambda_1 t]^{i-1-k}\frac{d^k}{dt^k}[1-2\lambda_2 t]^k|_{t=1/2\lambda_3} \tag{19}$$

$$= \frac{(\lambda_3/\lambda_1)^{1-i}}{(i-1)!}\sum_{k=0}^{i-1}\left[\binom{i-1}{k}\left(\prod_{j=1}^{i-1-k}-\frac{n}{2}-j+1\right)\left(1-\frac{\lambda_1}{\lambda_3}\right)^{-\frac{n}{2}-(i-1-k)}\right. \tag{20}$$

$$\left.\left(\prod_{j=1}^{k}-\frac{n}{2}-j+1\right)\left(1-\frac{\lambda_2}{\lambda_3}\right)^{-\frac{n}{2}-k}\left(\frac{\lambda_2}{\lambda_1}\right)^k\right] \tag{21}$$

In steps (a), (b) and (c), we used the following general Leibniz rule for the n^{th} derivative of a product that is

$$(fg)^n(x) = \sum_{k=0}^{n}\binom{n}{k}f^{n-k}(x)g^k(x) \tag{22}$$

Finally, the resulting probability distribution and cumulative distribution functions for $j=3$ are

$$f_3(x) = \sum_{i=1}^{n}\left(\frac{A_i}{\Gamma\left(\frac{n}{2}-i+1\right)(2\lambda_1)^{\frac{n}{2}-i+1}}\mathrm{e}^{-\frac{x}{2\lambda_1 i}}x^{\frac{n}{2}-i} + \frac{B_i}{\Gamma\left(\frac{n}{2}-i+1\right)(2\lambda_2)^{\frac{n}{2}-i+1}}\mathrm{e}^{-\frac{x}{2\lambda_2 i}}x^{\frac{n}{2}-i}\right.$$
$$\left. + \frac{C_i}{\Gamma\left(\frac{n}{2}-i+1\right)(2\lambda_3)^{\frac{n}{2}-i+1}}\mathrm{e}^{-\frac{x}{2\lambda_3 i}}x^{\frac{n}{2}-i}\right), \tag{23}$$

$$F_3(x) = \sum_{i=1}^{n}\left(\frac{A_i}{\Gamma\left(\frac{n}{2}-i+1\right)}\gamma\left(\frac{n}{2}-i+1,\frac{x}{2\lambda_1}\right) + \frac{B_i}{\Gamma\left(\frac{n}{2}-i+1\right)}\gamma\left(\frac{n}{2}-i+1,\frac{x}{2\lambda_2}\right)\right.$$
$$\left. + \frac{C_i}{\Gamma\left(\frac{n}{2}-i+1\right)}\gamma\left(\frac{n}{2}-i+1,\frac{x}{2\lambda_3}\right)\right), \tag{24}$$

for $x \geq 0$, respectively. The gamma $\Gamma(.)$ and incomplete gamma $\gamma(.)$ functions are reminded above. Partial fraction factors A_i, B_i and C_i are derived in (13)-(19), respectively.

3 Numerical Evaluation Results

Figure 1 presents the c.d.f. given in (9) with different positive and negative weighs as given in the legend.

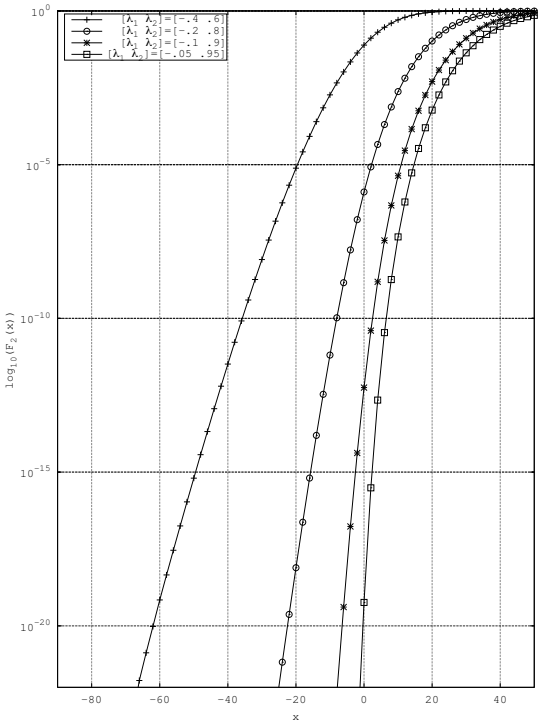


Figure 1: Numerical evaluation of $F_2(x)$ for different pairs of λ_1, λ_2 and $n = 50$.

References

[1] Grad, A. and Solomon, H. (1955). Distribution of Quadratic Forms and Some Applications *The Annals of Mathematical Statistics* **3** 464–477.

[2] Solomon, H. and Stephens, M. A. (1977) Distribution of a Sum of Weighted Chi-Square Variables *Technical Report, Stanford University*

[3] Moschopoulos, P.G. and Canada, W.B. (1984). The Distribution Function of a Linear Combination of Chi-Squares *Comp. & Maths. with Appls.* **10** 383–386.

[4] Kotz, S. and Johnson, N.L. and Boyd, D.W.(1967). Series Representations of Distributions of Quadratic Forms in Normal Variables. I. Central Case *The Annals of Mathematical Statistics* Vol. 38, No. 3 823–837.

[5] Kotz, S. and Johnson, N.L. and Boyd, D.W.(1967). Series Representations of Distributions of Quadratic Forms in Normal Variables. II. Non-Central Case *The Annals of Mathematical Statistics* Vol. 38, No. 3 838–848.

[6] Castaño-Martínez, A. and López-Blázquez, F.(2005). Distribution of a sum of weighted noncentral chi-square variables *TEST* Vol. 14, No. 2 397–415.

Détection de la périodicité du milieu aléatoire par l'observation d'une seule trajectoire d'une marche aléatoire

Jean Vaillancourt [1] & Bruno N. Rémillard [2]

[1] *HEC Montréal, 3000, chemin de la Côte-Sainte-Catherine, Montréal (Québec) Canada H3T 2A7 jean.vaillancourt@hec.ca*

[2] *HEC Montréal, 3000, chemin de la Côte-Sainte-Catherine, Montréal (Québec) Canada H3T 2A7 bruno.remillard@hec.ca*

Résumé. Pour les marches aléatoires simples univariées en milieu périodique, là où la probabilité de déplacement dépend d'une fonction périodique, nous montrons comment estimer la période et la fonction elle-même. Pour les marches aléatoires en milieu non-périodique, nous montrons que la limite asymptotique de l'estimateur est constante dans le cas ballistique, lorsque la marche aléatoire est transitoire et que la loi des grands nombres est valide avec une limite non nulle. Quelques exemples numériques sont offerts dans le cas récurrent, ainsi que dans le cas sous-ballistique (lorsque la marche aléatoire est transitoire mais que la loi des grands nombres a une limite nulle).

Mots-clés. Marche aléatoire, milieu aléatoire, fonctions périodiques, théorie ergodique

Abstract. For nearest neighbour univariate random walks in a periodic environment, where the probability of moving depends on a periodic function, we show how to estimate the period and the function. For random walks in non-periodic environments, we find that the asymptotic limit of the estimator is constant in the ballistic case, when the random walk is transient and the law of large numbers holds with a non zero limit. Numerical examples are given in the recurrent case, as well as in the sub-ballistic case (where the random walk is transient but the law of large numbers yields a zero limit).

Keywords. Random walks, Random environments, Periodic functions, Ergodic theory

1 Introduction

En l'absence d'information autre que l'observation de la trajectoire d'une marche aléatoire en milieu aléatoire (MAMA), on souhaite estimer la loi du milieu ambiant. Bien que, en règle générale, cela soit impossible, il y a des situations simples où plusieurs choix de modèles se prêtent bien à l'identification de la fonction de répartition, au moins asymptotiquement. Après une brève revue de la littérature spécifique à ce problème, notre

propos s'oriente uniquement vers la détection de périodicité à partir d'une seule et unique trajectoire d'une MAMA. Introduisons tout d'abord la notation requise.

Étant donné un espace probabilisé complet $(E, \mathcal{E}, P)$ sur lequel toutes les variables aléatoires sont construites, un milieu aléatoire est une suite bidirectionnelle $\alpha(e) = \alpha(\cdot, e) \in (0,1)^{\mathbb{Z}}$ indicée par $e \in E$ et on écrit $\mu = P \circ \alpha^{-1}$ pour sa répartition sur les boréliens de $[0,1]^{\mathbb{Z}}$. On écrit $\rho_i(e) = \frac{1-\alpha(i,e)}{\alpha(i,e)}$ tout en omettant e si le sens ne se perd pas. Soit $X = \{X_t\}_{t\geq 0}$ une MAMA sur $\mathbb{Z}$, à savoir, telle que

$$\mathbb{P}^e\left(\Delta X_t = X_t - X_{t-1} = 1 |\, X_{t-1} = i\right) = \mathbb{P}\left(X_t = i+1 |\, X_{t-1} = i, \mathcal{E}\right)(e) = \alpha(i,e), \quad (1.1)$$

avec $\mathbb{P}^e(\Delta X_t = -1 |\, X_{t-1} = i) = 1 - \alpha(i,e)$, pour tout choix de $i \in \mathbb{Z}$ et tout entier positif t. Conditionnellement à la réalisation e du milieu et au point de départ X_0 de la marche, les sites visités successivement $\{X_t : t \geq 1\}$ forment une chaîne de Markov homogène sur $\mathbb{Z}$ sous $\mathbb{P}^e$ et la loi conditionnelle de la marche entière est une probabilité $\mathbb{P}^e$ (sur l'ensemble puissance de $\mathbb{Z}^{\mathbb{N}}$) connue sous le vocable de loi trempée (quenched law) pour la MAMA. Le processus est donc encapsulé par la famille de lois conjointes $\mathcal{P}_\mu$ définies par $\mathcal{P}_\mu(F \times A) = \int_A \mathbb{P}^e(F)\ \mu(de)$. Notez que X n'est généralement pas un processus de Markov sous $\mathcal{P}_\mu$.

Afin de pouvoir utiliser ce genre de modèle, il est nécessaire de pouvoir estimer le processus $\mathcal{A} = \{\alpha(i)\}_{i\in\mathbb{Z}}$. Ce n'est bien sûr pas possible en général, mais en imposant une structure paramétrique il est possible d'y parvenir. Par exemple, Comets et al (2014) étudient les distributions asymptotiques pour un M-estimateur de leur choix qui s'apparente à celui de vraisemblance maximale pour la solution de (1.1). Les auteurs font l'hypothèse que les $\{\alpha(i)\}_{i\in\mathbb{Z}}$ sont indépendantes et identiquement distribuées (iid) avec des marges paramétrées, dans le cas dit balistique donc transitoire vers la droite avec $\mathbb{E}(\log \rho_0) < 0$ et $\mathbb{E}(\rho_0) < 1$. Le paramètre d'intérêt est estimé via un M-estimateur basé sur la suite $X_0, \ldots, X_{\tau_n}$, où τ_n est le temps de la première visite n. Le cas sous-balistique correspondant, i.e., $\mathbb{E}(\log \rho_0) < 0$ et $\mathbb{E}(\rho_0) \geq 1$, est analysé dans Falconnet et al (2014), tandis que le cas récurrent $\mathbb{E}(\log \rho_0) = 0$ fait l'objet de Comets et al (2016), sous la contrainte additionnelle d'un support fini pour la suite $\{\alpha(i)\}_{i\in\mathbb{Z}}$. Récemment, Diel et Lerasle (2018) ont construits des estimateurs nonparamétriques compétitifs. Le traitement du cas balistique est étendu dans Andreoletti et al (2015) à une classe de modèles de Markov cachés pour les $\{\alpha(i)\}_{i\in\mathbb{Z}}$. Historiquement Adelman et Enriquez (2004) ont proposé la solution initiale à ce problème, basée sur les moments du milieu aléatoire. Dans tous ces exemples, bien que $\{\alpha(i)\}_{i\in\mathbb{Z}}$ soit un processus stochastique très simple, la MAMA résultante X de (1.1) n'est jamais un processus de Markov.

Tout milieu périodique connu permet la construction d'un estimateur du maximum de vraisemblance (EMV). C'est notre estimateur de choix. Examinons tout d'abord son comportement dans un milieu périodique quelconque.

2 Estimation d'un milieu périodique

Pour chaque $p \in \mathcal{P}_d$, l'ensemble des fonctions périodiques sur $\mathbb{Z}$ à valeurs dans $(0,1)$ et de période d, dénotons les milieux résultants par $E_p = \{e_j = p(\cdot + j); 1 \leq j \leq d\}$. Ainsi donc $\{\alpha(\cdot, e_j) = p(\cdot + j); j \geq 0\}$ est une suite déterministe (sauf lorsque le point de départ $e_0 \in E_p$ est aléatoire) avec une seule probabilité invariante: la loi uniforme sur E_p. Ce cas particulier de MAMA constitue une marche aléatoire en milieu périodique (MAMP) au sens de Pyke (2003). Autrement dit, X est une MAMP s'il existe un $d \in \mathbb{N}$ et un $p \in \mathcal{P}_d$ tels que $X = \{X_t\}_{t \geq 0}$ soit la marche aléatoire simple définie par

$$\mathbb{P}\left(\Delta X_t = 1 |\ X_{t-1} = i\right) = p(i),\ i \in \mathbb{Z},\ t \geq 1. \tag{2.1}$$

Évidemment si $p \in \mathcal{P}_d$ alors $p \in \mathcal{P}_{kd}$ pour tout $k \in \mathbb{N}$. On cherche à estimer la fonction p ainsi que le plus petit d_0 tel que $p \in \mathcal{P}_{d_0}$, en n'observant que les n premiers pas de la trajectoire $X_0, \ldots, X_n$ de la MAMP X.

On écrit $(x)_d = i$ pour $x = i \mod (d)$, la classe des résidus modulo d. On remarque que le processus $\{(X_t)_d\}_{t \geq 0}$ à valeurs dans $S_d = \{1, \ldots, d\}$ n'est markovien que si et seulement si d est un multiple de la valeur effective d_0.

2.1 Estimation de p si $d = d_0$ est connue

Si $(X_t)_{t=0}^n$ satisfait à (2.1) avec $p \in \mathcal{P}_d$, l'EMV $p_n^{(d)}$ de p prend la forme

$$p_n^{(d)}(j) = A_{n,j}^{(d)} / \left(A_{n,j}^{(d)} + B_{n,j}^{(d)}\right), \quad j \in \{1, \ldots, d\}, \tag{2.2}$$

avec $A_{n,j}^{(d)} = \sum_{t=1}^n \mathbb{I}\{(X_{t-1})_d = j, \Delta X_t = 1\}$, $B_{n,j}^{(d)} = \sum_{t=1}^n \mathbb{I}\{(X_{t-1})_d = j, \Delta X_t = -1\}$ et $A_{n,j}^{(d)} + B_{n,j}^{(d)} = \sum_{t=1}^n \mathbb{I}\{(X_{t-1})_d = j\}$. La fonction Log-vraisemblance associée est

$$L_{n,d} \ = \ -\sum_{j=1}^d \left(A_{n,j}^{(d)} + B_{n,j}^{(d)}\right) H\left\{p_n^{(d)}(j)\right\}, \tag{2.3}$$

avec $H(x) = -x \log x - (1-x)\log(1-x) \geq 0$ l'entropie de Boltzmann pour la loi de Bernoulli de paramètre x. La convergence de l'EMV procède normalement ici, voir Rémillard et Vaillancourt (2019) pour le détail.

Proposition 2.1 *Si $p \in \mathcal{P}_d$ et π est l'unique loi invariante de la chaîne de Markov $(X_t)_d$ sur S_d associée à p, alors, quand $n \to \infty$ on obtient, pour tout $j \in S_d$, μ presque sûrement, la convergence de l'EMV $p_n^{(d)}(j) \to p(j)$ ainsi que de sa Log-vraisemblance $L_{n,d}\Big/n \to \mathcal{L}_d = -\sum_{j=1}^d \pi_j H\{p(j)\}$.*

En pratique, le défi est de déterminer la vraie valeur d_0 de d. L'illustration suivante permet de voir ce dont il en ressort.

Exemple. On engendre $n = 10000$ pas d'une marche aléatoire satisfaisant à (2.1), démarrée en $X_0 = 100$, avec $p = (0.099, 0.749, 0.749)$. Les résultats de l'estimation pour $d_0 = 3$ sont $p_n = (0.0954, 0.7593, 0.7430) \in \mathcal{P}_3$, avec une Log-vraisemblance de -4.6976×10^3. La Figure 1 montre le comportement de la Log-vraisemblance pour les périodes $d \in \{1, \ldots, 10\}$. Les maxima locaux sont atteints aux multiples de $d = 3$. Dans ces cas, le pointillé horizontal en $\mathcal{L}_3$ suggère bien la convergence de $L_{n,d}/n$ vers cette valeur quand n est assez grand.

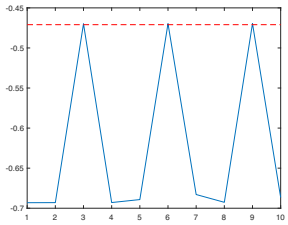


Figure 1: Graphe de $L_{n,d}/n$ (solide) et $\mathcal{L}_3$ (pointillé), pour $d \in \{1, \ldots, 10\}$.

2.2 Estimation de la période minimale

Sous $p \in \mathcal{P}_{d_0}$, qu'advient-il de $p_n^{(d)}$ quand $d \neq d_0$?

Proposition 2.2 *Soient $p \in \mathcal{P}_{d_0}$ avec d_0 minimale et π l'unique loi invariante de la chaîne de Markov $(X_t)_{d_0}$. Si $m = (d, d_0)$ est le plus grand commun diviseur de d et d_0, alors, quand $n \to \infty$ et pour tout $i \in S_d$,*

$$p_n^{(d)}(i) \overset{p.s.}{\to} p^{(d)}(i) = \sum_{j \in S_{d_0},\, (j)_m = (i)_m} \pi_j p(j) \Big/ \sum_{j \in S_{d_0},\, (j)_m = (i)_m} \pi_j, \tag{2.4}$$

$$L_{n,d} \Big/ n \overset{p.s.}{\to} \mathcal{L}_d = -\frac{m}{d} \sum_{i=1}^{d} \sum_{j \in S_{d_0},\, (j)_m = (i)_m} \pi_j H\left\{p^{(d)}(i)\right\}. \tag{2.5}$$

En particulier, si $m = 1$, alors $p_n^{(d)}(i)$ converge presque sûrement vers $\sum_{j=1}^{d_0} \pi_j p(j)$, qui ne dépend pas de $i \in S_d$. De plus, pour tout $d \in \mathbb{N}$, $\mathcal{L}_d < \mathcal{L}_{d_0}$ dès que $(d, d_0) < d_0$, alors que $\mathcal{L}_d = \mathcal{L}_{d_0}$ lorsque $(d, d_0) = d_0$.

Pour estimer d_0, Rémillard et Vaillancourt (2019) proposent donc la méthode suivante: pour $d = 1, \ldots,$ on estime $p^{(d)}$ et $\mathcal{L}_d$ selon (2.2)–(2.3) sous l'hypothèse du modèle (2.1). $\hat{d}_0$ est alors le premier maximum local de $\mathcal{L}_d$ et on prend $\hat{p} = p_n^{(\hat{d}_0)}$.

3 Comportement en milieux non périodiques

Que se passe-t-il si le milieu n'est pas périodique? Soit $X = (X_t)_{t\geq 0}$ une MAMA satisfaisant à (1.1). L'opérateur de translation sur $(0,1)^{\mathbb{Z}}$ est noté T et on définit les écritures suivantes: $S(e) = 1 + \sum_{j=1}^{\infty} \prod_{k=1}^{j} \rho_k(e)$ et $F(e) = 1 + \sum_{j=1}^{\infty} \prod_{k=1}^{j} \frac{1}{\rho_{-k}(e)}$. La suite $\alpha(i,\cdot)$ est dorénavant présumée stationnaire et ergodique par rapport à une mesure μ que T préserve. Le comportement asymptotique de X est dans ce cas complètement déterminé par l'espérance de $\log \rho_0$, un important résultat dû à Alili (1999). Celui de l'estimateur $p_n^{(d)}(j)$ ne l'est que partiellement. En voici un cas, où la MAMA est balistique vers la droite, i.e., on a μ presque sûrement et $\forall i \in \mathbb{Z}$, $\mathbb{P}_i^e(\lim_{n\to\infty} X_n = +\infty) = 1$. Il est extrait de Rémillard et Vaillancourt (2019). On en présentera d'autres, ainsi que des illustrations de calculs et des conseils d'application, mais aussi quelques problèmes non résolus.

Proposition 3.1 *Supposons ergodiques toutes les puissances de T. Si en plus $\mathbb{E}(S) < \infty$, il vient $p_n^{(d)}(j) \overset{p.s.}{\to} v\mathbb{E}(S)\ \forall j \in \mathbb{Z}$ et $L_{n,d}/n \overset{p.s.}{\to} -H\{v\mathbb{E}(S)\}$, avec $v = \frac{1}{2\mathbb{E}(S)-1}$.*

Bibliographie

Adelman, O. et Enriquez, N. (2004). Random walks in random environment: What a single trajectory tells. *Israel Journal of Mathematics*, 142(1), 205–220.

Alili, S. (1999). Asymptotic behaviour for random walks in random environments. *J. Appl. Probab.*, 36(2), 334–349.

Andreoletti, P., Loukianova, D. et Matias, C. (2015). Hidden Markov model for parameter estimation of a random walk in a Markov environment. *ESAIM Probab. Stat.*, 19, 605–625.

Comets, F., Falconnet, M., Loukianov, O. et Loukianova, D. (2016). Maximum likelihood estimator consistency for recurrent random walk in a parametric random environment with finite support. *Stochastic Process. Appl.*, 126(11), 3578–3604.

Comets, F., Falconnet, M., Loukianov, O., Loukianova, D. et Matias, C. (2014). Maximum likelihood estimator consistency for a ballistic random walk in a parametric random environment. *Stochastic Process. Appl.*, 124(1), 268–288.

Diel, R et Lerasle, D. (2018). Non parametric estimation for random walks in random environment. *Stochastic Process. Appl.*, 128(1), 132–155.

Falconnet, M., Gloter, A. et Loukianova, D. (2014). Maximum likelihood estimation in the context of a sub-ballistic random walk in a parametric random environment. *Math. Methods Statist.*, 23(3), 159–175.

Pyke, R. (2003). On random walks and diffusions related to Parrondo's games. *IMS Lecture Notes Monogr. Ser.* 42, 185–216.

Rémillard, B et Vaillancourt, J. (2019). Detecting periodicity from the trajectory of a random walk in random environment. *Statist. Probab. Letters*, 155, 108568.

Comportement asymptotique de tests de Sobolev sur la sphère unité.

Christine Cutting, Davy Paindaveine and Thomas Verdebout

ECARES and Mathematics Department, Université Libre de Bruxelles, Boulevard du Triomphe, CP210, B-1050 Brussels, Belgium, email: tverdebo@ulb.ac.be.

Résumé. Dans ce travail, nous considérons le problème de test d'uniformité sur l'hypersphère unité de $\mathbb{R}^p$. Nous obtenons de nouveaux résultats sur le comportement asymptotique de tests de Sobolev sous des contre-hypothèses locales à symmétrie rotationnelle.

Mots-clés. Données directionnelles, tests d'uniformité, tests de Sobolev.

Abstract. In this work, we tackle the problem of testing uniformity on the unit hypersphere of $\mathbb{R}^p$. We obtain new results on the asymptotic behavior of Sobolev tests under rotationally symmetric alternatives.

Keywords. Directional data, uniformity tests, Sobolev tests.

1 Directional data and testing for uniformity

Directional statistics are dealing with observations that belong to the unit hypersphere $\mathbb{S}^{p-1} := \{\mathbf{u} \in \mathbb{R}^p : \|\mathbf{u}\|^2 = \mathbf{u}'\mathbf{u} = 1\}$ of $\mathbb{R}^p$ or more generally on compact Riemannian manifolds. Instances of directional data happen in meteorology (wind directions), astronomy (directions of cosmic rays, positions of stars), paleomagnetism (remanence directions), biology (protein structure, studies of animal navigation), forest sciences (directions of wildfire propagation), medicine (head normal vectors), and text mining (quantitative representation of documents in high-dimensional hyperspheres), to cite but some. Classical monographs on directional statistics are Watson (1983) and Mardia and Jupp (2000); a recent book that overviews the usage of some modern methods in directional statistics is Ley and Verdebout (2017).

When modeling directional data, that is, unit-norm multivariate vectors, a first natural question is to ask whether the directions at hand are uniformly distributed or, on the contrary, whether there exist modes of variation significantly different from uniformity. On the basis of n i.i.d. observations $\mathbf{U}_1, \ldots, \mathbf{U}_n$ with common distribution P on $\mathbb{S}^{p-1}$, the problem we tackle in this work is the problem of testing $\mathcal{H}_0 : \mathrm{P} \equiv \mathrm{Unif}(\mathbb{S}^{p-1})$ against $\mathcal{H}_1 : \mathrm{P} \neq \mathrm{Unif}(\mathbb{S}^{p-1})$, where $\mathrm{Unif}(\mathbb{S}^{p-1})$ stand for the uniform probability measure on $\mathbb{S}^{p-1}$. We study in this work tests belonging to the class of Sobolev tests for this problem. Sobolev tests are introduced in the next Section.

2 Sobolev tests

The class of so-called *Sobolev tests* has been introduced by Beran (1968, 1969) and Gine (1975). Sobolev tests are obtained using the eigenfunctions of the *Laplace–Beltrami operator* (or *Laplacian*) Δ acting on $\mathbb{S}^{p-1}$. Using the n-tuple of observations $\mathbf{U}_1, \ldots, \mathbf{U}_n$, a Sobolev test rejects the null hypothesis of uniformity $\mathcal{H}_0$ for large values of

$$S_n := \frac{1}{n} \sum_{i,j=1}^{n} \sum_{k=1}^{\infty} v_k^2 \langle \mathrm{t}_k(\mathbf{U}_i), \mathrm{t}_k(\mathbf{U}_j) \rangle, \tag{2.1}$$

where $\mathbf{u} \to \mathrm{t}_k(\mathbf{u})$ is a mapping from $\mathbb{S}^{p-1}$ to the space of eigenfunctions associated with the kth non-zero eigenvalue of the Laplacian, the v_k's are weights and $\langle f, g \rangle := \int_{\mathbb{S}^{p-1}} f(\mathbf{u}) g(\mathbf{u}) \, \mathrm{d}\mu(\mathbf{u})$ denotes the inner product on $L^2(\mathbb{S}^{p-1}, \mu)$ (μ is the surface area measure on $\mathcal{S}^{p-1}$). An explicit form for $\langle \mathrm{t}_k(\mathbf{U}_i), \mathrm{t}_k(\mathbf{U}_j) \rangle$ on $\mathbb{S}^{p-1}$ exists. More precisely, given $\mathbf{u}, \mathbf{v} \in \mathbb{S}^{p-1}$,

$$\langle \mathrm{t}_k(\mathbf{u}), \mathrm{t}_k(\mathbf{v}) \rangle = \begin{cases} 2\cos(k \angle(\mathbf{u}, \mathbf{v})), & \text{if } p = 2, \\ \left(1 + \frac{2k}{p-2}\right) C_k^{(p-2)/2}(\mathbf{u}'\mathbf{v}), & \text{if } p > 2, \end{cases} \tag{2.2}$$

where $\cos \angle(\mathbf{u}, \mathbf{v}) = \mathbf{u}'\mathbf{v}$ and C_k^{α} denote the Gegenbauer polynomial of index α and order k. Well-known Sobolev tests are

- the *Rayleigh test*. Taking $v_1 = 1$ and $v_k = 0$ for $k \geq 2$ in (2.1) we obtain the Rayleigh test statistic on $\mathbb{S}^{p-1}$ given by

$$R_n = \frac{p}{n} \sum_{i,j=1}^{n} \mathbf{U}_i' \mathbf{U}_j. \tag{2.3}$$

  Under $\mathcal{H}_0$, R_n is asymptotically χ_p^2 distributed.

- the *Bingham test*. When $\mathbf{U} \sim \mathrm{Unif}(\mathbb{S}^{p-1})$, then $\mathbb{E}[\mathbf{U}\mathbf{U}'] = \frac{1}{p}\mathbf{I}_p$. The Bingham test evaluates this latter sphericity property of $\mathbf{U}$ by the test statistic

$$B_n := \frac{np(p+2)}{2} \left(\mathrm{tr}(\mathbf{S}^2) - \frac{1}{p} \right),$$

  where $\mathbf{S} := \frac{1}{n} \sum_{i=1}^{n} \mathbf{U}_i \mathbf{U}_i'$ is the empirical covariance matrix of the $\mathbf{U}_i$'s. Under $\mathcal{H}_0$, B_n is asymptotically $\chi^2_{(p-1)(p+2)/2}$ distributed. The statistic B_n is obtained by letting $v_2 = 1$ and $v_k = 0$ for $k \neq 2$ in (2.1).

While much is known about the asymptotic behavior of several Sobolev tests under the null hypothesis of uniformity and when the dimension is fixed, less is known about the asymptotic behaviour of such tests under local alternatives, even under the rotationally symmetric alternatives defined in the next section.

3 Asymptotic results for the Rayleigh test

In this section, we present several results obtained in Cutting *et al.* (2017). We consider specific alternatives to the null of uniformity over the p-dimensional unit sphere $\mathcal{S}^{p-1}$, namely rotationally symmetric alternatives. A p-dimensional unit vector $\mathbf{U}$ is said to be *rotationally symmetric about* $\boldsymbol{\theta}(\in \mathcal{S}^{p-1})$ if and only if $\mathbf{OU}$ is equal in distribution to $\mathbf{U}$ for any orthogonal $p \times p$ matrix $\mathbf{O}$ satisfying $\mathbf{O}\boldsymbol{\theta} = \boldsymbol{\theta}$. We actually restrict to rotationally symmetric densities of the form

$$\mathbf{u} \mapsto c_{p,\kappa,f} f(\kappa\, \mathbf{u}'\boldsymbol{\theta}), \qquad \mathbf{x} \in \mathcal{S}^{p-1}, \tag{3.4}$$

where $\boldsymbol{\theta}(\in \mathcal{S}^{p-1})$ is a location parameter, $\kappa(> 0)$ is a concentration parameter, and the function $f : \mathbb{R} \to \mathbb{R}^+$ is monotone strictly increasing, twice differentiable at 0, and satisfies $f(0) = f'(0) = 1$. Consider triangular arrays of observations $\mathbf{U}_{ni}$, $i = 1, \ldots, n$, $n = 1, 2, \ldots$ where the random vectors $\mathbf{U}_{ni}$, $i = 1, \ldots, n$ take values in $\mathcal{S}^{p_n-1}$. More specifically, for any $\boldsymbol{\theta}_n \in \mathcal{S}^{p_n-1}$, $\kappa_n > 0$ and f as above, we will denote as $\mathrm{P}^{(n)}_{\boldsymbol{\theta}_n,\kappa_n,f}$ the hypothesis under which $\mathbf{U}_{ni}$, $i = 1, \ldots, n$ are mutually independent and share the common density $\mathbf{u} \mapsto c_{p_n,\kappa_n,f} f(\kappa_n\, \mathbf{u}'\boldsymbol{\theta}_n)$; $\mathrm{P}^{(n)}_0$ will denote triangular arrays of uniformly distributed observations. We have the following result.

Proposition 3.1 *Let (p_n) be a sequence in $\{2, 3, \ldots\}$. Let $(\boldsymbol{\theta}_n)$ be a sequence such that $\boldsymbol{\theta}_n \in \mathcal{S}^{p_n-1}$ for all n, (κ_n) be a positive sequence such that $\kappa_n = O(\sqrt{\frac{p_n}{n}})$. Then, the sequence of alternative hypotheses $\mathrm{P}^{(n)}_{\boldsymbol{\theta}_n,\kappa_n,f}$ and the null sequence $\mathrm{P}^{(n)}_0$ are mutually contiguous.*

The sequences κ_n of the form $\kappa_n = O(\sqrt{\frac{p_n}{n}})$ therefore characterize the high-dimensional contiguous alternatives to the uniform distribution. We also obtain in Cutting *et al.* (2017) the following result.

Proposition 3.2 *Let (p_n) be a sequence in $\{2, 3, \ldots\}$ and let $(\boldsymbol{\theta}_n)$ be a sequence such that $\boldsymbol{\theta}_n \in \mathcal{S}^{p_n-1}$ for all n. Let $\kappa_n = \tau_n \sqrt{p_n/n}$, where the positive sequence (τ_n) is $O(1)$ but not $o(1)$. Then, as $n \to \infty$ under $\mathrm{P}^{(n)}_0$,*

$$\log \frac{d\mathrm{P}^{(n)}_{\boldsymbol{\theta}_n,\kappa_n,f}}{d\mathrm{P}^{(n)}_0} = \tau_n \Delta^{(n)}_{\boldsymbol{\theta}_n} - \frac{\tau_n^2}{2} + o_{\mathrm{P}}(1), \tag{3.5}$$

where $\Delta^{(n)}_{\boldsymbol{\theta}_n} := \frac{\sqrt{p_n}}{\sqrt{n}} \sum_{i=1}^n \mathbf{U}'_{ni}\boldsymbol{\theta}_n$ is asymptotically standard normal. In other words, the model $\{\mathrm{P}^{(n)}_{\boldsymbol{\theta}_n,\kappa,f} : \kappa \geq 0\}$ is locally asymptotically normal at $\kappa = 0$ with central sequence $\Delta^{(n)}_{\boldsymbol{\theta}_n}$, Fisher information 1, and contiguity rate $\sqrt{p_n/n}$.

Proposition 3.2 entails that the test $\phi^{(n)}_{\boldsymbol{\theta}_n}$ rejecting the null at asymptotic level α whenever

$$\Delta^{(n)}_{\boldsymbol{\theta}_n} = \sqrt{np_n}\, \bar{\mathbf{X}}'_n \boldsymbol{\theta}_n > z_\alpha \tag{3.6}$$

is locally asymptotically most powerful for the considered problem. The Rayleigh test is not locally and asymptotically optimal for testing uniformity against specified-$\boldsymbol{\theta}_n$ rotationally symmetric alternatives. It is actually blind to the contiguous alternatives. It detects alternatives with $\kappa_n = O(\frac{p_n^{3/4}}{\sqrt{n}})$. We show in Cutting *et al.* (2017) that it is locally and asymptotically most powerful for the unspecified-$\boldsymbol{\theta}_n$ problem within the class of rotation-invariant tests.

4 Perspectives

High-dimensional results dealing with the asymptotic behavior of the Bingham test have been obtained in Cutting *et al.* (2020). Our objective in a near future is to extend the results obtained in Cutting *et al.* (2017) and Cutting *et al.* (2020) to the entire class of Sobolev tests.

Bibliographie

Beran, R. J. (1968). Testing for uniformity on a compact homogeneous space. *Journal of Applied Probability*, 5, pp. 177-195.

Beran, R. J. (1969). Asymptotic theory of a class of tests for uniformity of a circular distribution. *Annals of Mathematical Statistics*, 40, pp. 1196-1206.

Cutting, C., Paindaveine, D., and Verdebout, T. (2017). Testing uniformity on high-dimensional spheres against monotone rotationally symmetric alternatives. *Annals of Statistics*, 45(3), pp. 1024-1058.

Cutting, C., Paindaveine, D., and Verdebout, T. (2020). Testing uniformity on high-dimensional spheres: the non-null behaviour of the Bingham test. *Submitted*.

Ley, C. and Verdebout, T. (2017). Modern Directional Statistics. Chapman and Hall/CRC.

Mardia, K. V. and Jupp, P. E. (2000). *Directional Statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester.

Watson, G. S. (1983). *Statistics on Spheres*, volume 6 of University of Arkansas Lecture Notes in the Mathematical Sciences. John Wiley & Sons, New York.

Prédictions géostatistiques avec des données censurées : application à la caractérisation radiologique pour le démantèlement des installations nucléaires

Martin Wieskotten[1,3], Marielle Crozet[1], Bertrand Iooss[2], Céline Lacaux[3], Nadia Pérot[4]

[1] *CEA, DES, ISEC, DMRC, Univ. Montpellier, Marcoule, France et marielle.crozet@cea.fr*
[2] *EDF R&D, 6 quai Watier, 78400, Chatou, France et bertrand.iooss@edf.fr*
[3] *LMA Université d'Avignon, 84029, Avignon, France et celine.lacaux@univ-avignon.fr*
[4] *CEA, DES, IRESNE, DER, SESI, Cadarache, France et nadia.perot@cea.fr/martin.wieskotten@cea.fr*

Résumé. La caractérisation radiologique est un enjeu primordial dans les problématiques d'assainissement et de démantèlement d'infrastructures nucléaires. Les statistiques spatiales offrent des solutions pour estimer la présence de radionucléides, mais ne permettent pas toujours d'intégrer les données censurées aux modélisations. Ces données correspondent à des résultats de mesure en limite de détection et sont souvent supprimées ou remplacées par la valeur de la limite de détection. Ces pratiques introduisent un biais dans les prédictions en modifiant la variance et la moyenne des estimations. Une alternative aux méthodes de remplacement est la méthode du package R CensSpatial qui permet la prise en compte de données censurées. Notre objectif est de comparer les méthodes courantes de remplacement de données censurées à la méthode CensSpatial développée par Ordoñez et al. (2018) sur des mesures provenant d'un projet du CEA de Marcoule.

Mots-clés. Statistique Spatiale, Géostatistique, Données Censurées, Limite de détection, Algorithme SAEM, Package CensSpatial

Abstract. The radiological characterization is one of the main stake of nuclear facilities' decommissioning and dismantling projects. Spatial statistics offer solutions for predicting the location of radionuclides, but can not always take into account censored data. These responses correspond to measurement inferior to the decision threshold and are often discarded or replaced with the value of the detection limit. These practices insert bias in predictions for they change the variance and the mean of estimations. An alternative to these practices is the R package CensSpatial's method that allows to take censored data into account. Our goal is to compare the usual practices of replacement to the CensSpatial methods implemented by Ordoñez et al. (2018) on a data set from a project of the CEA Marcoule.

Keywords. Spatial Statistics, Geostatistics, Censored Data, Detection limit, SAEM Algorithm, CensSpatial Package

1 Introduction

La caractérisation radiologique d'un sol ou d'une surface est un enjeu primordial dans les projets d'assainissement et de démantèlement des installations nucléaires, en vue d'organiser la gestion des déchets produits. Cette cartographie radiologique est fondée sur l'observation d'une ou plusieurs grandeurs physiques en certains points, ce qui la rend initialement incomplète. Les statistiques spatiales offrent des solutions pour compléter cette cartographie initiale, notamment avec la géostatistique qui utilise les corrélations spatiales entre observations pour réaliser des prédictions sur des sites non observés. Néanmoins la géostatistique ne permet pas de prendre en compte les données censurées, données qui sont très répandues en assainissement/démantèlement. Elles correspondent à des résultats de mesure inférieurs à un seuil de décision et qui sont en pratique renvoyés par les laboratoires comme inférieurs à une limite de détection, sans donner la valeur numérique qui a été mesurée lors du protocole expérimental. Les définitions du seuil de décision et de la limite de détection peuvent être trouvées dans le Vocabulaire Internationnal de Métrologie et davantage de détails sur ces définitions sont donnés chez Rivier et Crozet (2014). Ces données apportent des informations limitées mais non négligeables. Les pratiques actuelles en géostatistique consistent à remplacer simplement ces données par une valeur arbitraire, ce qui selon la valeur choisie introduit un biais affectant les prédictions finales. Le package R CensSpatial développé par Ordoñez et al. (2018) permet d'estimer les valeurs censurées et de limiter le biais lors de la réalisation de prédictions. Pour vérifier les avantages de cette méthode par rapport aux méthodes de remplacement, nous avons choisi de comparer la méthode implémentée par CensSpatial avec une méthode de remplacement par 0, et une autre de remplacement par la limite de détection.

Pour réaliser cette comparaison, nous avons choisi un jeu de données provenant d'un projet du CEA Marcoule correspondant à 70 mesures d'activité surfacique (en Bq/cm^2). Afin de se ramener à l'hypothèse de normalité, une transformation des données est réalisée par un logarithme translaté : $Y = \ln(1 + X)$. Pour faciliter les comparaisons, des points supplémentaires ont été obtenus à l'aide d'un krigeage simple. Un maillage de 100×100 points est créé et augmente le nombre de points sur lesquels nous pouvons travailler. Dans la suite ces points supplémentaires sont considérés comme des observations, et seront donc considérés comme des données valides. De plus on fait l'hypothèse que la limite de détection est la même pour chaque observation, ce qui implique un protocole exactement identique lors de la mesure d'activité surfacique en chaque point. Cette hypothèse est raisonnable puisque ce jeu de données correspond à un échantillonnage in-situ effectué avec le même appareil de mesure.

Considérons un processus gaussien réel $Z(x), x \in D$, avec D la région étudiée ($D \subset \mathbb{R}^2$), stationnaire à l'ordre 2. La moyenne, notée μ, est donc indépendante de la position x et la covariance ne dépend que du vecteur $x - x'$. On observe les réalisations de ce processus $\mathbf{Z} = (Z(x_1), Z(x_2), ..., Z(x_n))$ en des points connus $x_i, i = 1, ..., n$, correspondant ainsi à n variables aléatoires. L'expression de la matrice de covariance de ces variables aléatoires

$Z(x_i)$ peut s'écrire $\mathbf{\Sigma} = \tau^2\mathbf{I}_n + \sigma^2\mathbf{R}(\phi)$ d'après Diggle et Ribeiro (2007), τ^2 correspondant au terme de pépite, σ^2 à la variance, ϕ à la portée et le terme $\mathbf{R}$ à une fonction vérifiant plusieurs conditions détaillées par Chilès et Delfiner (1999). Ici nous avons choisi une fonction exponentielle isotrope en accord avec le modèle utilisé pour construire les points supplémentaires par krigeage simple :

$$\mathbf{R}(\phi) = [R(\phi, \|x_i - x_j\|)] = \left[\exp\left(-\frac{\|x_i - x_j\|}{\phi}\right)\right],\ i = 1, ..., n \text{ et } j = 1, ..., n.$$

Par souci de simplicité, nous considérons dans la suite que $\tau^2 = 0$. Cette hypothèse a pour but de simplifier la comparaison des méthodes. De plus elle est en accord avec la construction des points supplémentaires ce qui la rend donc raisonnable, mais elle n'est pas vraie en pratique. Nous reviendrons sur ce point dans la conclusion.
L'intégration des données censurées se fait en réorganisant les expressions du vecteur des données et de la covariance en ordonnant les données observées (d'indice o) et les données censurées (d'indice c) :

$$\mathbf{Z} = \begin{bmatrix}\mathbf{Z}^o \\ \mathbf{Z}^c\end{bmatrix}, \mathbf{\Sigma} = \begin{bmatrix}\mathbf{\Sigma}^{oo} & \mathbf{\Sigma}^{oc} \\ \mathbf{\Sigma}^{co} & \mathbf{\Sigma}^{cc}\end{bmatrix}$$

Le vecteur $\mathbf{Z}$ correspond aux données observées complétées par les données censurées. Cette mise en forme sera utile lors du calcul de l'estimation du maximum de vraisemblance de la méthode CensSpatial. On note $\theta = (\mu, \sigma^2, \phi)$ l'ensemble des paramètres à estimer. L'expression de la fonction de vraisemblance l_c utilisée dans la méthode CensSpatial est la suivante :

$$l_c(\theta) \propto -\frac{1}{2}\left[\log(|\mathbf{\Sigma}|) + (\mathbf{Z} - \boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{Z} - \boldsymbol{\mu})\right] - \frac{n}{2}\log(2\pi)$$

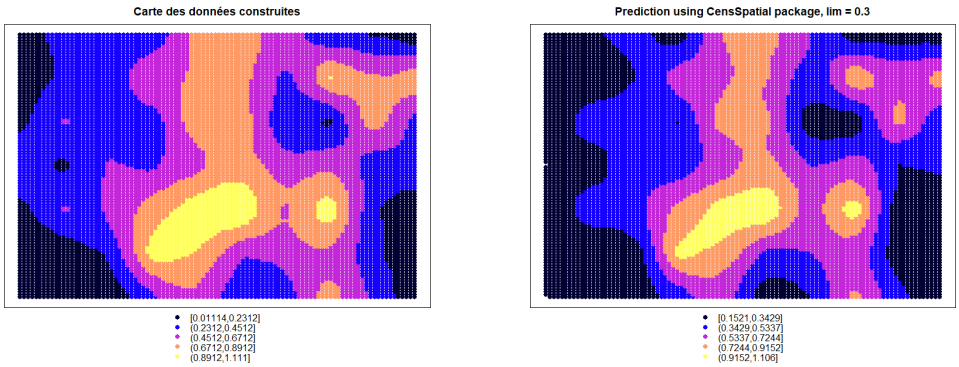
avec $\boldsymbol{\mu}$ le vecteur colonne contenant n répétitions de μ.

L'optimisation de cette fonction de vraisemblance se fait à l'aide de l'algorithme SAEM détaillé par Delyon et al. (1999). Cet algorithme est une variante de l'algorithme EM : "Expectation Maximisation" proposé par Dempster et al. (1977), qui est un algorithme itératif dont l'objectif est l'estimation des paramètres d'une loi de probabilité par maximum de vraisemblance. Sa particularité est sa capacité à prendre en compte une variable latente, dans notre cas les données censurées, pour réaliser cette estimation. Ces algorithmes se déroulent en 2 temps : la première étape estime l'espérance de la fonction de vraisemblance conditionnellement aux données (étape E) et la seconde estime les paramètres permettant de maximiser cette vraisemblance (étape M). L'algorithme SAEM rajoute un calcul d'approximation stochastique lors de l'étape E. Ordoñez et al. (2018) détaillent l'algorithme SAEM de la façon suivante :
Etape E-1 : On effectue un tirage aléatoire d'un vecteur $\mathbf{Z}^c$ représentant les données censurées d'une loi normale tronquée sur l'intervalle $[0; V_{lim}]$, V_{lim} étant la limite de détection. On forme ainsi le vecteur $\mathbf{Z}^k = \begin{bmatrix}\mathbf{Z}^o \\ \mathbf{Z}^c\end{bmatrix}$ qui contient le tirage des nouvelles données (censurées) et les données observées. On répète ce tirage un nombre M de fois pour obtenir

une séquence de vecteurs aléatoires : $(\mathbf{Z}^{(k,l)})_{l=1,\ldots,M}$.
Etape E-2 : On estime la valeur de l'espérance de la fonction de vraisemblance conditionnellement aux paramètres estimés à l'étape précédente et aux données complètes à l'aide d'une approximation stochastique comme décrit par Ordoñez et al. (2018).
Etape M : On maximise la log-vraisemblance et on obtient une nouvelle estimation des paramètres $\theta^{(k+1)}$ contenant la moyenne et la variance du processus ainsi que les paramètres de la covariance. Ces étapes sont répétées jusqu'à ce que la différence en valeur absolue des fonctions de vraisemblance aux étapes (k) et $(k+1)$ soit inférieure à un seuil donné. La convergence de l'algorithme est prouvée sous des hypothèses générales par Wu (1983).
La prédiction de la variable régionalisée en des points non observés est ensuite faite par krigeage simple.

2 Protocole de l'étude réalisée

La Figure 1 (a) correspond à la carte de la contamination (après transformation par le logarithme translaté, sans unité) obtenue avec les données construites initialement. Cette carte correspond à la « réalité » par rapport à laquelle nous comparons les 3 méthodes. La Figure 1 (b) donne un exemple de carte de prédiction de la contamination à l'aide de la méthode CensSpatial pour une limite de détection fixée à 0.3.
Les calculs réalisés se font en 2 temps : estimation des paramètres de la covariance (et



(a) (b)

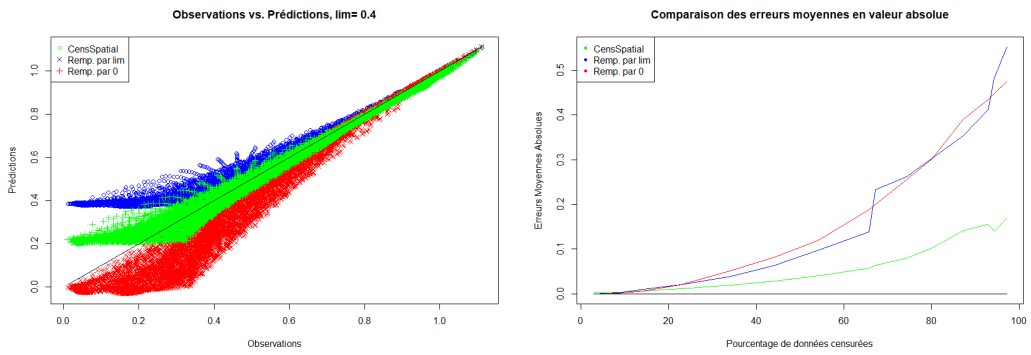
FIGURE 1 – Cartes de la contamination. (a) Carte initiale des données construites. (b) Carte des prédictions par CensSpatial.

des paramètres de la distribution pour la méthode CensSpatial), puis prédictions selon un maillage prédéfini par krigeage simple.
La moyenne des erreurs $m_{erreurs}$ est calculée sur plusieurs cas avec des limites de détection différentes correspondant à des pourcentages de données censurées différents, et ce pour

chaque méthode. Cette moyenne est calculée en faisant la différence entre la prédiction de la méthode z_{pred} et la valeur vraie z_{vrai} en valeur absolue et ce en chaque point de notre maillage de n_{pred} points : $m_{erreurs} = \sum \frac{|z_{pred} - z_{obs}|}{n_{pred}}$.

3 Résultats

La Figure 2 (a) représente les prédictions en fonction des observations pour une limite de détection égale à 0.4 (par exemple), avec la droite identité correspondant à une prédiction parfaite. Cette figure permet donc d'évaluer les tendances qui peuvent apparaître dans les prédictions par les 3 méthodes étudiées. Les différentes erreurs moyennes obtenues ont été représentées sur la Figure 2 (b). Cette figure permet d'étudier l'évolution de la moyenne des erreurs avec la variation du pourcentage de données censurées.



(a) (b)

FIGURE 2 – Résultats obtenus. (a) Graphe des prédictions en fonction des observations. (b) Moyenne des erreurs selon la limite de détection.

Pour comparer les différents résultats, nous avons isolé plusieurs éléments importants : les paramètres estimés de la covariance (variance et portée) et la qualité des prédictions faites, et ce selon le pourcentage de données censurées.
On observe dans le cas du remplacement par la limite une sur-estimation de la variance et dans le cas du remplacement par 0 une sous-évaluation de la variance de la variable régionalisée quand le pourcentage de données censurées augmente. La méthode CensSpatial estime une variance proche de celle réellement observée. De plus cette variance est plus stable lorsque le pourcentage de données censurées varie. La portée suit des évolutions similaires mais inversées : elle est sous-estimée par la méthode de remplacement par 0 et et sur-estimée par la méthode de remplacement par la limite.

Lorsque le nombre de données en limite de détection est faible, les 3 méthodes donnent des résultats quasi identiques. En effet la technique de prédiction (krigeage simple) étant

commune aux 3 méthodes et les jeux de données étant proches pour les 3 méthodes, les prédictions sont toutes quasi parfaites. En revanche lorsque la limite de détection augmente, les 3 méthodes offrent des résultats assez différents, la méthode de remplacement par 0 sous-estime les valeurs réelles et la méthode de remplacement par la limite surestime ces valeurs. La méthode CensSpatial offre un compromis aux 2 autres méthodes en générant des estimations comprises entre 0 et la limite de détection (conditionnellement aux données). La Figure 2(a) indique également que la méthode CensSpatial surestime les valeurs faibles, ce qui implique tout de même un gain en terme de sûreté dans le cadre de l'assainissement démantèlement (mais une perte en terme d'optimisation des coûts).

En conclusion, cette application de la méthode développée par Ordoñez et al. montre l'intérêt de la prise en compte des données censurées lors de réalisation de prédictions ou l'évaluation d'une structure de covariance. La méthode CensSpatial fournit des résultats similaires aux méthodes de remplacement pour un faible nombre de données censurées mais son intérêt augmente pour des pourcentages de données censurées supérieurs à 15%. Cette étude n'est cependant qu'une première étape car comme nous l'avons déjà évoqué, l'hypothèse d'incertitude de mesure nulle n'est pas raisonnable, en particulier avec un échantillonnage in-situ où le cadre expérimental est moins contrôlé qu'en laboratoire. La suite envisagée est d'étudier l'effet de ces incertitudes de mesure sur les modélisations spatiales et la mise en œuvre de méthodes comme CensSpatial. Une autre perspective de travail est la comparaison de la méthode CensSpatial avec une méthode remplaçant les données censurées par un tirage selon une loi normale tronquée à la limite de détection (et de paramètres identiques à ceux identifiés lors de l'inférence sur les données non censurées).

Bibliographie

Ordoñez J.A., Bandyopadhyay D., Lachos V.H., Cabral C.R.B. (2018), Geostatistical estimation and prediction for censored responses, *Spatial Statistics*, 23, 109-123.

Rivier C., Crozet M. (2014) Limite de détection de méthodes d'analyse et termes apparentés, *Techniques de l'ingénieur*, p262

Diggle P.J., Ribeiro P.J. (2007). Model-based Geostatistics, *Springer Series in Statistics*, Springer.

Chilès J.P., Delfiner P. (1999). Geostatistics : Modeling Spatial Uncertainty, *Wiley Series in Probability and Statistics*, Wiley.

Delyon B., Lavielle M., Moulines E., (1999). Convergence of a stochastic approximation version of the EM algorithm, Ann. Statist. 27(1), 94-128.

Dempster A., Laird N., Rubin D. (1977), Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society*, 39, 1-38.

Wu C.J. (1983) On the convergence properties of the EM algorithm, Ann. Statist., 11(1)

Ordoñez A., Galarza C.E., Lachos V.H. (2017), CensSpatial : Censored Spatial Models. R package version 2.1 URL https ://CRAN.R-project.org/package=CensSpatial

Critères de comparaison des algorithmes de génération de population synthétique à deux niveaux: application aux données françaises du recensement

Boyam Fabrice Yaméogo [1, 2, 3] & Pierre-Olivier Vandanjon [3]&
Pascal Gastineau [3]& Pierre Hankach [4]

[1] *Agence de l'Environnement et de la Maîtrise de l'Energie*
[2] *SNCF TER Mobilité Pays de la Loire.*
[3] *AME-EASE, Univ Gustave Eiffel, IFSTTAR,*
F-44344 Bouguenais, France
boyam-fabrice.yameogo@univ-eiffel.fr
pierre-olivier.vandanjon@univ-eiffel.fr
pascal.gastineau@univ-eiffel.fr
[4] *MAST-LAMES, Univ Gustave Eiffel, IFSTTAR,*
F-44344 Bouguenais, France
pierre.hankach@univ-eiffel.fr

Résumé.

La modélisation multi-agents, qui tend à se développer notamment dans le champ des transports, nécessite des données complètes sur les caractéristiques démographiques et socio-économiques des individus et des ménages. Toutefois, pour des raisons de confidentialité et de respect de la vie privée, il n'existe pas de base de données exhaustives renseignant les caractéristiques socio-démographiques des individus notamment à une échelle géographique fine. L'unique alternative est donc de générer une "population synthétique" d'individus (rattachés à des ménages), que l'on souhaite la plus représentative possible de la population réelle. Pour simuler cette population, les caractéristiques de l'ensemble des individus d'une zone géographique donnée sont généralement inférées à partir des caractéristiques des individus d'un échantillon de cette zone. Plusieurs méthodes sont disponibles pour générer des populations synthétiques cohérentes et précises.

Dans cette communication, un cadre méthodologique est proposé pour comparer onze méthodes de génération de population synthétique selon certains critères tels que le respect de la structure hiérarchique des données (association de variables individuelles et de ménages), les données d'entrée requises et le nombre d'attributs potentiels pouvant être générés. Ce cadre peut être utilisé comme référence pour le choix d'une méthode de synthèse en fonction de la quantité, du type et de la qualité des données disponibles. Sur la base de ce cadre analytique, quatre méthodes appropriées aux données françaises ont été mises en œuvre pour générer une population synthétique de l'aire urbaine de Nantes avec le logiciel libre R. Nos résultats montrent que les algorithmes testés produisent des populations synthétiques réalistes. Toutefois, le Hierarchical Iterative Proportionnal Fitting

(HIPF) génère la population synthétique qui représente le plus fidèlement la population réelle tout en préservant la cohérence entre les attributs des individus et les attributs des ménages auxquels ces individus sont rattachés.

Mots-clés. Population synthétique; Reconstruction synthétique; Microsimulation; Données françaises.

Abstract.

Agent-Based Models which have grown in popularity particularly in the field of transport, require comprehensive data on the demographic and socio-economic characteristics of individuals and households. However, for privacy reasons, there is no complete data about the socio-demographic characteristics of individuals at a small geographical scale. The only solution consists of generating a "synthetic population" of individuals (linked to households) that is representative of the real population. During this process, the characteristics of (all) the individuals in a given study area are usually inferred from the characteristics of the individuals in the sample from that area. Several well established methods exist to build consistent and accurate synthetic populations. In this communication, a methodological framework is proposed to compare eleven generation methods following certain criteria such as the respect of the hierarchical structure of the data (association of individual and household variables), the required input data and the number of potential attributes that can be handled. This framework can be used as a reference for choosing a synthesis method connected with the amount,the type and quality of available data. On the basis of this analytical framework, four appropriate methods to the French data were implemented to generate a synthetic population of the city of Nantes with the free R software. Our results show that the tested algorithms produce realistic synthetic populations. However, the Hierarchical Iterative Proportionnal Fitting (HIPF) generate the synthetic population that most faithfully represents the real population while maintaining consistency between the attributes of individuals and the attributes of the households to which these individuals are assigned.

Keywords. Population synthesis; Synthetic reconstruction; Microsimulation; French data.

1 Introduction

Les modèles multi-agents permettent de simuler les décisions et les activités des individus. Toutefois, pour un certain nombre de décisions (mobilité résidentielle, choix de mobilité,...), il est généralement admis que:

- les individus interagissent au sein des ménages auxquels ils appartiennent;
- les décisions prises ou les activités menées au sein d'un ménage influencent également le comportement des individus.

Dans ces cas, les décisions et activités d'un agent dépendent à la fois de ses caractéristiques propres mais aussi de celles du ménage auquel il appartient. La plupart du temps (notamment dans le domaine du transport), les modèles multi-agents ne considèrent qu'une seule de ces deux dimensions. Notre communication s'intéressera à la possibilité de considérér conjointement la dimension "individuelle" et la dimension "ménage". Les individus et les ménages peuvent alors être conceptualisés comme un système hiérarchique à deux niveaux distincts: un individu possède des caractéristiques propres (âge, statut professionnel...) et tous les individus d'un même ménage ont les mêmes attributs (possession d'une voiture, taille du ménage...).

Par conséquent, une description fine et exhaustive de la composition de la population prenant en compte les niveaux individuel et ménage peut s'avérer indispensable. La très grande majorité des instituts statistiques ne mettant à la disposition du public qu'un échantillon d'individus associés à leurs caractéristiques, la modélisation multi-agents nécessite le plus souvent, la génération d'une "population synthétique" représentative (en caractéristiques et en effectifs) des individus et des ménages réels. Le principe de base de la population synthétique est d'inférer les caractéristiques des individus d'une zone d'étude donnée à partir de distributions de probabilités appropriées, obtenues le plus souvent grâce aux caractéristiques des individus d'un échantillon de cette zone. La population synthétique apparaît alors comme une représentation microscopique simplifiée de la population actuelle car seules des variables d'intérêt sont sélectionnées pour l'inférence [Chapuis et Taillandier., 2019].

Cette communication propose une évaluation des principales méthodes permettant de générer des individus et des ménages synthétiques afin de déterminer laquelle est la plus appropriée aux données françaises à des fins de modélisation multi-agents des déplacements pour l'évaluation de politiques de transport. Les critères retenus sont entre autres le respect de la structure hiérarchique des données en associant au mieux les variables individuelles et ménages, la reproduction des interdépendances parmi les individus d'un même ménage, le respect de l'hétérogénéité de la distribution des ménages et des individus notamment au niveau des zones géographiques retenues [Sun et al., 2018]. Notre cas d'étude utilise les données du recensement français et s'applique à l'aire urbaine de Nantes.

2 Méthodologie

Trois grandes familles de génération de population synthétique sont mentionnées dans la littérature: Reconstruction Synthétique, Optimisation Combinatoire et Apprentissage Statistique [Sun et al., 2018]. Pour générer les individus et les ménages d'une zone donnée:

- l'approche de la Reconstruction Synthétique (SR) utilise à la fois l'échantillon et les données agrégées de cette zone afin de calculer des poids qui reflètent la représentativité de chaque ménage et individu de l'échantillon dans cette zone. Les méthodes SR nécessitent un effort de prétraitement et restent strictes en termes de données (échantillon et données agrégées à la fois au niveau individuel et au niveau ménage);
- les méthodes d'Optimisation Combinatoire (CO) utilisent également l'échantillon et les données agrégées afin de sélectionner une combinaison appropriée de ménages et d'individus qui correspondent le mieux aux données agrégées. Ces méthodes sont moins contraignantes en termes de données que les méthodes SR. Toutefois, elles ne peuvent pas toujours garantir la possibilité d'une solution optimale en ce qui concerne la correspondance avec les données agrégées. Elles peuvent également nécessiter un temps de calcul conséquent. Cette catégorie est mieux adaptée à la génération de population d'une taille réduite;
- la famille de l'Apprentissage Statistique (SL) considère uniquement l'échantillon et cherche à estimer la distribution jointe de tous les attributs en estimant une probabilité pour les différentes combinaisons possibles. La famille SL est capable de produire des résultats cohérents même pour des échantillons de petite taille (2,5% voire même 1%). Par contre, les méthodes SL ne permettent pas une correspondance avec les données agrégées, ce qui constitue un inconvénient majeur lorsque ces données sont disponibles.

Parmi les principales méthodes associées à ces trois grandes familles, onze méthodes ont été analysées au regard des critères suivants:

- principe de fonctionnement de l'algorithme de traitement (complexité, temps de calcul, etc);
- le respect de la structure hiérarchique des données (individus et ménages);
- les caractéristiques de la population synthétique obtenue;
- les données d'entrée requises;
- l'adéquation aux données agrégées disponibles à un niveau géographique fin;
- le niveau géographique;

- le nombre de variables qui peuvent être incluses dans le processus de génération.

Au regard de ces critères, les méthodes de Reconstruction Synthétique semblent plus adaptées pour la génération d'une population synthétique d'individus et de ménages à partir de données françaises du recensement. Quatre algorithmes de cette famille ont alors été testés: l'Iterative Proportional Updating (IPU) [Ye et al., 2009], le Hierarchical Iterative Proportional Fitting (HIPF) [Müller, 2017], la maximisation de l'entropie [Bar-Gera et al., 2009 ; Fu et al., 2011] et l'estimation par calages sur marges (Generalized Raking) [Deville et al., 1993]. Les performances de ces algorithmes sont comparées à partir des données du recensement de l'Institut National de la Statistique et des Etudes Economiques (INSEE) pour l'aire urbaine de Nantes. Une population synthétique d'individus et de ménages (949.000 individus rattachés à 418.000 ménages) a été générée pour chaque iris (307 iris au total) de l'aire urbaine avec chacun des algorithmes à partir d'un échantillon de 287.000 individus et 136.000 individus.

Les attributs suivants ont été utilisés dans la génération de la population synthétique:

- caractéristiques individuelles: sexe, âge, catégorie socio-professionnelle, condition d'emploi, temps de travail, personne de référence du ménage, diplôme le plus élevé;
- caractéristiques des ménages: structure familiale, catégorie socio-professionnelle de la personne de référence du ménage, nombre de voitures, type de ménage, taille du ménage.

3 Principaux résultats obtenus

La performance de ces quatre algorithmes de synthèse a été évaluée en utilisant les indicateurs statistiques fréquemment utilisés dans la littérature sur la microsimulation spatiale:

- Le coefficient de détermination R^2. Cet indicateur varie entre 0 et 1 et mesure la corrélation linéaire entre les données simulées x et les données réelles y (données agrégées du recensement) au sein de chaque iris i.

$$R^2 = \left(\frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x} \cdot \bar{y}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2} \quad \sqrt{\frac{1}{n}\sum_{i=1}^{n} y_i^2 - \bar{y}^2}} \right)^2$$

  Plus le R^2 est proche de 1, meilleure est la correspondance entre les données simulées et les données réelles.

- L'erreur absolue standardisée (SAE). Cet indicateur donne le pourcentage d'erreurs entre les données simulées et les données réelles.

$$\text{SAE}= \frac{\sum_{ij} |x_{ij} - y_{ij}|}{\text{total population} \times \text{nombre de variables}}$$

avec x et y qui représentent respectivement les données simulées et observées pour chaque attribut j au sein de l'iris i .

- L'erreur quadratique moyenne standardisée (SRMSE).

$$\text{SRMSE}= \frac{\sqrt{\frac{1}{J}\sum_{j=1}^{J}(x_i - y_i)^2}}{\frac{1}{J}\sum_{i} y_i}$$

avec J qui représente le nombre total d'attributs (toutes variables confondues). Cet indicateur analyse la dispersion des erreurs et est utilisé pour évaluer la qualité de l'ajustement entre la population synthétique estimée et les données agrégées. Plus sa valeur est faible, plus la qualité de l'ajustement est meilleure.

En plus de ces indicateurs, les algorithmes ont également été évalués en fonction de leur capacité à:

- conserver la structure hiérarchique entre individus et ménages;
- maintenir l'hétérogénéité des ménages et des individus entre les iris;
- reproduire les interdépendances entre les agents d'un même ménage;

Au regard des premiers résultats obtenus, l'HIPF semble produire la population synthétique représentant le plus fidèlement la population réelle tout en préservant la cohérence entre les attributs des individus et les caractéristiques des ménages auxquels ils appartiennent.

Bibliographie

Bar-Gera, H., Konduri, K., Sana, B., Ye, X., and Pendyala, R. M. (2009). Estimating survey weights with multiple constraints using entropy optimization methods, *88th Annual Meeting of the Transportation Research Board*, Washington, DC.
Chapuis, K. and Taillandier, P. (2019). A brief review of synthetic population generation practices in agent-based social simulation, *SSC2019, Social Simulation Conference*, 27-29 septembre 2019, Mainz, Germany.

Deville, J.-C., Särndal, C.-E., and Sautory, O. (1993). Generalized raking procedures in survey sampling, *Journal of the American statistical Association*, 88, pp.1013–1020.
Lee, D.-H. and Fu, Y. (2011). Cross-entropy optimization model for population synthesis in activity-based microsimulation models, *Transportation Research Record*, 2255,1, pp. 20-27.
Ma, L. and Srinivasan, S. (2015) Synthetic population generation with multilevel controls: A fitness-based synthesis approach and validations, *Computer-Aided Civil and Infrastructure Engineering*, 30, 2, pp. 135-150.
Müller, K. (2017). A generalized approach to population synthesis, PhD thesis, ETH Zurich.
Müller, K. and Axhausen, K. W. (2012) Multi-level fitting algorithms for population synthesis, *Arbeitsberichte Verkehrs-und Raumplanung*, 821, IVT, ETH Zürich.
Saadi, I., Mustafa, A., Teller, J., Farooq, B., and Cools, M. (2016) Hidden Markov Model, *Transportation Research Part B: Methodological*, 90, pp.1-21.
Sun L., Erath A., Cai M. (2018). A hierarchical mixture modeling framework for population synthesis, *Transportation Research Part B: Methodological*, 114, pp. 199-212.
Ye, X., Konduri, K., Pendyala, R. M., Sana, B., and Waddell, P. (2009). A methodology to match distributions of both household and person attributes in the generation of synthetic populations, *88th Annual Meeting of the Transportation Research Board*, Washington, DC.

RÉGRESSION AVEC OPTION REJET

Ahmed ZAOUI & Christophe DENIS & Mohamed HEBIRI

Laboratoire LAMA, Université Gustave Eiffel,
Ahmed.Zaoui/Christophe.Denis/Mohamed.Hebiri@univ-eiffel.fr

Résumé. Nous étudions le problème de la régression où l'on est autorisé à s'abstenir de prédire. Ce cadre d'étude vise à prévenir les conséquences d'une prise de décision erronée. Nous appelons ce nouveau cadre la *régression avec option rejet*. Nous considérons le cas où le taux de rejet est fixe et établissons la *règle avec option rejet* optimale pour ce problème. Cette règle optimale conduit naturellement à considérer une procédure semi-supervisée d'estimation de type plug-in : un premier ensemble de données étiquetées est utilisé pour estimer les fonctions de régression et de variance conditionnelle ; un deuxième ensemble de données *non étiquetées* est exploité pour calibrer le taux de rejet à celui souhaité. Il s'avère que, lorsque l'estimation de la fonction de variance conditionnelle est basée sur l'algorithme des K-plus proches voisins (K-PPV), le prédicteur avec option rejet résultant a des performances proches du prédicteur avec option rejet optimal, à la fois en termes de risque et de taux de rejet. Par ailleurs, la procédure proposée offre de bonnes performances numériques.

Mots-clés. K-PPV, Régression, Régression avec option rejet, Plug-in.

Abstract. We investigate the problem of regression where one is allowed to abstain from predicting. We refer to this new framework as the *regression with reject option* as an extension of the classification with reject option. In this context, we focus on the case where the rejection rate is fixed and derive the optimal rule which relies on thresholding the conditional variance. Based on the plug-in principle, we provide an estimation procedure of the optimal rule involving two datasets : a first *labeled* dataset is used to estimate both regression function and conditional variance by the k-NN algorithm ; a second *unlabeled* dataset is exploited to calibrate the desired rejection rate. The resulting k-NN predictor with reject option is shown to be almost as good as the optimal predictor with reject option both in terms of risk and rejection rate. Finally, a numerical study is performed to illustrate the benefit of using the proposed procedure.

Keywords. K-nn, Regression, Regression with reject option, Plug-in.

1 Introduction

Prédire avec confiance est un enjeu majeur en apprentissage statistique. Ainsi, de nombreuses procédures d'estimation proposées visent à réduire les erreurs de prédiction, *e.g.,* les réseaux de neurones, les méthodes à noyaux, l'approche des K plus proches voisins

(K-PPV), les moindres carrés régularisés pour n'en nommer qu'un petit nombre. Cependant, même les méthodes les plus performantes commettent des erreurs qui peuvent dans certains cas avoir de lourdes conséquences. Le domaine médical est un champs d'application notable dans lequel il est souvent souhaitable de s'abstenir plutôt que de prendre une mauvaise décision. C'est dans ce but que nous introduisons le problème de *régression avec option rejet* qui vise à construire des procédures d'estimation capables de ne pas prédire lorsque le doute dans la valeur prédite est trop grand. En particulier, nous nous intéressons au cadre de travail dans lequel le taux de rejet est fixe, laissant ainsi la possibilité à l'humain d'agir sur une proportion des données jugées, par l'algorithme, trop délicate à traiter par l'algorithme lui-même. Cette décision de rejeter et donc de ne pas traiter certaines données par la machine est prise par l'algorithme et vise à rentabiliser l'effort de l'humain.

La prédiction avec option rejet a déjà fait l'objet de nombreux travaux dans le cadre de la classification. On peut citer par exemple les références suivantes : Chow (1970), Vovk *et al.* (1999), Herbei et Wegkamp (2006), Naadeem *et al.* (2010), Lei (2014), Denis et Hebiri (2019). A notre connaissance, ce travail est le premier à étendre la notion de rejet au problème de régression. Il apporte entre autres une meilleure compréhension des situations où il est judicieux d'avoir recours à une méthode autorisant le rejet. En particulier, la fonction de variance conditionnelle est au centre du problème de rejet et pour cette raison, nous nous focalisons sur des problèmes où la variance n'est pas constante (sur l'espace des individus). Ainsi, en marge de l'étude du problème de rejet, nous proposons une estimation de la fonction de variance conditionnelle fondée sur l'algorithme des K-PPV et en étudions les propriétés statistiques.

2 Cadre de travail

Cette section a pour objet d'introduire les différentes notations et paramètres utiles à la définition des prédicteurs avec option rejet. Nous nous plaçons dans le cadre de la régression. On considère $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$ un couple admettant une loi de probabilité $\mathbb{P}$ sur $\mathbb{R}^d \times \mathbb{R}$. On définit la fonction de régression f^* et la fonction de variance σ^2 respectivement comme suit : $f^*(x) = \mathbb{E}[Y|X = x]$ et $\sigma^2(x) = \mathbb{E}[(Y - f^*(X)^2|X = x]$ pour tout $x \in \mathbb{R}^d$. L'exemple le plus répandu de ce cadre d'étude est la régression hétéroscédastique où l'on suppose que $Y = f^*(X) + \sigma(X)\,\xi$ pour une variable aléatoire ξ, dite de bruit, centrée et indépendante de X.

Un prédicteur avec option rejet est une fonction mesurable de $\mathbb{R}^d$ dans $\mathcal{P}(\mathbb{R})$, l'ensemble des parties de $\mathbb{R}$. Dans notre méthodologie, un prédicteur est toujours associé à une fonction mesurable de $\mathbb{R}^d$ dans $\mathbb{R}$. Pour $f : \mathbb{R}^d \to \mathbb{R}$ une telle fonction, un *prédicteur avec option rejet* $\Gamma^{(f)} : \mathbb{R}^d \to \mathcal{P}(\mathbb{R})$, est une application qui à tout $x \in \mathbb{R}^d$ associe $\Gamma^{(f)}(x) \in \{\emptyset, \{f(x)\}\}$. L'utilisation d'un prédicteur avec option rejet offre donc deux possibilités : soit $\Gamma^{(f)}(x) = \{f(x)\}$ et on prédit $f(x)$ pour la variable x ; soit $\Gamma^{(f)}(x) = \emptyset$ et

l'option rejet a été utilisée. Notons que dans ce dernier cas, le cardinal de $\Gamma^{(f)}(x)$, noté $|\Gamma^{(f)}(x)|$, est nul (et $|\Gamma^{(f)}(x)| = 1$ dans le cas de la prédiction).

L'objet de l'étude est de savoir quand utiliser l'option rejet. Ainsi, pour un prédicteur avec option rejet $\Gamma^{(f)}$, nous introduisons un risque prenant en compte d'une part l'erreur de prédiction induite par la fonction f lorsque $\Gamma^{(f)}$ décide de prédire, et d'autre part un prix à payer pour l'utilisation de l'option rejet :

$$\mathcal{R}_\lambda\left(\Gamma^{(f)}\right) = \mathbb{E}\left[(Y - f(X))^2 \mathbb{1}_{\{|\Gamma^{(f)}(X))|=1\}}\right] + \lambda\,\mathbb{P}\left(|\Gamma^{(f)}(X)| = 0\right),$$

où $\mathbb{1}_{\{\cdot\}}$ est la fonction indicatrice et $\lambda > 0$ est un paramètre de régularisation contrôlant le poids que l'on souhaite donner à l'emploi de l'option rejet. En particulier, une forte valeur de λ impose une contrainte importante sur l'utilisation de l'option rejet, alors qu'à l'inverse, on aura facilement recours au rejet si λ est faible.

Alors que l'objet de ce manuscrit porte sur l'étude de prédicteur avec option rejet dont le taux de rejet est fixe (ce qui sera l'objet de la section suivante), nous établissons ici un premier résultat qui consiste à identifier la règle avec option rejet optimale pour λ fixe.

Théorème 1. *Soit $\lambda > 0$ et soit le* prédicteur avec option rejet optimal *définit par $\Gamma^*_\lambda \in \operatorname{argmin}_{\Gamma^{(f)}} \mathcal{R}_\lambda\left(\Gamma^{(f)}\right)$, où l'infimum est pris sur l'ensemble de tous les prédicteurs avec option rejet. On a alors*

$$\Gamma^*_\lambda(X) = \begin{cases} \{f^*(X)\} & si\ \ \sigma^2(X) \leq \lambda\ , \\ \emptyset & sinon\ . \end{cases} \tag{1}$$

Le résultat ci-dessus décrit un aspect important de Γ^*_λ et des prédicteurs avec option rejet en général. Le fait de rejeter ou non dépend essentiellement de la fonction de variance : plus le bruit en une région de l'espace est fort plus on a tendance à s'abstenir de prédire. Il est à noter que ce premier élément n'a pas été observé dans la littérature sur la classification avec option rejet. Deuxièmement, le Théorème 1 suggère qu'une bonne estimation de f^* et de σ^2 est requise pour bien estimer Γ^*_λ.

3 Prédicteurs avec taux de rejet fixe

Nous avons considéré dans la section précédente le cas où le rejet a un coût fixe λ. Toutefois, ce paramètre de seuillage de la fonction de variance (*cf.* Eq. (1)) est crucial et le déterminer peut s'avérer difficile. Pour cette raison, inspiré de la classification avec option rejet (Denis et Hebiri, 2019), nous présentons dans cette section le cadre d'étude où le prédicteur avec option rejet a un taux de rejet fixé à l'avance. Ceci traduit typiquement des situations où l'on dispose de ressources humaines limitées pour traiter toutes les tâches. Pour ce faire, on ajoute une contrainte sur le taux de rejet du prédicteur. Plus précisément,

soit $\varepsilon \in (0,1]$. Le *prédicteur optimal à taux de rejet* ε, ou le *ε-prédicteur optimal* pour raccourcir est défini par :

$$\Gamma_\varepsilon^* \in \operatorname{argmin} \left\{ \mathbb{E}\left[(Y - f(X))^2 \,|\, |\Gamma^{(f)}(X)| = 1\right] \;;\; r\left(\Gamma^{(f)}\right) \leq \varepsilon \right\} \;,$$

où pour tout f, $r\left(\Gamma^{(f)}\right) = \mathbb{P}\left(|\Gamma^{(f)}(X)| = 0\right)$. Contrairement au cas où λ est fixe, la caractérisation du ε-prédicteur optimal requiert une hypothèse supplémentaire :

Hypothèse 1. *La fonction de répartition F_{σ^2} de la variable aléatoire $\sigma^2(X)$ est continue.*

Cette hypothèse a été introduite par Denis et Hebiri (2019) dans le contexte de la classification binaire avec option rejet. Sous l'Hypothèse 1 et les propriétés de la fonction quantile [1] $F_{\sigma^2}^{-1}$, il est clair que $F_{\sigma^2}(\sigma^2(X)) \leq 1-\varepsilon \iff \sigma^2(X) \leq F_{\sigma^2}^{-1}(1-\varepsilon)$. Cette équivalence permet de déduire le résultat suivant :

Théorème 2. *Soit $\varepsilon \in (0,1]$. Sous l'Hypothèse 1, les assertions suivantes sont vérifiées :*
(i) le ε-prédicteur optimal *est défini par :*

$$\Gamma_\varepsilon^*(X) = \begin{cases} \{f^*(X)\} & \text{if } \sigma^2(X) \leq \lambda_\varepsilon \;, \\ \emptyset & \text{sinon} \;, \end{cases} \tag{2}$$

où $\lambda_\varepsilon = F_{\sigma^2}^{-1}(1-\varepsilon)$. De plus, $r\left(\Gamma_\varepsilon^*\right) = \mathbb{P}\left(|\Gamma_\varepsilon^*(X)| = 0\right) = \varepsilon$.
(ii) l'excès de risque $\mathcal{E}\left(\Gamma^{(f)}\right) := \mathcal{R}_{\lambda_\varepsilon}(\Gamma^{(f)}) - \mathcal{R}_{\lambda_\varepsilon}(\Gamma_\varepsilon^*)$ de tout prédicteur avec option rejet $\Gamma^{(f)}$ peut s'écrire explicitement sous la forme :

$$\mathcal{E}\left(\Gamma^{(f)}\right) = \mathbb{E}\left[(f^*(X) - f(X))^2 \mathbb{1}_{\left\{|\Gamma^{(f)}(X)|=1\right\}}\right] + \mathbb{E}\left[|\sigma^2(X) - \lambda_\varepsilon| \mathbb{1}_{\left\{|\Gamma^{(f)}(X)| \neq |\Gamma_\varepsilon^*(X)|\right\}}\right] \;.$$

Ce théorème établit un point crucial, à savoir la valeur de λ pour laquelle le prédicteur a exactement un taux de rejet égal à ε. De plus, le point (ii) du Théorème 2 donne une formulation explicite de l'excès de risque d'un prédicteur $\Gamma^{(f)}$. Il est montré que celui-ci se décompose en deux parties. La première dépend de l'erreur L^2 de la fonction f alors que la seconde, plus originale, dépend du comportement de la fonction de variance σ^2 autour du seuil λ_ε.

3.1 Stratégie d'estimation

L'objectif de cette section est de fournir un estimateur de Γ_ε^* ayant des propriétés statistiques proches. La nature de l'oracle donné par (2) suggère une approche semi-supervisée de type plug-in qui repose sur l'estimation de f^* et σ^2, ainsi que du seuil λ_ε. Dans ce but, nous utilisons deux échantillons indépendants : $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$ qui consiste en n copies indépendantes de (X,Y), et $\mathcal{D}_N = \{X_i\}_{i=n+1}^{n+N}$ qui contient N observations indépendantes,

1. ou encore inverse généralisé de F_{σ^2}

et *non étiquetées*, de même loi que X. L'échantillon $\mathcal{D}_n$ sera exploité pour construire des estimateurs $\hat{f}$ et $\hat{\sigma}^2$ de f^* et σ^2. Nous reportons cet aspect de l'étude à la Section 3.2 et nous concentrons pour l'heure sur l'utilisation de l'échantillon $\mathcal{D}_N$, en partant du principe que les estimateurs $\hat{f}$ et $\hat{\sigma}^2$ sont déjà construits. Considérons la fonction de répartition empirique de $\hat{\sigma}^2$ sur $\mathcal{D}_N$ donnée par $\hat{F}_{\hat{\sigma}^2}(\cdot) = \frac{1}{N}\sum_{i=n+1}^{n+N} \mathbb{1}_{\{\hat{\sigma}^2(X_i)\leq\cdot\}}$. Nous sommes à présent en mesure de définir le *prédicteur plug-in de niveau* ε :

$$\hat{\Gamma}_\varepsilon(x) = \begin{cases} \left\{\hat{f}(x)\right\} & \text{if } \hat{F}_{\hat{\sigma}^2}(\hat{\sigma}^2(x)) \leq 1-\varepsilon \ , \\ \emptyset & \text{sinon } . \end{cases} \tag{3}$$

3.2 Estimation de σ^2 par K-PPV

Le problème de l'estimation de la fonction de régression f^* est classique et largement étudié, voir par exemple Wand et Jones (1994), Györfi et al (2002), ou encore Biau et Devroye (2015). Dans la suite, nous supposons que l'on dispose d'un estimateur $\hat{f}$ de f^*, consistant en norme infinie. La question de l'estimation de σ^2 est plus originale et peu étudiée dans le cas de *design* aléatoire (*cf.* Hall et Caroll (1989), Fan et Yao (1998) ou encore Cai et Wang (2008)). Dans ce papier, nous considérons l'algorithme populaire des K-PPV pour estimer ce paramètre.

Soit $x \in \mathbb{R}^d$. Introduisons la notation suivante $\left\{\left(X_{(i,n)}(x), Y_{(i,n)}(x)\right)\right\}_{i=1}^n$, caractérisant la suite ordonnée des voisins de x. Ici $X_{(i,n)}(x)$ est vu comme le $i^{\text{ème}}$ plus proche voisin de x au sens de la distance Euclidienne [2]. On estime alors σ^2 par :

$$\hat{\sigma}^2(\cdot) := \frac{1}{k}\sum_{i=1}^{k}\left(Y_{(i,n)}(x) - \hat{f}\left(X_{(i,n)}(x)\right)\right)^2 \ . \tag{4}$$

La proposition suivante établit la consistance de l'estimateur (4) en norme L^1 :

Proposition 1. *Supposons que les fonction f^* et σ^2 soient bornées et que $\mathbb{E}[Y^4] < \infty$. Supposons de plus que $\mathbb{E}\left[\left\|\hat{f} - f^*\right\|_\infty^2\right] \underset{n\to+\infty}{\longrightarrow} 0$. Alors, pour $k = k_n$ tel que $k_n \to \infty$, et $k_n/n \to 0$ quand $n \to \infty$, on a*

$$\mathbb{E}\left[\left|\hat{\sigma}^2(X) - \sigma^2(X)\right|\right] \underset{n\to+\infty}{\longrightarrow} 0.$$

3.3 Consistance

On étudie dans cette section les propriétés de consistance du *prédicteur plug-in de niveau* ε défini par l'Eq. (3). Pour cela, nous imposons l'hypothèse suivante dont l'importance a été démontrée empiriquement par Denis et Hebiri (2019).

2. En cas d'égalité, on convient que le le voisin le plus proche est celui ayant l'indice le plus petit.

Hypothèse 2. *la fonction de répartition $F_{\hat{\sigma}^2}$ de $\hat{\sigma}^2(X)$ est continue.*

L'Hypothèse 2 n'est pas restrictive puisqu'elle contraint l'estimateur $\hat{\sigma}^2$ et que celui-ci peut être choisi par le praticien [3]. Nous pouvons à présent établir notre résultat principal :

Théorème 3. *Supposons que σ^2 est Lipschitz, $\hat{f}$ est un estimateur consistant de f^* en norme L^1 et $\hat{\sigma}^2$ est un estimateur consistant de σ^2 en norme L^2. Sous les Hypothèses 1-2, le prédicteur plug-in de niveau ε donné par* (3) *satisfait les propriétés suivantes :*

$$\mathbb{E}\left[\mathcal{E}\left(\hat{\Gamma}_\varepsilon\right)\right] \underset{n,N\to+\infty}{\longrightarrow} 0, \quad \textit{et} \quad \mathbb{E}\left[r(\hat{\Gamma}_\varepsilon)\right] \underset{n,N\to+\infty}{\longrightarrow} \varepsilon \ .$$

En résumé, sous de bonnes conditions, le Théorème 3 illustre le fait que le prédicteur $\hat{\Gamma}_\varepsilon$ est asymptotiquement aussi bon que le prédicteur optimal et de niveau ε.

Bibliographie

Biau, G. et Devroye, L. (2015). Lectures on the Nearest Neighbor Method, *Springer Series in the Data Sciences, Springer New York*.

Cai, T. et Wang, L. (2008). Adaptive variance function estimation in heteoscedastic nonparametric regression, *The annals of statistics*, 36(5), pp. 2025-2054.

Chow, C (1970). On optimum error and reject trade-off, *IEEE Transactions on Information Theory*, 16 :41–46.

Chzhen, E., Denis C. et Hebiri, M. (2019). Minimax semi-supervised confidence sets for multi-class classification. *preprint*.

Denis, C. et Hebiri, M. (2019). Consistency of plug-in confidence sets for classification in semi-supervised learning, *Journal of Nonparametric Statistics*.

Fan, J. et Yao, Q. (1998). Efficient estimation of conditional variance functions in stochastic regression, *Biometrika*, 85(3), pp. 645-660.

Györfi, L., Kohler, M., Krzyżak, A. et Walk, H. (2002). A distribution-free theory of nonparametric regression, *Springer Ser. Statist.*

Hall, P. et Carroll, R.J. (1989). Variance function estimation in regression : The mean effect of estimating the mean, *Journal of the Royal Statistical Society : Series B (Methodological)*, 51(1), pp. 3-14.

Herbei, R. et Wegkamp, M. (2006). Classification with reject option, *Canad. J. Statist.*, 34(4) :709–721.

Lei, J. (2014). Classification with confidence. *Biometrika*, 101(4) :755–769.

Naadeem, M., Zucker, J. et Hanczar, B. (2010). Accuracy-rejection curves (ARCs) forcomparing classification methods with a reject option. *InMLSB*, pages 65–81.

Vovk, V., Gammerman, A. et Saunders, C. (1999). Machine-learning applications of algorithmic randomness. *Proc. 16th International Conf. on Machine Learning*, pages 444–453.

Wand, M. P. et Jones M.C. (1994). Kernel smoothing, *CRC Press.*

3. Quitte à modifier légèrement l'estimateur (Chzhen *et al.* 2019).

Modèle de Markov multi-états pour estimer l'incidence du VIH à partir des données de notification en France: 2008-2018

Charlotte Castel [1,2] & Cécile Sommen [1] & Yann Le Strat [1] & Ahmadou Alioum [3,4]

[1] *Santé Publique France, Direction Appui, Traitements et Analyses des données, Saint-Maurice F-94417, France*
[2] *Université Paris-Est, Champs sur Marne F-77420, France*
[3] *Equipe Biostatistique, Centre Inserm U1219-Bordeaux Population Health, Université de Bordeaux, F-33000 Bordeaux, France*
[4] *Université de Bordeaux, ISPED, Centre Inserm U1219-Bordeaux Population Health, F-33000 Bordeaux, France*

Résumé.
Trente-cinq ans après la découverte du virus de l'immunodéficience humaine (VIH), l'épidémie se poursuit en France. Afin d'orienter les stratégies de prévention du VIH et de suivre leur impact, il est indispensable de connaître la dynamique de l'épidémie. L'indicateur permettant de rendre compte de l'évolution des nouvelles infections est l'incidence. En France, depuis 2003, Santé publique France a mis en place un système de surveillance des nouveaux diagnostics VIH couplé à une surveillance virologique, permettant notamment de disposer du stade clinique du patient au moment du diagnostic VIH ainsi que des valeurs de deux marqueurs d'infection au diagnostic. Nous avons adapté l'approche pour l'estimation de l'incidence du VIH proposée par Sommen et al. [2009] en utilisant une vraisemblance pénalisée pour obtenir une estimation lisse de la courbe d'incidence du VIH. La vraisemblance pénalisée a été calculée dans le cadre d'un modèle de Markov multi-états non homogène. La courbe d'incidence du VIH a été approximée à l'aide de M-splines cubiques et une approximation du critère de validation croisée a été utilisée pour estimer le paramètre de lissage. La méthode est illustrée sur 200 simulations reproduisant les données de la déclaration obligatoire du VIH collectées à Santé publique France et sur les données françaises de la déclaration obligatoire du VIH. Le modèle présenté ici est un nouvel outil pour estimer l'incidence de l'infection à VIH à partir des données de surveillance en utilisant l'information des stades cliniques au moment du diagnostic, et en tenant compte de l'évolution du dépistage au cours du temps.

Mots-clés. VIH, incidence, model de Markov multi-états, vraisemblance pénalisée, données de notification, surveillance

Abstract.
Thirty-five years after the discovery of the human immunodeficiency virus (HIV), the epidemic is still going in France. In order to guide HIV prevention strategies and monitor their impact, it is essential to know the dynamics of the epidemic. The indicator for reporting the progress of new infections is incidence. In France, since 2003, the National Institute for Public Health Surveillance has set up a surveillance system for new HIV diagnoses coupled with a virological surveillance, allowing to collect the clinical stage of the patient at the time of HIV diagnosis, as well as the value of two markers of infection at diagnosis.We adapted the approach for estimating HIV incidence proposed by Sommen et al. [2009] using a penalized likelihood to obtain a smooth estimate of the HIV incidence curve. The penalized likelihood was calculated within the framework of a non-homogeneous multi-state Markov model. The HIV incidence curve was approximated using cubic M-splines and an approximation of the cross-validation criterion was used to estimate the smoothing parameter. The method is illustrated on 200 simulations reproducing the French HIV mandatory notification data and also on real data using the French HIV mandatory notification data.The model presented here is a new tool for estimating the incidence of HIV infection from surveillance data using information from clinical stages at the time of diagnosis, and taking into account the evolution of screening during time.

Keywords. HIV, incidence, multi-state Markov model, penalized likelihood, notification data, surveillance

1 Introduction

En épidémiologie, l'incidence est un indicateur majeur pour estimer la dynamique d'une maladie. Depuis plusieurs années, différentes méthodes pour estimer l'incidence de l'infection par le VIH ont été utilisées dans les pays développés et en développement à partir de différents modèles épidémiologiques et de différentes méthodes statistiques présentées dans le rapport de UNAIDS/WHO [2015].
Nous proposons une approche basée sur un modèle multi-états qui prend en compte l'histoire naturelle du VIH et distingue les individus non diagnostiqués des individus diagnostiqués. Nous avons adapté aux données françaises l'approche proposée par Sommen et al. [2009] en tenant compte du stade de primo-infection. Une approche par vraisemblance pénalisée a été utilisée afin d'obtenir une courbe lisse d'incidence du VIH. La vraisemblance pénalisée a été calculée dans le cadre d'un modèle de Markov non homogène. La courbe d'infection du VIH a été approximée à l'aide de M-splines cubiques et une approximation de validation croisée a été utilisée pour choisir le paramètre de lissage. Ce nouveau modèle prend en compte toutes les informations fournies par le stade clinique déterminé par un médecin au moment du diagnostic VIH ainsi que les changements potentiels de comportement face au dépistage au cours du temps.
La méthode est illustrée sur les données de surveillance du VIH en France. En outre, un jeu de données simulées se rapprochant au maximum des données de déclaration obligatoire du VIH a été créé afin d'évaluer les performances du modèle pour reconstruire la courbe d'incidence du VIH. La section 2 présente les données de la déclaration obligatoire du VIH et la méthode de simulation. Le modèle de Markov est présenté dans la section 3 et les résultats sont décrits dans la section 4. Enfin, nous discutons des avantages et des inconvénients de ce modèle ainsi que des perspectives envisagées.

2 Matériels

Depuis mars 2003, le VIH est à déclaration obligatoire en France. Certaines informations épidémiologiques et cliniques concernant le patient, telles que la profession, la nationalité, la raison du test, les antécédents négatifs ou positifs de sérologie, le stade clinique ou le mode de contamination sont collectées par le médecin au moment du diagnostic. De plus, une surveillance virologique est effectuée pour déterminer le type de virus au moment du diagnostic et déterminer si l'infection est récente (moins de 6 mois) ou non récente. Pour la période de 2004 à 2018, nous disposons, pour chaque semaine, du nombre estimé de personnes diagnostiquées avec le VIH ainsi que de leur stade clinique au moment du diagnostic.
Nous avons réalisé une étude de simulation afin de créer un jeu de données réaliste, aussi proche que possible des données de la déclaration obligatoire du VIH, permettant d'évaluer les performances du modèle pour estimer l'incidence du VIH et les autres indicateurs épidémiologiques d'intérêt (nombre d'individus diagnostiqués / non diagnostiqués).

L'étude de simulation a permis de générer 200 jeux de données simulées.

3 Méthode

3.1 Description du modèle

Le modèle multi-états utilisé pour décrire l'histoire naturelle de l'infection par le VIH, l'accès au diagnostic et la mortalité pré-sida est illustré en Figure 1.

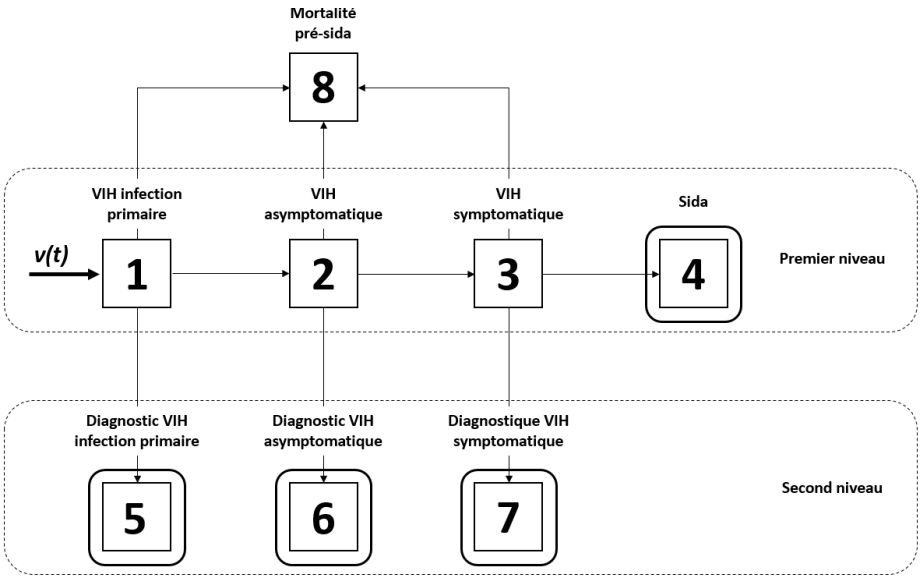


Figure 1: Modèle de Markov multi-états décrivant la progression de l'infection à VIH. Le premier niveau correspond à l'histoire naturelle du VIH avec $\nu(t)$ représentant l'intensité de l'infection du VIH au temps t. Le deuxième niveau représente l'accès aux diagnostics. Les états entourés (4, 5, 6 et 7) correspondent aux états pour lesquels des données sont disponibles.

Nous supposons que les individus nouvellement infectés entrent en temps continu dans l'état 1 selon un processus de Poisson d'intensité $\nu(t)$ représentant l'incidence de l'infection par le VIH. Les individus infectés peuvent ensuite progresser successivement à travers quatre stades cliniques.
L'objectif principal est d'estimer l'intensité de transition $\nu(t)$ du processus de Poisson non-homogène à partir des données de la déclaration obligatoire du VIH. Les intensités de transition du premier niveau sont connues à partir de l'article de Alioum et al. [2005] et nous faisons l'hypothèse que le taux de mortalité pré-sida est le même que le taux de

mortalité dans la population générale.
On considère un processus de Markov non homogène à temps discret, avec des probabilités de transition qui dépendent du temps. Comme les arrivées en 1 se font selon un processus de Poisson, les arrivées en 4, 5, 6 et 7 se font également selon un processus de Poisson.

3.2 Vraisemblance du modèle

La vraisemblance du modèle est basée sur un processus de Poisson à temps discret dont l'intensité est estimée à partir de la relation de récurrence décrite dans l'approche de Aalen et al. [2007], et qui se base à la fois sur l'incidence $\nu(t)$ mais aussi sur les probabilités de transition d'un état à l'autre. En utilisant cette approche, il est possible d'écrire le nombre attendu d'individus dans les états 1 à 8 au temps t_i, $i = 1, ..., K$.

La déclaration obligatoire du VIH fournit les informations sur le nombre estimé de diagnostics VIH. On note n_i^k le nombre de sujets diagnostiqués VIH positifs et inclus dans le système de surveillance dans l'état k dans l'intervalle $T_i = (t_{i-1}, t_i]$, $i = 1, 2, ..., K$. On note e_i^k le nombre de sujets attendu dans l'état k dans l'intervalle T_i, $i = 1, 2, ..., K$ calculé à partir de la formule de récurrence de Aalen et al. [2007] et Birrel et al. [2012] .

En utilisant l'hypothèse de Poisson, la log-vraisemblance du modèle, permettant d'estimer la courbe d'incidence du VIH $\nu(t)$, s'exprime comme suit :

$$l = \sum_{i=H}^{K} \sum_{j=4}^{7} n_i^j \log(e_i^j) - e_i^j$$

La somme commence au temps $T_i = T_H$ qui correspond au début de l'année 2004 et se termine pour $T_i = T_K$ correspondant à la fin de l'année 2018, période au cours de laquelle nous disposons des données.

3.3 Vraisemblance pénalisée

Il est souhaitable de produire une estimation lisse de la courbe d'incidence $\nu(t)$ de sorte que cette fonction n'ait pas de valeurs négatives, soit continue, et dans le cas de l'épidémie de VIH, ait de petites variations locales. Une méthode largement utilisée consiste à pénaliser la vraisemblance par un terme de pénalisation. Le facteur de pénalité λ est utilisé comme paramètre de lissage pour déterminer le degré de lissage de la courbe d'infection $\nu(t)$. La fonction de pénalisation utilisée $\int \nu''(u)^2 \mathrm{d}u$ est basée sur les dérivées secondes de $\nu(t)$ car cette pénalisation permet de lisser la courbure de la fonction d'après le travail de Good and Gasking [1971] et Good and Gasking [1980].
On considère pl la log-vraisemblance pénalisée de la forme :

$$pl = l - \lambda \int \nu''(u)^2 \mathrm{d}u$$

Les paramètres à estimer sont les vecteurs de probabilités de diagnostic du VIH $\alpha_{1,5} = (\alpha^1_{1,5}, \alpha^2_{1,5}, \ldots, \alpha^K_{1,5})$, $\alpha_{2,6} = (\alpha^1_{2,6}, \alpha^2_{2,6}, \ldots, \alpha^K_{2,6})$ et $\alpha_{3,7} = (\alpha^1_{3,7}, \alpha^2_{3,7}, \ldots, \alpha^K_{3,7})$ pour les différents temps et la courbe d'incidence du VIH $\nu(t)$. Pour une valeur fixe de λ, la maximisation de la vraisemblance pénalisée pl en $\Theta_\lambda = (\nu(.), \alpha_{1,5}, \alpha_{2,6}, \alpha_{3,7})$ fournit des estimateurs de maximum de vraisemblance pénalisée $\hat{\nu}(.)$, $\hat{\alpha}_{1,5}$, $\hat{\alpha}_{2,6}$ et $\hat{\alpha}_{3,7}$ pour le modèle illustré en Figure 1.
Toutefois $\hat{\nu}(.)$ ne peut être déterminée de forme explicite. L'estimateur du maximum de vraisemblance $\hat{\nu}(.)$ est donc approché par une forme $\tilde{\nu}(.)$ à l'aide de M-splines d'ordre 4 d'après Joly and Commenges [1998]:

$$\tilde{\nu}(.) = \sum_{j=1}^{Q+2} \theta_j M_j(.)$$

Avec Q le nombre de nœuds uniformément réparti sur la période d'étude.
Pour une valeur fixée de λ, nous estimons le vecteur de paramètres $\hat{\Theta}_\lambda = (\hat{\theta}, \hat{\alpha}_{1,5}, \hat{\alpha}_{2,6}, \hat{\alpha}_{3,7})$ où $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, ...\hat{\theta}_{Q+2})$, qui maximise la log-vraisemblance pénalisée. Le facteur de pénalisation λ est estimé en utilisant une approximation du critère de validation croisée.

4 Résultats

La période d'étude commence à la première semaine de 2004 et se termine à la dernière semaine de 2018, et toutes les durées ont été calculées en semaines à partir du début de l'année 2004. Nous considérons 5 périodes de temps de 3 ans dans lesquelles les probabilités de transition étaient supposées constantes (mais varient d'une période à l'autre). Ce découpage a été choisi arbitrairement pour obtenir des périodes de même taille et un nombre raisonnable de paramètres à estimer.
Sur les 200 simulations effectuées, l'algorithme a convergé vers le même maximum en partant de différentes valeurs initiale de paramètres. Concernant les données de la déclaration obligatoire du VIH, nous disposons de 75 bases imputées après traitement des données manquantes. Nous avons appliqué le modèle sur ces 75 bases. L'algorithme de maximisation de la vraisemblance du modèle a convergé pour chacune des bases et vers le même maximum en partant de différentes valeurs de paramètres initiaux. Les règles de Rubin ont été appliquées pour obtenir une estimation de la courbe d'incidence $\nu(t)$ ainsi que du nombre attendu de personnes dans les différents états. Afin d'approximer la variance de nos estimateurs, nous avons utilisé l'approche introduite par Whaba [1983] qui propose une méthode bayésienne pour obtenir des bandes de confiance point par point de l'incidence du VIH.

5 Discussion

Le modèle multi-états présenté ici est une approche pour estimer conjointement l'incidence, le nombre de personnes vivant avec le VIH et ne connaissant pas leur statut de séropositivé, ainsi que le délai entre infection et diagnostic. Ce nouveau modèle prend en compte l'information apportée par les stades cliniques au moment du diagnostic ainsi que les changements dans le temps de comportement face au dépistage. Des splines ont été utilisées pour modéliser la courbe d'incidence du VIH. Les splines ont l'avantage de fournir une estimation lisse de la courbe d'infection sans hypothèses paramétriques fortes sur la courbe d'infection d'après les travaux de Bellocco and Pagano [2001] et de Rosenberg and Goedert [1998]. Nous avons créé un jeu de données simulées aussi proche que possible des données de la déclaration obligatoire du VIH pour évaluer la capacité du modèle à reconstruire la courbe d'incidence du VIH. Notre idée était de tester le modèle sur des données qui ressemblent autant que possible aux données réelles, tout en sachant que ce choix de simulation implique de fortes contraintes non spécifiques à notre modèle. Les avantages de cette méthode sont un temps de calcul raisonnable (48 heures pour la validation croisée et environ 3 heures pour la maximisation de la vraisemblance pénalisée), une estimation conjointe de l'incidence, du nombre de personnes ne connaissant pas leur statut de séropositivé, la distribution du temps entre infection et diagnostic, ainsi qu'une estimation plus précise de la courbe d'incidence dans les 3 dernières années par rapport aux méthodes similaires.
La méthode a été illustrée sur les données de la déclaration obligatoire du VIH dans la population générale en France. Les résultats obtenus sont cohérents avec les estimations précédentes de l'incidence du VIH en France utilisant d'autres approches telles que celles proposées par Le Vu et al. [2010] ou Marty et al. [2018].
Cette méthode pourra être utilisée en routine pour estimer les indicateurs chaque année et si nécessaire plus fréquemment. Cette méthode n'est pas basée sur des hypothèses fortes propres aux données françaises du VIH. Elle peut être appliquée à tout pays qui dispose d'un système de surveillance du VIH. Ces travaux pourraient potentiellement fournir un cadre pour appliquer cette méthode à d'autres maladies infectieuses en modifiant l'histoire naturelle de la maladie.
En termes d'analyses futures, l'incidence du VIH est significativement associée au sexe, au mode de transmission, à l'origine géographique et également à la région de résidence. Une perspective intéressante sera de décliner les indicateurs épidémiologiques en fonction de ces variables, ce qui fera l'objet d'un futur travail.

References

C. Sommen, A. Alioum, and D. Commenges. A multistate approach for estimating the incidence of human immunodeficiency virus by using HIV and AIDS french surveillance data. *Stat Med*, 28(11):1554–68, 2009.

UNAIDS/WHO. Estimating hiv incidence using hiv case surveillance. 2015.

A. Alioum, D. Commenges, and R. Thiébault. A multistate approach for estimating the incidence of human immunodeficiency virus by using data from a prevalent cohort study. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 54:739–752, 2005.

O. O. Aalen, V. T. Farewell, D. De Angelis, N. E. Day, and O. N. Gill. A Markov model for HIV disease progression including the effect of hiv diagnosis and treatment: application to aids prediction in England and Wales. *Stat Med*, 16(19):2191–2210, 2007.

P. Birrel, T. Chadborn, O. Noel Gill, Valerie. Delpech, and D. De Angelis. Estimating trends in incidence, time-to-diagnosis and undiagnosed prevalence using a cd4-based bayesian back-calculation. *Statistical Communications in Infectious Diseases*, 4, 2012.

I.J. Good and R.A. Gasking. Nonparametric roughness penalties for probability densities. *Biometrika*, 58(2):255–277, 1971.

I.J. Good and R.A. Gasking. Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data. *Journal of the American Statistical Association*, 75(369):42–56, 1980.

P. Joly and D. Commenges. A penalized likelihood approach for arbitrarily consored and truncated data: application to age-specific incidence of dementia. *Biometrics*, 54: 203–212, 1998.

G. Whaba. Bayesian confidence intervals for the cross-validated smoothing splines. *Journal of the Royal Statistical Society, B*, 45:133–150, 1983.

R. Bellocco and M. Pagano. Multinomial analysis of smoothed hiv back-calculation models incorporating uncertainty in the aids incidence. *Stat Med*, 20(13):2017–2033, 2001.

P. Rosenberg and J. Goedert. Estimating the cumulative incidence of hiv infection among persons with haemophilia in the united states of america. *Stat Med*, 17(2):155–168, 1998.

S. Le Vu, Y. Le Strat, F. Barin, J. Pillonel, F. Cazein, V. Bousquet, S. Brunet, D. Thierry, C. Semaille, L. Meyer, and J.C. Desenclos. Population-based hiv-1 incidence in france, 2003–08:a modelling analysis. *Lancet infectious diseases*, 10, 2010.

L. Marty, F. Cazein, H. Panjo, J. Pillonel, D. Costagliola, V. Supervie, and Hermetic Study Group. Revealing geographical and population heterogeneity in HIV incidence, undiagnosed HIV prevalence and time to diagnosis to improve prevention and care: estimates for france. *J Int AIDS Soc*, 21(3), 2018.

Functional Peaks-over-threshold Analysis and its Applications in Environment

Raphaël de Fondeville [1] & Anthony C. Davison[2]

[1] *raphael.de-fondeville@epfl.ch - Swiss Data Science Center, INN 218, Station 14, 1015 Lausanne, Switzerland*

[2] *anthony.davison@epfl.ch - Chair of Statistics, École polytechnique fédérale de Lausanne, Station 8, 1015 Lausanne, Switzerland*

Résumé. Les techniques de quantification du risque naturel se sont largement démocratisées ces dernières années. Elles sont encore toutefois majoritairement limitées à la simple utilisation de catalogues d'évènements historiques, dont la taille excède rarement 40 à 50 ans, ainsi qu'à l'exploitation de modèles numériques, impliquant de lourds calculs tout en n'étant pas fiable pour l'extrapolation. La théorie des valeurs extrêmes définit les principes d'analyse statistiques nécessaires à l'estimation de la fréquence d'évènements rares tout en donnant un cadre formel pour extrapoler au-delà des niveaux historiques d'intensité. Toutefois son application s'est jusque-là principalement limitée au cadre univarié. Ainsi, une majorité des études traitant du risque naturel ont négligé sa nature spatio-temporelle.

Dans cette présentation, nous introduisons une extension de l'analyse de dépassements de seuil au cadre fonctionnel, dans lequel il est possible de caractériser un évènement extrême complexe à travers une notion généralisée d'excès, et décrivons ensuite la limite de leur queue de distribution, appelée processus de r-Pareto généralisé. Nous présentons un modèle dérivé de fonctions aléatoires log-Gaussiennes qui utilise les structures classiques de covariance pour caractériser la dépendance extrémale. Ensuite, nous décrivons un générateur stochastique d'évènements extrêmes, capable de quantifier la récurrence d'évènements passés ainsi que d'en générer des nouveaux dont l'intensité va au-delà des niveaux historiques. La méthodologie est ensuite appliquée à plusieurs risques naturels tels que les tempêtes et la pluie.

Mots-clés. Analyse de dépassements de seuil, Extrêmes spatio-temporels, Processus de r-Pareto généralisé, Risque naturel.

Abstract. Estimating the risk of single occurrences of natural hazards has become important in recent decades, but up until now it has been largely limited to re-using catalogues of historical events, which usually do not exceed 40 to 50 years in length, and to numerical models, which require heavy computation and are often unreliable for extrapolation. Extreme value theory provides statistical methods for estimating the frequency of past extreme events as well as for extrapolating beyond observed severities, but it has mostly been focused on studying univariate quantities. Consequently the majority of its applications to natural hazards have neglected their spatio-temporal characteristics.

We present an extension of peaks-over-threshold analysis to functions which allows one to define complex extreme events as special types of exceedances, and then obtain their limit tail distribution, namely the generalized r-Pareto process. We focus on a specific model based on log-Gaussian random functions using classical covariance structures to characterize extremal dependence. Then, we describe a stochastic weather generator for extreme events, capable of quantifying the recurrence of past events as well as generating completely new ones. The methodology is applied to several natural hazards such as windstorms and rainfall.

Keywords. Generalized r-Pareto process, natural hazards, peaks-over-threshold analysis, spatio-temporal extremes.

1 Extended summary

Extreme Value Theory (EVT) provides a theoretical framework to describe and model tails of statistical distributions within which estimating the frequency of past extreme events as well as to extrapolating beyond observed severities is possible. These have been extensively studied in a univariate framework (Fisher and Tippett, 1928; Gnedenko, 1943; Davison and Smith, 1990) especially for independent identically distributed replicates, and applications have been developed in fields such as finance, insurance, hydrology and telecommunications (Hosking and Wallis, 1987; Katz et al., 2002; Embrechts et al., 1997). Due to recent extreme events, there has been a surge of interest in environmental applications, motivated by the necessity to better understand the impact of global warming. Floods, windstorms, heatwaves have a complex spatio-temporal structure that cannot be modelled using univariate extreme value theory.

Max-stable processes (de Haan and Ferreira, 2006, Section 9.2), which provide a functional extension of the generalized extreme value distribution (Coles, 2001, p.47-48), have successfully been used to study the extremal behaviour of monthly and annual maxima, but applications have been limited due to the mathematical and computational complexity of such models (Huser and Davison, 2013). Also, the study of maxima discards a fair amount of information, making detection of mixtures in tail behaviour very difficult. For example, in some regions, rainfall can be divided into two classes: convective rain, which is local and marginally very intense, and cyclonic spells generating larger spatial accumulations of water but with lower local intensities. These phenomena are driven by different independent weather conditions that may both cause severe floods and their tail marginal distribution and spatio-temporal structure are likely to differ. With block maxima, marginally intense events naturally dominate and thus impose a focus on convective rainfall, while disregarding potential extreme cyclonic events. For risk mitigation, studying extremes of different natures is crucial, and max-stable processes are inappropriate

for modelling such complex phenomena, since taking maxima largely eliminates certain types of events.

Univariate peaks-over-thresholds analysis, associated to the generalized Pareto distribution, define extreme events as exceedances over a threshold. In this context, reduction of multivariate datasets to univariate structural variables, such as $\max(X_1, X_2)$ or $X_1^2 + X_2^2$, on which generalized Pareto distributions are fitted (Coles and Tawn, 1994), is common to study complex multivariate extreme events. However, this approach does not give insight on the combination of events yielding an exceedance and is hindered by the fact that different univariate summaries may lead to different tail behaviour. One way to understand these differences is to suppose that the observations are generated by an underlying mixture of generative processes, which are disentangled by computing these univariate summaries. Thus if the summary captures only one of these processes, for instance only cyclonic rain, it is not surprising that we obtain different tail behaviours. Functional peaks-over-threshold analysis generalizes this methodology for a better understanding of the underlying dependence structure and gives a theoretical foundation to detect mixtures of tail behaviour through different definitions of exceedances tailored to the type of extreme events of interest.

In univariate extreme value theory, the generalized Pareto distribution gives a unified framework to describe directly the tail decay of the original data, and encompasses the Weibull, Gumbel and Fréchet tail decay regimes. This work provides a similar unified formulation for functional peaks-over-threshold analysis under the assumption that the process has the same tail decay over its domain. In this context, we extend Dombry and Ribatet (2015) by introducing the generalized r-Pareto process, allowing more flexible excess definitions and generalized Pareto tail margins. The generalized r-Pareto process is the only limit of exceedances of a properly rescaled regularly varying process and for some specific definitions of exceedance, it can be factorized to enable simulation of events with a fixed intensity, i.e., events for which the risk measure equals a pre-determined return level.

We first review classical results for univariate extremes and introduces functional peaks-over-threshold analysis. We present convergence results for the three possible regimes of tail decay, under a generalized regular variation hypothesis, i.e., for a stochastic process X, we assume that there exist a tail index $\xi \in \mathbb{R}$ and sequences of functions $a_n > 0$ and b_n such that

$$n\Pr\left\{\left(1 + \xi\frac{X - b_n}{a_n}\right)^{1/\xi} \in \cdot\right\} \to \Lambda(\cdot), \quad n \to \infty,$$

where Λ is a non-zero measure on the space of non-negative continuous functions. We then introduce the class of generalized r-Pareto process, characerized by

$$P = r(a)\xi^{-1}R^{\xi}\frac{W}{\mathrm{r}(W)} + b - \xi 1a,$$

where R is a unit Pareto random variable, W is a stochastic process on the space of continuous functions with unit norm and $a > 0$ and b are continuous functions. For linear risk functionals, we prove that generalized r-Pareto processes are the only limit of increasingly large r-exceedances, i.e., events $\{r(X) > u\}$ with increasing threshold $u \in \mathbb{R}$. The previous result is then applied to develop of stochastic weather generator of windstorms over Europe. Finally we illustrate the importance of risk definition when studying the risk of flooding in the city of Zurich.

References

Coles, S. G. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer, London.

Coles, S. G. and Tawn, J. A. (1994). Statistical Methods for Multivariate to Structural Design Extremes: an Application to Structural Design. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 43(1):1–48.

Davison, A. C. and Smith, R. L. (1990). Models for Exceedances over High Thresholds (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological)*, 52(3):393–442.

de Haan, L. and Ferreira, A. (2006). *Extreme Value Theory: An Introduction*. Springer, New York, USA.

Dombry, C. and Ribatet, M. (2015). Functional Regular Variations, Pareto Processes and Peaks Over Thresholds. *Statistics and Its Interface*, 8(1):9–17.

Embrechts, P., Klüppelberg, C., and Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*. Springer-Verlag, Berlin.

Ferreira, A. and de Haan, L. (2014). The Generalized Pareto Process; with a View Towards Application and Simulation. *Bernoulli*, 20(4):1717–1737.

Fisher, R. A. and Tippett, L. H. C. (1928). Limiting Forms of the Frequency Fistribution of the Largest or Smallest Member of a Sample. *Mathematical Proceedings of the Cambridge Philosophical Society*, 24(2):180–190.

Gnedenko, B. (1943). Sur la Distribution Limite du Terme Maximum d'une Série Aléatoire. *Annals of Mathematics*, 44(3):423–453.

Hosking, J. R. M. and Wallis, J. R. (1987). Parameter and Quantile Estimation for Generalized Pareto Distribution. *Technometrics*, 29(3):339–349.

Huser, R. and Davison, A. C. (2013). Composite Likelihood Estimation for the Brown–Resnick Process. *Biometrika*, 100(2):511–518.

Katz, R. W., Parlange, M. B., and Naveau, P. (2002). Statistics of Extremes in Climatology. *Advances in Water Resources*, 25(8-12):1287–1304.

Klüppelberg, C. and Resnick, S. I. (2008). The Pareto Copula, Aggregation of Risks, and the Emperor's Socks. *Journal of Applied Probability*, 45(1):67–84.

MODELISATION DE NON STATIONNARITES PAR LE VARIOGRAMME EMPIRIQUE

Chantal de Fouquet[1], Léa Pannecoucke[1], Mathieu Le Coz[2]& Xavier Freulon [1]

[1] *Mines ParisTech, Ecole des Mines de Paris, Centre de géosciences, 35 rue Saint-Honoré, 77305 Fontainebleau Cedex. lea.pannecoucke@mines-paristech.fr, xavier.freulon@mines-paristech.fr, chantal.de_fouquet@mines-paristech.fr*

[2] *Institut de Radioprotection et de Sûreté Nucléaire (IRSN), PSE-ENV/SEDRE, 31 avenue de la Division Leclerc, 92262 Fontenay-aux-Roses Cedex. mathieu.lecoz@irsn.fr*

Résumé. Plusieurs modèles permettent de représenter des phénomènes non stationnaires en vue du krigeage : FAI-K, covariables en dérive externe ou plus récemment, les modèles de type SPDE. Mais ces modèles ne décrivent pas l'ensemble des non stationnarités d'ordre un ou deux. Dans les méthodes multi-points, la loi spatiale ou la covariance empiriques ne sont pas nécessairement des modèles admissibles, lorsqu'elles sont calculées sur une seule image d'entraînement.

Calculés par couples de points *(x,x')* en moyenne sur *N* réalisations, la covariance *C(x,x')* ou le variogramme $\gamma(x,x')$ non stationnaires empiriques admettent les propriétés mathématiques adéquates. Soit alors un phénomène modélisable à l'aide d'un code numérique. La randomisation des paramètres d'entrée fournit un lot de sorties des modèles numériques (spatiaux ou spatio-temporels) qui constituent autant de réalisations d'une Fonction Aléatoire. La covariance ou le variogramme non stationnaires empiriques reflètent alors la variabilité spatiale (spatio-temporelle) de la variable de sortie, ce qui permet d'introduire cette variabilité dans le krigeage.

Les propriétés de la covariance et du variogramme empiriques sont présentées. L'application à la modélisation d'un panache de pollution est discutée.

Mots-clés. Non stationnarité, krigeage, covariance empirique, variogramme.

Abstract. Various non stationary models are available to describe non stationary phenomena for kriging, such as IRF-k, covariates as external drifts, and more recently SPDE. However, these models are not suited for all non stationary phenomena. In addition, in the multi-point approaches the empirical spatial covariance or distribution are not necessary admissible when computed on a single realization.

The covariance *C(x,x')* and the empirical variogram $\gamma(x,x')$ computed on the average of pairs of points *(x,x')* over *N* realizations have the relevant mathematical properties. Let us thus consider a phenomenon that can be modelled using a numerical code. The randomisation of the input parameter fields provides a set of outputs from the numerical models (spatial or spatio-temporal), which constitute as many realisations of a Random Function not necessarily stationary. The numerical covariance or variogram reflect the spatial (spatio-temporal) variability of the studied variable, allowing this variability to be introduced into the kriging.

The properties of the empirical covariance and variogram are presented. An application to the modelling of a contamination plume is discussed.

Keywords. Non stationarity, kriging, empirical covariance, variogram

Introduction

Lorsque les données sont peu nombreuses, le résultat du krigeage ne reflète généralement pas la structure spatiale de la variable. Par exemple, avec seulement quelques données, la carte des concentrations estimées ne reproduit pas le développement d'un panache suivant le sens d'écoulement de la nappe. Par ailleurs, les résultats des modèles d'écoulement et de transport ne coïncident pas avec les observations, et les méthodes inverse sont nécessaires pour recaler les sorties des modèles aux données.

On propose ici d'utiliser un code de calcul pour générer un lot de champs physiquement réalistes de la variable étudiée (la concentration), sur lesquels calculer un variogramme ou une covariance empiriques, qui reflètent la variabilité spatiale attendue. Ce variogramme empirique est ensuite utilisé pour le krigeage.

1. Krigeage avec variogramme empirique

La démarche est illustrée pour l'estimation d'un panache de pollution dans un sol non saturé (Pannecoucke et al., 2020), et récapitulée à la Figure 1.

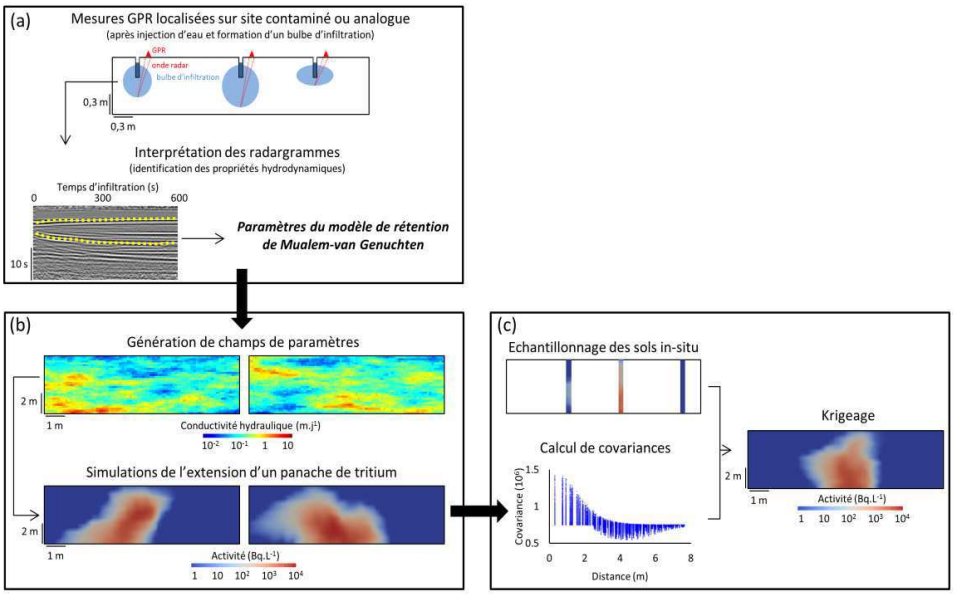


Figure 1 : caractérisation des paramètres hydrodynamiques de la zone non saturée (a) ; modélisation hydrogéologique à base physique à partir d'une source ponctuelle, par le code MELODIE (b) ; modélisation géostatistique pour la spatialisation par krigeage des concentrations (c).

1.1 Covariance et variogramme non stationnaires empiriques

Il est facile de montrer que la covariance empirique, calculée sur un lot de réalisations (Figure 2), est de type positif, et le variogramme empirique de type négatif conditionnel (de Fouquet, 2019).

Figure 2 : Schéma de principe du calcul de la covariance *C(x,t ;x',t')* ou du variogramme $\gamma(x,t;x',t')$ non stationnaires spatio-temporels empiriques à partir de *N* simulations.

1.2 Krigeage

Ces propriétés ont été utilisées précédemment pour le krigeage, lorsque la structure spatiale de la variable est essentiellement contrainte par les équations décrivant la physique de phénomène (Roth, 1995 ; Schwede et Cirpka, 2010).

Les tests montrent peu de différences selon que le krigeage est effectué en variogramme ou en covariance, et selon l'hypothèse de stationnarité sur la moyenne spatiale (ou spatio-temporelle). Les résultats reflètent le fort contraste des concentrations entre le panache et les zones non contaminées (Pannecoucke et al., 2019).

2. Résultats

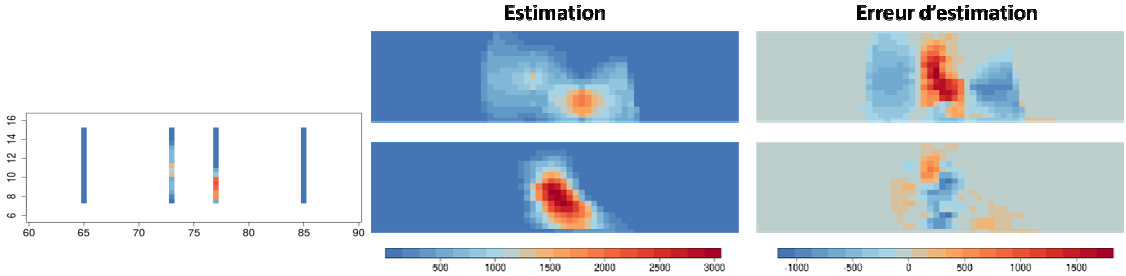


Figure 3 : Cas synthétique avec quatre sondages (à gauche). Krigeage avec un variogramme classique (en haut), et avec le variogramme empirique (en bas) ; erreurs d'estimation associées.

Les tests détaillés indiquent logiquement que le variogramme empirique est d'autant plus intéressant que les données sont peu nombreuses. La méthode permet de tenir compte des incertitudes sur les conditions aux limites et sur les paramètres du modèle (portée du champ de perméabilités, par exemple), par randomisation des paramètres d'entrée des simulations utilisées dans le calcul du variogramme empirique. Pour la prévision du développement du panache de la contamination, les résultats sont proches de ceux obtenus par le filtre de Kalman d'ensemble.

Le krigeage spatio-temporel est aussi utilisable pour « rechercher » la source, sous réserve de données proches de l'origine de la contamination, spatialement et temporellement.

Remerciements

Le projet Kri-Terres est financé par l'ANDRA dans le cadre du Programme d'Investissements d'Avenir.

Bibliographie

de Fouquet C. (2019) *Exercices corrigés de géostatistique.* Presses des Mines, Paris.

Pannecoucke L., Le Coz M. , Houzé C., Saintenoy A., Cazala C. et de Fouquet C. (2019) A Combining geostatistics and simulations of flow and transport to characterize contamination within the unsaturated zone. *Jounal of Hydrology,* 574, pp.160-168. https://doi.org/10.1016/j.jhydrol.2019.04.016

Pannecoucke L., Le Coz M., Freulon X., de Fouquet C. (2020) Combining geostatistics and simulations of flow and transport to characterize contamination within the unsaturated zone. *Science of the Total Environment* 699. https://doi.org/10.1016/j.scitotenv.2019.134216

Roth C. (1995) *Contribution de la géostatistique à la résolution du problème inverse en hydrogéologie.* Thèse de doctorat. École Nationale Supérieure des Mines de Paris.

Schwede R. et Cirpka O.A. (2010) Interpolation of steady-state concentration data by inverse modeling. *Groundwater* 48(4), pp. 569-579

Transfer learning to improve predictive models of product performances

Antoine de Mathelin [1,2] & François Deheeger [2] & Mathilde Mougeot [1,3]

[1] *Centre Borelli - ENS Paris-Saclay, 61 Avenue du Président Wilson 94235 Cachan*
[2] *Michelin, 23 place des Carmes Déchaux 63000 Clermont-Ferrand*
[3] *ENSIIE, 1 Rue de la Résistance 91000 Évry*

Résumé. Tout modèle construit à partir de données observées est fortement dépendant de son ensemble de données d'apprentissage et une question inhérente aux modèles de machine learning est la question de leur généralisation aux nouvelles observations. Cette question est d'autant plus critique pour les modèles de machine learning appliqués à la conception de produits, car, dans un soucis d'innovation, on cherche le plus souvent à estimer les performances de nouveaux produits qui peuvent avoir des spécifications sensiblement différentes des précédents. Dans ce travail, nous montrons sur le cas réel de la conception de pneus que les méthodes de transfer learning permettent d'améliorer de manière significative les performances des modèles de conception sur de nouveaux pneus.

Mots-clés. Transfer learning, Modèle de conception, Machine learning pour l'industrie

Abstract. Any data driven model is strongly dependent on its learning dataset and an inherent question of machine learning models is the issue of their generalization on new observations. In fact this issue is very critical when applying machine learning techniques to design models since one most often seeks to estimate the performances of new prototype products that may have significantly different specifications from previous products. In this work, we show on the real case of tire design space exploration that transfer learning methods help to notably improve the performances of design models on new tires.

Keywords. Transfer learning, Design model, Machine learning for infustry

1 Introduction

For the past few years, machine learning methods have proven to be powerful tools to solve complex tasks as for example image classification, events prediction, natural language processing... Because of these noteworthy results, the industrial sector has become particularly interested to apply these techniques in order to improve their own processes, as for instance quality management, maintenance or predictive models for design... Nevertheless, the use of machine learning for industrial problems encounter challenging issues which highlight the limitations of traditional machine learning methods. In a recent paper Wuest et al. (2016) recall the main difficulties: the lack of labeled data and the data obsolescence due to technological drifts... These issues are even more critical when building predictive models for design space exploration as designers are looking for new

innovative products, which may be significantly different from previous ones. They often need to explore in unknown design space (where a limited number of observations are available) and machine learning models, which are strongly dependent on their learning dataset, could unlikely provide pertinent results in these areas. However, based on the assumption that "new" products should have a behaviour somewhat similar to previous ones, we propose to use *transfer learning* methods to handle this kind of generalization issue in design predictive models. As we will develop in the next part, the goal of our work is to "transfer" the information contained in previous tire lines data in order to build a model performing well on data from a "new" tire line. See Figure 1.

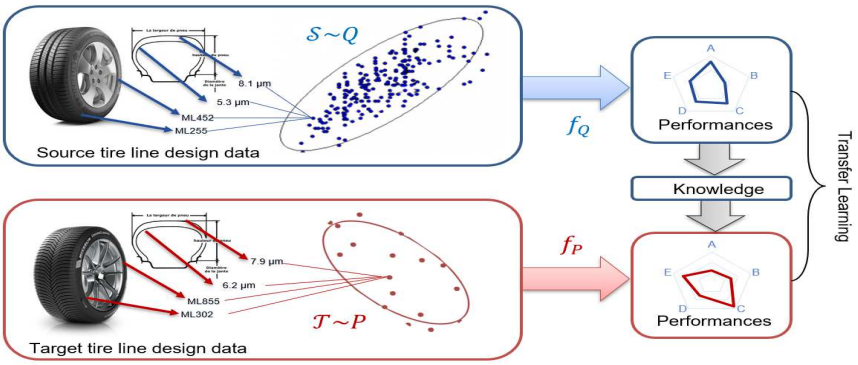


Figure 1: **Transfer learning between tire lines**: Information on the relationship between tire components and performances are *transferred* from a *source* tire line with a large amount of observations available to a *target* tire line with few observation available.

2 Transfer learning

2.1 Framework

Our work focuses on improving the generalization of design predictive models of tire performances. For this purpose we consider the familiar supervised learning setting where the learning algorithm receives a sample of m labeled points $\mathcal{S} = (z_1, ..., z_m) = ((x_1, y_1), ..., (x_m, y_m)) \in (X \times Y)^m$, where $X \subset \mathbb{R}^p$ is the input space such that $x_i \in X$ are input instances vector encoding the tire features (width, height, dimension of rubber layers, materials properties...) and p the number of features. $Y \subset \mathbb{R}$ is the output space, where y_i are a corresponding performance observed for the tire encoded by x_i (for example the tire rolling resistance). Building a predictive model for tire design consists in building an estimator or hypothesis $h \in H$ which approximates the function $f : X \to Y$ modeling the relationship between the tire design features and one performance. H is a class of hypotheses, for example the set of fully-connected regressive neural-network functions.

As mentioned previously, predictive models of tire performances encounter a generalization issue when applying one model built on "existing" tire lines data on data of a "new" tire line. To characterize this issue of generalization in the transfer learning framework, we consider $\mathcal{S}$ as the "existing" tire lines dataset and we denote by $\mathcal{T} = ((x_1, y_1), ..., (x_n, y_n)) \in (X \times Y)^n$ the labeled instances of a "new" tire line from which only a few observations are available ($n << m$). In the transfer learning framework, one makes the assumption that the sample $\mathcal{S}$ is drawn according to a source distribution Q, while the sample $\mathcal{T}$ is drawn according to a target distribution P that may somewhat differ from Q.

Predictive model for tire design should perform well on new tire lines, i.e for instances drawn according to P. Our main goal consists then in selecting a hypothesis h out of a hypothesis set H with a small expected loss according to the target distribution P, i.e minimizing $\mathcal{L}_P(h, f) = \mathrm{E}_{x\sim P}[L(h(x), f(x))]$ where $L : Y \times Y \to \mathbb{R}$ is a loss function defined over pair of performance labels.

Transfer learning framework has been first theoretically introduced in the works of Ben-David et al. (2010) and Y. Mansour et al. (2009). The latest presents interesting bounds of the expected error on target distribution $\mathcal{L}_P(h, f)$ in the transfer learning setting. These bounds provide valuable information about the conditions needed to make successful transfer. In particular, based on the bounds provided in Theorem 8, Corollary 5 and 7 from Mansour et al. (2009) we are able to adapt the following formula for our transfer issue:

$$\mathcal{L}_P(h, f) \leq \mathcal{L}_{\widehat{P}}(h, f) + \mathcal{L}_{\widehat{Q}}(h, h^*_Q) + \mathrm{disc}_{L_q}(\widehat{P}, \widehat{Q}) + (4q + 1)\left(\widehat{\mathfrak{R}}_{\mathcal{S}}(H) + \widehat{\mathfrak{R}}_{\mathcal{T}}(H)\right) + 4M\left(\sqrt{\frac{\log\frac{4}{\delta}}{2m}} + \sqrt{\frac{\log\frac{4}{\delta}}{2n}}\right) + \mathcal{L}_Q(h^*_Q, h^*_P) \text{ with probability at least } 1 - \delta \quad (1)$$
$$\text{for any } 0 < \delta < 1$$

Where:

- $h \in H$ is a hypothesis and H is a hypothesis set bounded by some $M > 0$ for the loss function L_q, i.e $L_q(h, h') \leq M$ for all $h, h' \in H$.
- Q, P are the source and target distributions and $\widehat{Q}$, $\widehat{P}$ their empirical estimators.
- h^*_Q, h^*_P the optimal hypotheses on respectively Q and P.
- $\mathrm{disc}_L(P, Q) = \max_{h',h\in H} |\mathcal{L}_P(h, h') - \mathcal{L}_Q(h, h')|$ is called the *discrepancy distance* and characterizes the ability to find on class H a hypothesis which would perform well on source domain and poorly on target one or inversely.

- $\widehat{\mathfrak{R}}_{\mathcal{S}} = \frac{2}{m}\mathrm{E}_{\sigma}\left[\sup_{h\in H} | \sum_{i=1}^{m} \sigma_i h(x_i) | \mathcal{S} = (x_1, ..., x_m)|\right]$ with σ_i independent uniform random variables, is called the *Rademacher* distance and characterizes the ability of hypotheses from class H to fit noise on $\mathcal{S}$.

We deduce from this bound that efficient transfer is possible when h performs well on target empirical data ($\mathcal{L}_{\widehat{P}}(h, f_P)$), stays close to the best hypothesis on the source distribution ($\mathcal{L}_{\widehat{Q}}(h, h^*_Q)$) and minimizes the discrepancy distance between empirical distributions ($\mathrm{disc}_{L_q}(\widehat{P}, \widehat{Q})$). Besides, the hypothesis class H should be regularized enough to avoid over-fitting ($\widehat{\mathfrak{R}}_{\mathcal{S}}(H) + \widehat{\mathfrak{R}}_{\mathcal{T}}(H)$). Notice that the last term $\mathcal{L}_Q(h^*_Q, h^*_P)$ informs that source and target optimal hypotheses should be close enough on the source domain to expect successful transfer.

2.2 Methods

Applied to transfer between tire lines, our work particularly studies and compares the ability of several transfer learning methods to improve design space exploration models. We selected transfer methods according to the three classes defined by Pan and Yang (2010): *parameter-based*, *feature-based* and *instance-based*:

- *parameter-based* methods aim to transfer the parameters of a hypothesis built on the source domain to build a new hypothesis for the target domain. For instance, Chelba et al. (2007) develop a transfer method where the parameters β_P of a hypothesis h_P built on $\mathcal{T}$ are penalized by their euclidean distance with the parameters β_Q of a hypothesis h_Q built on $\mathcal{S}$. In this manner, Chelba et al. try to obtain a good hypothesis on target empirical data which stays close to a good hypothesis learned on source data.

- In *feature-based* methods, the goal is to find a common feature space in which target and source data have the same behavior according to the output data. Oquab et al. (2014), for instance, use some target data to train the last layers of a neural network built on source data. In this way, they build ϕ such that $\mathrm{disc}_{L_q}(\phi(\widehat{P}), \phi(\widehat{Q}))$ is minimized, ϕ reducing the number of liberty degrees.

- In *instance-based* methods, weights of source data are modified to help the learning process on target data. These methods search for $\widehat{Q}'$ which reduces $\mathrm{disc}_{L_q}(\widehat{P}, \widehat{Q}')$. For instance Pardoe and Stone (2010) developed a method *TrAdaBoostR2* for transfer in regression cases. This method is based on a reverse boosting principle where the weights of source instances poorly predicted are decreased at each boosting iteration.

3 Application

We conduct the experiments on tire design data. The purpose of these experiments is to build a machine learning model mapping the component features of one tire (i.e. the characteristics of the materials, polymers and textiles in the tire) to one of its performances (here the rolling resistance).

The dataset used contains around 1000 row instances, each corresponds to one specific tire from which we have recorded the component features and the rolling resistance.

Each tire belongs to one tire line which is a set of product specifications that defines a design scope. As mentioned previously, the goal of this work is to build a machine learning model with good generalization property between tire lines. More specifically, we aim to build a model performing well on a "new" tire line using data from previous tire lines.

In the experiments, we select arbitrarily one tire line and withdraw all its corresponding instances from the dataset (around 100). This tire line will be considered as a "new" tire line for which we aim to have a good model despite the lack of observations. This tire line will be called *target* in the following and its 100 corresponding instances will be used to evaluate the performances of the different transfer methods. The remaining data, referred as *source*, are used in the training process. Additionally, as the considered transfer methods use a few observations from the target domain, we add 10 target instances in the training set.

The set of hypotheses H considered to build the machine learning models is the set of fully-connected regressive neural-network functions with two hidden layers of 100 neurons each and a *sigmoid* activation. We consider, besides, the mean squared error to compare the performances of the transfer methods.

Finally, we divide the experiments in two operational scenarii:

- First, we consider the case where tire designers have access to source data and to a few target observations. In this case, they are able to train a neural network with the set composed of both source and target data available. The model built in this manner is called *all* model. We compare its performances to the ones of the transfer method *TrAdaBoostR2* which also needs both source and target data.

- Then we consider the case where tire designers do not have access to source data but only to a few target observations and to a source model previously trained on source data. This scenario occurs in industry when source data are confidential or because the overwhelming number of source data induces computational burden in the training process. In this case, the transfer methods from Chelba and from Oquab are used. To evaluate the benefit of transfer methods, we compare their performances to the ones of the two basic models trained respectively on source data only (*source-only*) and on target data only (*target-only*).

Table 1 reports the mean squared error of the model predictions on the target data. It appears in both scenario that using transfer methods improve considerably the model performances. In particular *TrAdaBoostR2* presents the best results compared to the other methods. We observed also that using source data instead of pre-trained source models is a better approach in this case as the basic *all* method performs better than Chelba and Oquab transfer methods. However, as mentioned before, training a model with source and target data may not be possible for confidentiality reason or may lead to computational burden due to large source samples.

Scenario	Source data not available				Source data available	
Method	Srce-Only	Tgt-Only	Chelba	Oquab	All	Pardoe
RMSE	52 (6)	49 (2)	**29 (2)**	38 (3)	28 (2)	**23 (1)**

Table 1: **Summary of methods results:** 1000 source data and 10 target data are chosen randomly for the training phase, 100 others target data are used to compute the results. Experiments are conducted 10 times to obtain the standard deviation in brackets. In the "source data not available scenario", a source hypothesis is first pre-trained on the 1000 source data and used in a second time with the 10 target training data to build the target hypothesis. RMSE $\times$ 100 are given to mesure methods performances.

Bibliography

Ben-David, S., Blitzer J., Crammer, K., Kulesza, A., Pereira, F. and Vaughan, J. W. (2010). A theory of learning from different domains, *Machine Learning*, 79(1-2), 151–175.

Chelba, C., Silva, J., and Acero, A. (2007). Soft indexing of speech content for search in spoken documents, *Computational Speech and Language*, 21(3), 458–478.

Mansour, Y., Mohri, M., and Rostamizadeh, A. (2009). Domain adaptation: Learning bounds and algorithms, *COLT*.

Oquab, M., Bottou, L., Laptev, I. and Sivic, J. (2014). Learning and transferring mid-level image representations using convolutional neural networks, *CVPR*.

Pan, S. J. and Yang, Q. (2010). A survey on transfer learning, *IEEE Transactions on Knowledge and Data Engineering*, 22(10), 1345–1359.

Pardoe D. and Stone, P. (2010). Boosting for regression transfer, *Proceedings of the 27th International Conference on Machine Learning (ICML)*.

Wuest, T., Weimer, D., Irgens, C., and Thoben, K. D. (2016). Machine learning in manufacturing: advantages, challenges, and applications, *Production and Manufacturing Research*, 4(1), 23–45, 2016.

Sélection de variables sous contraintes de confidentialité différentielle locale

Amandine Dubois [1], Cristina Butucea [2] et Adrien Saumard [1]

[1] *CREST, ENSAI, Campus de Ker-Lann, Rue Blaise Pascal, BP 37203, 35712 Bruz cedex, France. amandine.dubois@ensai.fr; adrien.saumard@ensai.fr*
[2] *CREST, ENSAE, IPP, 5 avenue Henry Le Chatelier, F-91120 Palaiseau. cristina.butucea@ensae.fr*

Résumé. Nous nous intéressons à la sélection de variables pour l'espérance d'une loi normale d-dimensionnelle sous la contrainte supplémentaire que seules des données privatisées sont disponibles. À cette fin, nous adoptons une récente généralisation de la théorie minimax classique au cadre de la confidentialité différentielle locale. Nous donnons une borne inférieure et une borne supérieure sur la vitesse de convergence, pour le risque lié à la distance de Hamming, sur des classes de vecteurs s-sparses dont les coordonnées non nulles sont séparées de 0 par une constante $a > 0$. Nous en déduisons des conditions nécessaires et des conditions suffisantes pour que l'identification presque parfaite du support soit possible. En particulier, la borne inférieure obtenue montre que l'identification presque parfaite est impossible en grande dimension quand on impose des contraintes de confidentialité différentielle locale.

Mots-clés. Sélection de variables, confidentialité différentielle locale, identification presque parfaite du support.

Abstract. We address the problem of variable selection in the Gaussian mean model in $\mathbb{R}^d$ under the additional constraint that only privatised data are available for inference. For this purpose, we adopt a recent generalisation of classical minimax theory to the framework of local α-differential privacy. We provide a lower bound and an upper bound on the rate of convergence for the expected Hamming loss over classes of at most s-sparse vectors whose non-zero coordinates are separated from 0 by a constant $a > 0$. As corollaries, we derive necessary conditions and sufficient conditions for almost full recovery. In particular, our lower bound shows that almost full recovery is impossible under local differential privacy constraints in the high-dimensional setting.

Keywords. Variable selection, local differential privacy, almost full recovery.

1 Introduction

Problème. De nos jours, une grande quantité de données, telles que les dossiers médicaux, les informations de localisation des téléphones portables, l'historique de navigation

sur Internet, est collectée et stockée. Un enjeu majeur consiste à caractériser et à équilibrer l'utilité statistique de ces données et la protection de la vie privée des personnes auprès desquelles elles sont obtenues.

Nous nous intéressons ici au problème de la sélection de variables sous contrainte de confidentialité différentielle locale. Pour $i = 1, \dots, n$, le i-ème détenteur de données observe un vecteur aléatoire X_i de $\mathbb{R}^d$ distribué selon la loi $\mathcal{N}(\theta, \sigma^2 I_d)$. Le paramètre θ est supposé s-sparse et ses coordonnées non nulles sont supposées plus grandes qu'une certaine constante $a > 0$. L'objectif est que chaque détenteur de données publie une version anonymisée Z_i de X_i de sorte que la notion de confidentialité différentielle locale définie ci-dessous soit satisfaite et qu'on puisse identifier les coordonnées non nulles de θ à partir des données $Z_1, \dots, Z_n$ d'une manière optimale.

Confidentialité différentielle locale. La notion de *confidentialité différentielle locale* regroupe deux concepts différents, à savoir la confidentialité *locale* et la confidentialité *différentielle*. La notion qualitative de confidentialité *locale* caractérise la manière dont les différentes entités détenant les données $X_1, \dots, X_n$ peuvent interagir pour générer un échantillon privé Z qui pourra être communiqué. Elle s'oppose au concept de confidentialité *globale* où les détenteurs de données font confiance à une même autorité qui a accès à l'ensemble des données non masquées $X_1, \dots, X_n$ et qui génère, à partir de cette information complète, des données privées communicables. Dans la configuration *locale*, une telle autorité, en qui toutes les parties ont confiance, n'existe pas. Cependant un certain degré d'interaction entre les différentes parties est permis. Les données privées $Z_1, \dots, Z_n$ sont obtenues de la manière suivante : sachant $X_i = x_i$ et $Z_1 = z_1, \dots, Z_{i-1} = z_{i-1}$, le i-ème détenteur de données génère

$$Z_i \sim Q_i(\cdot \mid X_i = x_i, Z_1 = z_1, \dots, Z_{i-1} = z_{i-1})$$

pour un noyau de Markov Q_i. Un cas particulier important est celui de la confidentialité locale *non-interactive* où Z_i dépend seulement de X_i. La notion de confidentialité *différentielle* est une notion quantitative. Nous donnons sa définition dans le cas local et renvoyons à l'article de Wasserman et Zhou (2010) pour une définition dans le cas global.

Définition 1.1. Une suite de noyaux de Markov Q_i garantit la α-confidentialité différencielle locale si

$$\sup_{A \in \sigma(\mathcal{Z})} \frac{Q_i(A \mid X_i = x, Z_1 = z_1, \dots, Z_{i-1} = z_{i-1})}{Q_i(A \mid X_i = x', Z_1 = z_1, \dots, Z_{i-1} = z_{i-1})} \leq \exp(\alpha), \quad \forall x, x' \in \mathcal{X}.$$

Plus $\alpha \in (0, \infty)$ est petit, plus la contrainte de confidentialité ci-dessus est forte. Wasserman et Zhou (2010) proposent une interprétation de la confidentialité différentielle en terme de risque d'identification.

Cadre minimax privé. Duchi et al. (2018) ont expliqué comment étendre le cadre minimax classique pour prendre en compte des contraintes de confidentialité différentielle

locale. Nous présentons maintenant le risque minimax privé que nous allons étudier par la suite.

Soient X_i, $i = 1, \dots, n$ des vecteurs aléatoires de $\mathbb{R}^d$ indépendants et identiquement distribués selon la loi $\mathcal{N}(\theta, \sigma^2 I_d)$. Les vecteurs $X_i = (X_{i,j})_{j=1,\dots,d}$ pour $i = 1, \dots, n$ sont observés par n personnes distinctes qui refusent de partager leurs observations respectives. Le statisticien n'a pas accès à ces données mais seulement à des versions privatisées $Z_1, \dots Z_n$ de ces données. On suppose que θ appartient à l'ensemble suivant:

$$\Theta_d^+(s,a) = \{\theta \in \mathbb{R}^d : \text{il existe } S \subseteq \{1,\dots,d\} \text{ de cardinal au plus } s \text{ tel que } \theta_j \geq a \text{ pour tout } j \in S, \text{ et } \theta_j = 0 \text{ pour tout } j \notin S\}.$$

L'objectif est de sélectionner les coordonnées non nulles de θ, c'est-à-dire d'estimer le vecteur

$$\eta = \eta(P_\theta) = (I(\theta_j \neq 0))_{j=1,\dots,d},$$

où $I(\cdot)$ est la fonction indicatrice. On appellera *sélecteur* toute fonction mesurable $\hat{\eta} = \hat{\eta}(Z_1, \dots, Z_n)$ à valeurs dans $\{0,1\}^d$. Notre but est d'estimer le vecteur η par un sélecteur $\hat{\eta}$ appartenant à $\mathcal{Q}_\alpha$, l'ensemble des procédures vérifiant la contrainte de confidentialité différentielle locale de niveau α. Précisément, on s'intéresse ici aux mécanismes vérifiant la contrainte de confidentialité différentielle de niveau α qui transforment chaque X_i en un vecteur Z_i de $\mathbb{R}^d$. On dit qu'un sélecteur $\hat{\eta} = (\hat{\eta}_1, \dots, \hat{\eta}_d)$ est *séparable* si pour tout $j = 1, \dots, d$ sa $j$ème coordonnée $\hat{\eta}_j$ dépend seulement des variables $(Z_{i,j})_{i=1,\dots,n}$. On note $\mathcal{T}$ l'ensemble des sélecteurs séparables. Plus particulièrement, on supposera qu'il existe un mécanisme Q vérifiant la contrainte de confidentialité différentielle locale de niveau α/d tel que $Z_{i,j} \sim Q(\cdot \mid X_{i,j} = x)$ pour tout $i \in [\![1,n]\!]$ et tout $j \in [\![1,d]\!]$. On note P_θ la distribution des X_i et $Q^d P_\theta$ la distribution des Z_i. On s'intéresse au risque minimax privé suivant,

$$\inf_{Q \in \mathcal{Q}_{\alpha/d}} \inf_{\tilde{\eta} = \tilde{\eta}(Z^{(1)},\dots,Z^{(d)}) \in \mathcal{T}} \sup_{\theta \in \Theta_d^+(s,a)} \frac{1}{s} \mathbb{E}_{(Q^d P_\theta)^{\otimes n}} |\tilde{\eta}(Z^{(1)}, \dots, Z^{(n)})) - \eta|, \tag{1}$$

où $\mathcal{Q}_{\alpha/d}$ est l'ensemble des mécanismes vérifiant la contrainte de confidentialité différentielle locale de niveau α/d et

$$|\hat{\eta} - \eta| := \sum_{j=1}^d |\hat{\eta}_j - \eta_j| = \sum_{j=1}^d I(\hat{\eta}_j \neq \eta_j)$$

est la distance de Hamming entre $\hat{\eta}$ and η. On dit que l'*identification presque parfaite du support est possible* pour $(\Theta_d^+(s_d, a_d))_{d \geq 1}$ s'il existe un mécanisme $Q \in Q_{\alpha/d}$ et un sélecteur séparable $\hat{\eta}$ tel que

$$\lim_{d \to \infty} \sup_{\theta \in \Theta_d^+(s_d,a_d)} \frac{1}{s_d} \mathbb{E}_{(Q^d P_\theta)^{\otimes n}} |\hat{\eta} - \eta| = 0.$$

On dit que l'*identification presque parfaite du support est impossible* pour $(\Theta_d^+(s_d, a_d))_{d\geq 1}$ si

$$\liminf_{d\to\infty} \inf_{Q\in\mathcal{Q}_{\alpha/d}} \inf_{\hat{\eta}=\hat{\eta}(Z_1,\dots,Z_n)\in\mathcal{T}} \sup_{\theta\in\Theta_d^+(s_d,a_d)} \frac{1}{s_d}\mathbb{E}_{(Q^dP_\theta)^{\otimes n}}|\hat{\eta}-\eta| > 0.$$

Dans le cadre non confidentiel, Butucea et al. (2018) ont mis en évidence l'existence d'une valeur a_d^* telle que l'identification presque parfaite est possible pour a_d plus grand que a_d^* et impossible pour a_d plus petit que a_d^*. Nous montrons dans la suite que lorsqu'on ajoute des contraintes de confidentialité différentielle locale, l'identification presque parfaite du support devient impossible pour $n \leq d$ peu importe la valeur de a.

2 Bornes inférieures

Nous donnons dans cette partie une borne inférieure sur le risque minimax (1) et montrons que l'identification presque parfaite est impossible en grande dimension sous contrainte de confidentialité différentielle locale.

Théorème 2.1. *Pour tout $a > 0$, $\alpha > 0$, $1 \leq s \leq d$, $n \geq 1$, on a*

$$\inf_{\mathcal{Q}\in\mathcal{Q}_{\alpha/d}} \inf_{\tilde{\eta}=\tilde{\eta}(Z^{(1)},\dots,Z^{(n)})\in\mathcal{T}} \sup_{\theta\in\Theta_d^+(s,a)} \frac{1}{s}\mathbb{E}_{(Q^dP_\theta)^{\otimes n}}|\tilde{\eta}-\eta|$$
$$\geq \left(1-\frac{s}{d}\right)\exp\left(-4n(e^{\alpha/d}-1)^2\min\left\{\frac{a^2}{4\sigma^2},1\right\}\right).$$

Pour que la contrainte de confidentialité ait un sens en pratique, le paramètre de confidentialité α ne doit pas être trop grand. En particulier, $\alpha/d \to 0$ quand $d \to +\infty$. On a alors $n(e^{\alpha/d}-1)^2 \sim n\alpha^2/d^2$ et le théorème 2.1 montre donc que l'identification presque parfaite est impossible sous contrainte de confidentialité différentielle locale si la quantité $n\alpha^2/d^2$ est bornée supérieurement. En particulier, l'identification presque parfaite est impossible sous contrainte de confidentialité différentielle locale dans le cadre de la grande dimension, c'est-à-dire quand $n \leq d$, et ce, peu importe la valeur de a, ce qui contraste avec les résultats obtenus dans le cadre non confidentiel par Butucea et al. (2018). Le théorème 2.1 montre aussi que si $n\alpha^2/d^2 \to +\infty$, l'identification presque parfaite du support est impossible pour $a \leq C(\sigma d)/(\sqrt{n}\alpha)$.

La borne inférieure donnée par le théorème 2.1 est complétée dans la suite par une borne supérieure nous permettant d'obtenir des conditions suffisantes pour que l'identification presque parfaite soit possible.

3 Borne supérieure

Pour obtenir une borne supérieure sur le risque minimax (1) nous construisons un estimateur à partir de versions privatisées $Z_1, \dots Z_n$ des données X_i, $i = 1 \dots d$ et calculons

son risque.

Pour tout $i \in [\![1, n]\!]$ et $j \in [\![1, d]\!]$ définissons

$$Z_{i,j} = [X_{i,j}]_T + \frac{2Td}{\alpha} W_{i,j},$$

où les $W_{i,j}$'s sont des variables aléatoires indépendantes et identiquement distribuées selon la loi Laplace(1), $W_{i,j}$ est indépendante de $X_{i,j}$, et $[x]_T = \max\{-T, \min\{x, T\}\}$ désigne la projection de x sur l'intervalle $[-T, T]$ où T est un paramètre à choisir.

Proposition 3.1. *Pour tout $i \in [\![1, n]\!]$ et $j \in [\![1, d]\!]$, $Z_{i,j}$ est une version privatisée de $X_{i,j}$ vérifiant la contrainte de confidentialité locale de niveau α/d. Par conséquent, pour tout $i \in [\![1, n]\!]$ $Z_i = (Z_{i,j})_{j=1,\dots,d}$ est une version privatisée de X_i vérifiant la contrainte de confidentialité locale de niveau α.*

Définissons le sélecteur $\hat{\eta}$ ayant pour coordonnées

$$\hat{\eta}_j = I\left(\frac{1}{n}\sum_{i=1}^{n} Z_{i,j} \geq \tau\right), \quad j = 1, \dots, d, \tag{2}$$

où le paramètre τ est à choisir. Le résultat suivant donne une majoration du risque de ce sélecteur.

Théorème 3.2. *Si τ, T sont tels que $\tau < T$, $\tau\alpha/(8Td) < 1$ et $\alpha(T-\tau)/(4Td) < 1$, alors on a pour tout $\theta \in \Theta_d^+(s, a)$,*

$$\mathbb{E}\left[\frac{1}{s}|\hat{\eta} - \eta|\right] \leq \frac{d}{s}\left[\mathbb{P}\left(\xi \geq \frac{\tau\sqrt{n}}{2\sigma}\right) + \exp\left(-\frac{\tau^2 n\alpha^2}{128T^2d^2}\right) + n\mathbb{P}\left(|\xi| \geq \frac{T}{\sigma}\right)\right] + \exp\left(-\frac{(T-\tau)^2 n\alpha^2}{32T^2d^2}\right) + n\mathbb{P}\left(\xi \leq \frac{T-a}{\sigma}\right),$$

où ξ est une variable aléatoire gaussienne centrée réduite.

Ce résultat nous permet d'obtenir des conditions suffisantes pour que l'identification presque parfaite du support soit possible.

Corollaire 3.3. *Supposons $\alpha/d \to 0$, $n\alpha^2/d^2 \to +\infty$, $nd/s \to +\infty$ et $\limsup 2^9 \frac{\log(d/s)}{n\alpha^2/d^2} < 1$. Si $T = \sqrt{2\sigma^2(1+\delta)[\log(2n) + \log(d/s)]}$ pour un certain $\delta > 0$, alors le sélecteur $\hat{\eta}$ défini par (2) avec $\tau = T/2$ vérifie*

$$\sup_{\theta \in \Theta_d^+(s,a)} \frac{1}{s}\mathbb{E}_{(Q^d P_\theta)^{\otimes n}}|\hat{\eta}(Z^{(1)}, \dots, Z^{(d)})) - \eta| \to 0,$$

pour tout $a \geq \sqrt{2\sigma^2(1+\delta)}\left(\sqrt{\log n} + \sqrt{\log(2n) + \log(d/s)}\right)$.

Bibliographie

Butucea, C., Ndaoud, M., Stepanova, N.A. et Tsybakov, A.B. (2018). Variable selection with Hamming loss. *The Annals of Statistics*, 46(5):1837-1875.
Duchi, J. C., Jordan, M.I. et Wainwright, M.J. (2018). Minimax optimal procedures for locally private estimation, *Journal of American Statistical Association*, 113(521):182-201.
Wasserman, L, et Zhou, S. (2010). A statistical framework for differential privacy, *Journal of American Statistical Association* 105, 375-389.

Sparse Multiple Correspondence Analysis

Vincent Guillemot[1,*], Julie Le Borgne[1], Arnaud Gloaguen[2], Arthur Tenenhaus[2], Gilbert Saporta[3], Sylvie Chollet[4], Derek Beaton[5], Hervé Abdi[6,*]

[1] *Bioinformatics/Biostatistics Hub, CBD, Institut Pasteur, USR 3756 CNRS, Paris, FR*
[2] *Laboratoire des Signaux et Systèmes, CentraleSupelec, Gif-Sur-Yvette, FR*
[3] *CEDRIC, Conservatoire National des Arts et Métiers, Paris, FR*
[4] *Institut Supérieur d'Agriculture de Lille, Lille, FR*
[5] *The Rotman Institute at Baycrest, Toronto, Canada*
[6] *The University of Texas at Dallas, Richardson, TX, USA*
[*] *E-mail: vincent.guillemot@pasteur.fr, herve@utdallas.edu*

Résumé. L'Analyse des Correspondances Multiples (ACM) est la méthode de choix pour l'analyse des données catégorielles multivariées. Basée sur la décomposition en valeurs singulières (SVD), l'ACM bénéficie naturellement des extensions de cette dernière, dont celles qui permettent de réaliser des analyses parcimonieuses. L'algorithme permettant de réaliser l'ACM parcimonieuse nécessite deux propriétés particulières additionnelles: l'inclusion des matrices de métriques (masses et poids) caractéristiques de l'ACM, et la possibilité de sélectionner des groupes entiers de variables (un groupe étant constitué du codage disjonctif complet d'une variable catégorielle). Nous proposons un algorithme pour l'ACM parcimonieuse basé sur la décomposition en valeurs singulières généralisée et une projection sur la boule $\ell_{1,2}$. Nous illustrons notre méthode avec les résultats d'une enquête par questionnaires sur les connaissances alimentaires et la perception de fromages.

Mots-clés. Parcimonie, Analyse Multivariée, Analyse des Correspondances Multiples

Abstract. Multiple Correspondence Analysis (MCA) is the method of choice for the multivariate analysis of categorical data. In MCA each qualitative variable is represented by a group of binary variables (with a coding scheme called "complete disjunctive coding") and each binary variable has a weight inversely proportional to its frequency. The data matrix concatenates all these binary variables, and once normalized and centered this data matrix is analyzed with a generalized singular value decomposition (GSVD) that incorporates the variable weights as constraints (or "metric"). The GSVD is, of course, based on the plain SVD and so MCA can be sparsified by extending algorithms designed to sparsify the SVD. To do so requires two additional features: to include weights and to be able to sparsify entire groups of variables at once. Another important feature of such a sparsification should be to preserve the orthogonality of the components, Here, we integrate all these constraints by using an exact projection scheme onto the intersection of subspaces (i.e., balls) where each ball represents a specific type of constraints. We illustrate our procedure with the data from a questionnaire survey on the perception of cheese in two French cities.

Keywords. Sparsity, Multivariate Analysis, Multiple Correspondence Analysis

Akin to principal component analysis (PCA), Multiple Correspondence Analysis (MCA, see for reviews [1, 6, 8]) is a multivariate analysis method that analyzes the structure of a set of I observations described by K qualitative variables each comprising J_k modalities. In MCA, the response of an observation to a qualitative variable is represented by a binary vector. The I by J data matrix to analyze (denoted $\mathbf{Y}$), called the disjunctive table, is the concatenation of K matrices each with I observations and J_k columns. In MCA, the matrix actually analyzed is the centered probability matrix denoted $\mathbf{X}$, and computed as

$$\mathbf{X} = \mathbf{Y} \times (IK)^{-1} - \mathbf{r}\mathbf{c}^\top \text{ with } \mathbf{r} = \mathbf{Y}\mathbf{1} \times (IK)^{-1} \text{ and } \mathbf{c} = \mathbf{Y}^\top\mathbf{1} \times (IK)^{-1} \ . \quad (1)$$

MCA is obtained from the generalized singular value decomposition (GSVD) of $\mathbf{X}$ as

$$\mathbf{X} = \mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^\top \text{ with } \mathbf{P}^\top\mathbf{M}\mathbf{P} = \mathbf{Q}^\top\mathbf{W}\mathbf{Q} = \mathbf{I} \text{ where } \mathbf{M} = \text{diag}\{\mathbf{r}\} \text{ and } \mathbf{W} = \text{diag}\{\mathbf{c}\} \quad (2)$$

with the diag operator transforming a vector into a diagonal matrix. In MCA, the interpretation of the dimensions is greatly facilitated when variables have either a very large or a very small (rather than a medium) contribution to a given dimension. In standard PCA such a pattern is obtained by rotation of the dimensions (e.g., with VARIMAX) or by sparsification (see [5]). To extend sparsification to MCA (for previous work on Sparse MCA see [2, 7]), the procedure needs to rely on extensions of sparsification that can incorporate 1) the constraints imposed by the matrices $\mathbf{M}$ and $\mathbf{W}$ and 2) a group constraint which imposes that the entire block of columns representing a variable is is selected or discarded by the sparsification procedure. In addition, as for standard sparsification of PCA, the interpretation of the results is improved when the sparsified dimensions are pairwise orthogonal, but previous sparsification methods for MCA do not implement the orthogonality constraint and sparsify only single variables. In this paper we present a generalization of the sparsified SVD (the SGSVD) that implements all the aforementioned constraints. We first present the method as an optimization problem, derive the algorithm for this optimization problem, and illustrate this new method with the analysis of a questionnaire on food preference.

1 Method

The algorithm that we propose solves the following optimization problem

$$\underset{\mathbf{P},\boldsymbol{\Delta},\mathbf{Q}}{\arg\min} \tfrac{1}{2} \left\|\mathbf{X} - \mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^\top\right\|_2^2 \text{ with } \begin{cases} \mathbf{P}^\top\mathbf{M}\mathbf{P} = \mathbf{I} \\ \mathbf{Q}^\top\mathbf{W}\mathbf{Q} = \mathbf{I} \end{cases}, \text{ and } \forall \ell = 1, ..., R \begin{cases} \|\mathbf{p}_\ell\|_{\mathcal{G}} \leq c_{1,\ell} \\ \|\mathbf{q}_\ell\|_{\mathcal{G}} \leq c_{2,\ell} \end{cases} \quad (3)$$

where $\|.\|_2$ is the Euclidean norm, $c_{1,\ell}$ and $c_{2,\ell}$ are positive scalars controlling the sparsity constraint for the pseudo-singular vectors, and where the group norm is defined as: $\|\mathbf{x}\|_{\mathcal{G}} = \sum_{g=1}^{G} \|\mathbf{x}_{\iota_g}\|_2$. In other words, it is the ℓ_1-norm of the vector containing the ℓ_2-norm of the sub-vectors defined by the groups. The $\ell_{1,2}$-ball associated with this norm is noted $\mathscr{B}_{1,2}(\cdot)$.

Data: $\mathbf{X}$, ε, R
Result: Sparse MCA of $\mathbf{X}$
Define $\mathbf{P} = \mathbf{0}$;
Define $\mathbf{Q} = \mathbf{0}$;
Apply weights and masses to $\mathbf{X}$;
for $\ell = 1, \ldots, R$ **do**
  $\mathbf{p}^{(0)}$ and $\mathbf{q}^{(0)}$ are randomly initialized;
  $\delta^{(0)} \leftarrow 0$;
  $\delta^{(1)} \leftarrow \mathbf{p}^{(0)\top}\mathbf{X}\mathbf{q}^{(0)}$;
  $s \leftarrow 0$;
  while $\left|\delta^{(s+1)} - \delta^{(s)}\right| \geq \varepsilon$ **do**
    $\mathbf{p}^{(s+1)} \leftarrow \text{proj}(\mathbf{X}\mathbf{q}^{(s)}, \mathscr{B}_{1,2}(c_{1,\ell}) \cap \mathscr{B}_2(1) \cap \mathbf{P}^{\perp})$;
    $\mathbf{q}^{(s+1)} \leftarrow \text{proj}(\mathbf{X}^{\top}\mathbf{p}^{(s+1)}, \mathscr{B}_{1,2}(c_{2,\ell}) \cap \mathscr{B}_2(1) \cap \mathbf{Q}^{\perp})$;
    $\delta^{(s+1)} \leftarrow \mathbf{p}^{(s+1)\top}\mathbf{X}\mathbf{q}^{(s+1)}$;
    $s \leftarrow s + 1$;
  end
  $\delta_\ell \leftarrow \delta^{(s+1)}$;
  $\mathbf{P} \leftarrow \text{vec}\left(\mathbf{P}, \mathbf{p}^{(s+1)}\right)$;
  $\mathbf{Q} \leftarrow \text{vec}\left(\mathbf{Q}, \mathbf{q}^{(s+1)}\right)$;
end
Apply inverse weights and masses to $\mathbf{P}$ and $\mathbf{Q}$;

Algorithm 1: General algorithm of the sparse MCA.

1.1 The sparse MCA algorithm

The sparse MCA algorithm is presented in Algorithm 1. It is essentially an alternate projection algorithm. Its key component is the projection onto the intersection between the ball defined by the group constraint, the ℓ_2-ball, and the space orthogonal to the already estimated singular triplets (left or right).

To achieve the projections on $\mathscr{B}_{1,2}(c_{1,\ell}) \cap \mathscr{B}_2(1) \cap \mathbf{P}^{\perp}$ and $\mathscr{B}_{1,2}(c_{2,\ell}) \cap \mathscr{B}_2(1) \cap \mathbf{Q}^{\perp}$, we perform a Projection Onto Convex Sets (POCS) [3] with two components: the projection onto the intersection of the group ball and the ℓ_2-ball, and the projection onto the orthogonal spaces defined by the already estimated pseudo-singular vectors. We detail the first projection in the next section.

1.2 Projection onto the intersection of the group and the ℓ_2-balls

Here, we present a fast and exact algorithm for the projection of $\mathbf{x}$, a fixed vector of $\mathbb{R}^n$ that comprises K non-overlapping groups, onto the intersection of an $\ell_{1,2}$-ball of radius c and the ℓ_2-ball of radius 1. This generalizes the projection onto $\mathscr{B}_1(c) \cap \mathscr{B}_2(1)$ [4].

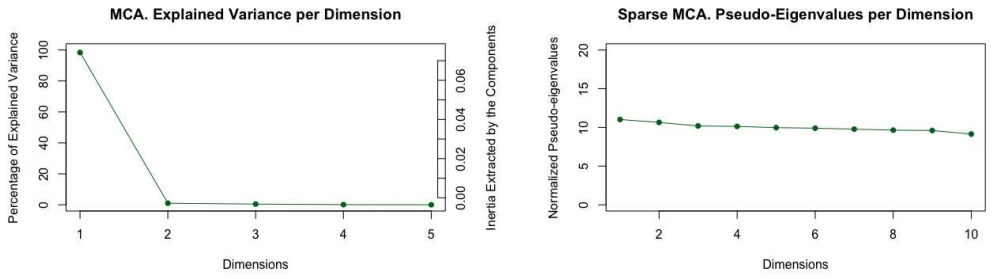
Denote by $\mathcal{G}$ the set of the indices defining the groups: $\mathcal{G} = \{\iota_k, k = 1, ..., K\}$, where ι_k indicates the variables contained in group k and K is the number of groups. Let $\mathbf{v}$ be the vector containing all the ℓ_2-norms of the sub-vectors $\mathbf{x}_{\iota_k}$, $k = 1, ..., K$. The real and positive value λ^* is such that $\| \operatorname{prox}_{\lambda^* \|\|\|_{\mathcal{G}}}(\mathbf{x})\|_{\mathcal{G}} = c$ if and only if $\| \operatorname{prox}_{\lambda^* \|\|\|_1}(\mathbf{v})\|_1 = c$, where prox_f is the proximal operator of a convex function f. Recall that the projection onto a ball is intimately linked to the proximal operator of the ball's norm. So, projecting $\mathbf{x}$ onto $\mathscr{B}_{1,2}(c) \cap \mathscr{B}_2(1)$ is equivalent to projecting $\mathbf{v}$ onto $\mathscr{B}_1(c) \cap \mathscr{B}_2(1)$, which can be achieved with the algorithm presented in [4].

2 Results

We analyzed the answers to two sets of questions of a survey on cheese answered by a sample of French participants from the two French cities of Angers and Lille. The 8 questions from the first set evaluate knowledge with answers coded as either correct or incorrect. The 23 questions from the second set evaluate the behaviors, opinions, or attitudes of the respondents toward cheese that are either farm-made or industrial. These questions are answered with a 4 point Likert scale (from 1 meaning "I totally agree" to 4 meaning "I totally disagree"). In addition we have information about the respondents: Sex, Age (coded in 4 categories), and the city where they live (Angers or Lille).

A regular MCA was applied to this data, as well as a sparse MCA with the constraints that each dimension be based on only one group of variables (i.e., a unique categorical variable) and only one group of observations (i.e., city of origin).

The Scree-plots on Figures 1a and 1b show that the first dimension of the regular MCA captures most of the variability of the data, whereas for the sparse MCA the pseudo-eigenvalues (normalized by the total inertia) are almost all equal.



(a) MCA scree-plot.

(b) Sparse MCA scree plot.

Figure 1: Scree-plots of the regular (left) and sparse (right) MCA.

The first two dimensions obtained with the regular MCA are shown in Figures 2a and 2b that confirm that city of origin is the main source of variability from this data set.

The results of the sparse MCA (see Figures 2c, 2d, 2e, and 2f) reveal that: Age is an important component on the structure of the Angers group, followed by "knowledge" and gender whereas the second city of origin (Lille) is associated with the fourth dimension.

3 Conclusion

We developed a sparse version of the MCA that incorporates into the GSVD, the group constraints imposed on the different modalities of a qualitative variable, and illustrated with real data that this new approach can simplify the interpretation of the factorial dimensions as well as reveal deeper insights. Future directions will include taking into account a hierarchical structure of either variables or observations such as overlapping groups of grouped variables as can be found, for example, for SNP data structured into pathways.

References

[1] H. Abdi and D. Valentin. Multiple Correspondence Analysis. In *Encyclopedia of Measurement and Statistics*. SAGE Publications, Inc., Thousand Oaks, 2007.

[2] A. Bernard, C. Guinot, and G. Saporta. Sparse principal component analysis for multiblock data and its extension to sparse multiple correspondence analysis. In *Proceedings of 20th International Conference on Computational Statistics (COMPSTAT 2012)*, pages 99–106, 2012.

[3] P. Combettes. The foundations of set theoretic estimation. *Proceedings of the IEEE*, 81(2):182–208, 1993.

[4] A. Gloaguen, V. Guillemot, and A. Tenenhaus. An efficient algorithm to satisfy ℓ_1 and ℓ_2 constraints. In *49èmes Journées de statistique*, Avignon, France, 2017.

[5] V. Guillemot, D. Beaton, A. Gloaguen, T. Löfstedt, B. Levine, N. Raymond, A. Tenenhaus, and H. Abdi. A constrained singular value decomposition method that integrates sparsity and orthogonality. *PLOS ONE*, 14(3):e0211463, mar 2019.

[6] B. Le Roux and H. Rouanet. *Multiple Correspondence Analysis*. Sage, Thousand Oaks, CA, 2010.

[7] Y. Mori, M. Kuroda, and N. Makino. Sparse Multiple Correspondence Analysis. In *Nonlinear Principal Component Analysis and Its Applications*, chapter 5, pages 47–56. Springer Singapore, 2016.

[8] G. Saporta. *Probabilités, Analyse des Données et Statistique*. Technip, Paris, France, 3rd edition, 2011.

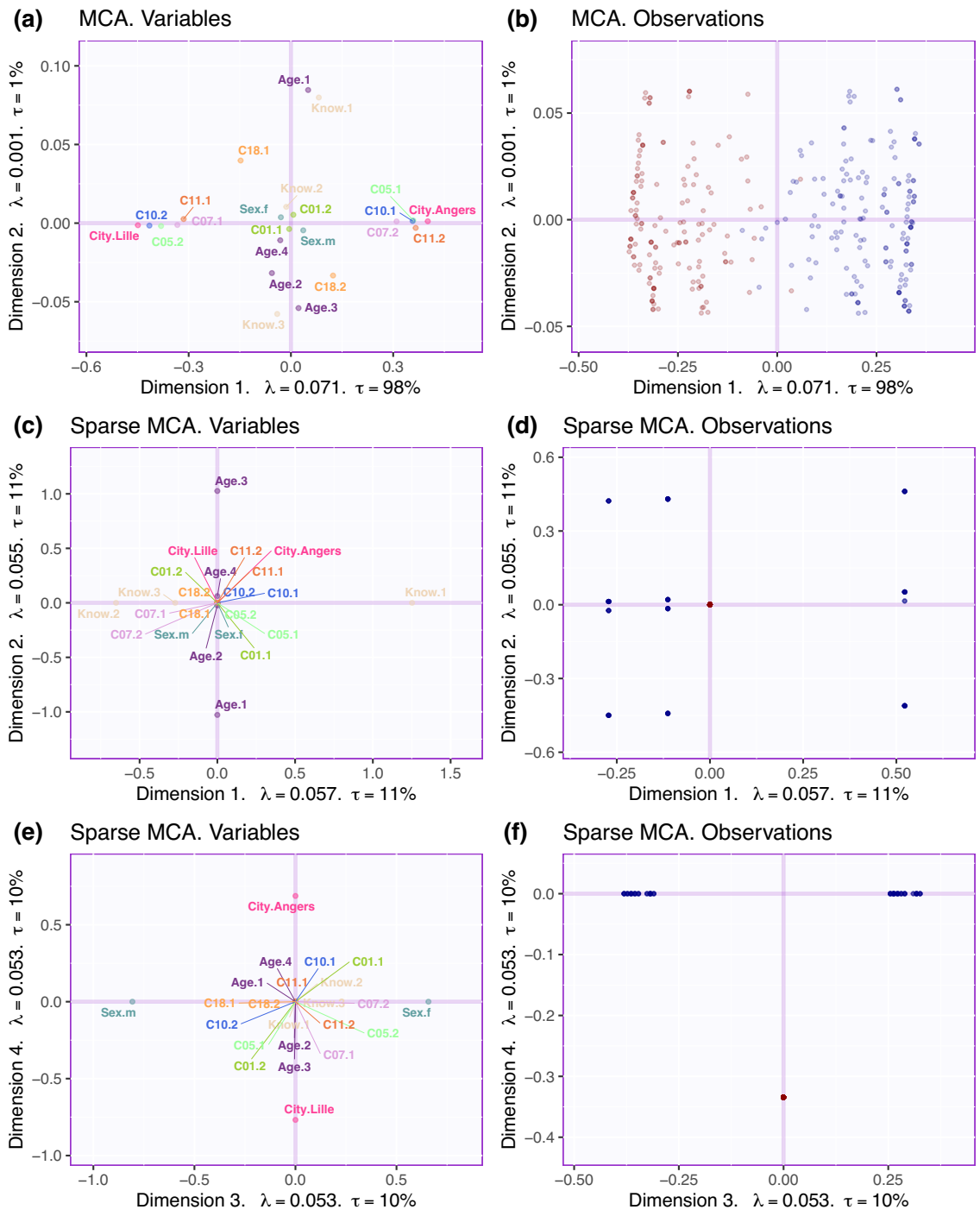


Figure 2: Variable and observation maps for MCA and sparse MCA. Only Dimensions 1 and 2 are shown for the regular MCA. For the sparse MCA, we show Dimensions 1 and 2 (middle figures) and Dimensions 3 and 4 (lower figures).

A median test for functional data

Zaineb Smida & Lionel Cucala & Ali Gannoun

Institut Montpelliérain Alexander Grothendieck, Université de Montpellier, France.

E-mail: zaineb.smida@umontpellier.fr ; lionel.cucala@umontpellier.fr ; ali.gannoun@umontpellier.fr

Résumé. Le test de la médiane est plus puissant que les tests de Student et de Wilcoxon-Mann-Whitney dans le cas des distributions à queues lourdes pour des données univariées. Pour les données multivariées de dimension finie, le test de signe est plus efficace que les tests de Hotelling et de Wilcoxon-Mann-Whitney lorsque les distributions sont aussi à queues lourdes et que l'espace est de grande dimension.
Dans ce travail, nous construisons un test de la médiane basé sur les rangs spatiaux pour des données fonctionnelles. Ensuite, nous le comparons avec le test de Wilcoxon-Mann-Whitney en utilisant des données fonctionnelles simulées.

Mots-clés. Données fonctionnelles, Dérivée au sens de Gâteaux, Espace de Banach, Espace de Hilbert séparable.

Abstract. The median test is more powerful than the Student and the Wilcoxon-Mann-Whitney tests in heavy-tails cases for univariate data. For finite multivariate data, the sign test is more efficient than the Hotelling and the Wilcoxon-Mann-Whitney tests for high dimensions and in very heavy-tailed cases.
In this work, we construct a median type test based on spatial ranks for functional data. Then, we compare it to the Wilcoxon-Mann-Whitney one using simulated functional data.

Keywords. Functional data, Gâteaux derivative, Separable Hilbert space, Smooth Banach space.

1 Introduction

Parametric and nonparametric statistical hypothesis testing play an essential role in statistics (Lehmann (1986) and Lehman and Romano (2005)). Here, we consider the nonparametric procedures to construct tests. These procedures are applicable in many cases where the data are not drawn from a population with a specific distribution. These type of tests can be used to verify that two or more datasets come from identical populations.
For univariate data, Wilcoxon (1945) and Mann and Whitney (1947) proposed nonparametric tests based on ranks. Each of them defined their own test statistic which lead to the same test named Wilcoxon-Mann-Whitney. Another test of hypothesis of the location

problem is assigned to Mood (1950) and it is called the median test. Another version of this test based on ranks (see, Capéraà and Cutsem (1988)) has been proposed by Hàjek, Ŝidák and Sen (1999). Nowadays, the median test is not often used, because it is less powerful than the Wilcoxon-Mann-Whitney test when applied to Gaussian distributions (Mood (1954)). However, this test is more efficient, when using symmetrical distributions with heavy-tails, than the Wilcoxon-Mann-Whitney one (Capéraà and Cutsem (1988)).
For multivariate data, several versions of the Hotelling, Wilcoxon-Mann-Whitney and median tests have been studied.
For functional data, the main difficulty is the infinite dimension of the space data like the Banach and the Hilbert spaces. Appropriate statistical tools are necessary to handle these type of data, for example to decide whether two samples of curves are issued from the same distribution. In this context, Horváth, Kokoszka, and Reeder (2013) proposed two test statistics for testing the equality of mean functions and one of them is the same as the Hotelling statistic in finite dimensioanl space.
In a nonparametric setting, Chakraborty and Chaudhuri (2015) proposed a Wilcoxon-Mann-Whitney test based on spatial ranks.
In the following, we propose a median test statistic based on spatial ranks in separable Banach space and especially in separable Hilbert space.

2 Construction of the test

2.1 The univariate case

Let X and Y be two $\mathbb{R}$-valued random variables. We consider $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ two random samples of X and Y with distribution functions F and F_μ respectively, such that $\forall x \in \mathbb{R}; F_\mu(x) = F(x - \mu)$. The constant μ is called *translation parameter*.
We want to test :

$$H_0 : \mu = 0 \quad \text{against} \quad H_1 : \mu \neq 0.$$

Now, we present two nonparametric tests which are curently used.

- **The Wilcoxon test**: It is a rank test which is defined by the test statistic

$$W = \frac{1}{n} \sum_{i=1}^{n} R_i.$$

- **The median test**: It is a rank test which is defined by the test statistic

$$M = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{R_i > 0\}}.$$

In both statistics, $R_i = 1 + (\sum_{j=1}^{m} \mathbb{1}_{\{Y_i > X_j\}} + \sum_{k=1}^{n} \mathbb{1}_{\{Y_i > Y_k\}} - \frac{N+1}{2})$ is the centered rank of Y_i when $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ are ordered together in the same sample of size $N = m + n$.

2.2 The functional case

Now, let X and Y be two independant random elements in a separable Banach space χ. We denote by χ^* its dual space, i.e., the space of the linear continuous functions on χ with values in $\mathbb{R}$, and χ^{**} its bidual space, i.e., the space of the linear continuous functions on χ^* with values in $\mathbb{R}$. We denote by $\|.\|_\chi$ (resp. $\|.\|_{\chi^*}$) a norm on χ (resp. on χ^*).
Then, we consider $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ independent random samples of X and Y from two probability measures P and Q on χ. We suppose that P and Q differ by a shift $\Delta \in \chi$.
We want to test :

$$H_0 : \Delta = 0 \quad \text{against} \quad H_1 : \Delta \neq 0.$$

To construct the Wilcoxon-Mann-Whitney test, Chakraborty and Chaudhuri (2015) assumed that the space χ is smooth, i.e., $\|.\|_\chi$ is Gâteaux differentiable at each $x \neq 0, x \in \chi$ with Gâteaux derivative called $SGN_x \in \chi^*$.

- **The existing Wilcoxon-Mann-Whitney test**: It is defined by the test statistic

$$T_{\text{WMW}} = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \text{SGN}_{\{Y_i - X_j\}}.$$

  - The T_{WMW} is an unbaised estimator of the spatial rank of Y which is equal to $E(\text{SGN}_{\{Y-X\}})$.
  - We reject the null hypothesis for large values of $\|T_{WMW}\|_{\chi^*}$.

  Remark: In particular case, when the space χ is assumed to be an Hilbert one, the Wilcoxon-Mann-Whitney test statistic becomes

$$T_{\text{WMW}} = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{Y_i - X_j}{\|Y_i - X_j\|_\chi}.$$

In the univariate case, the median test is more powerful than the Wilcoxon-Mann-Whitney one in heavy-tails cases. For these reasons, we decided to construct a median test in this infinite dimensional space χ. A hypothesis needed to construct it is that the space χ^* is smooth, i.e., $\|.\|_{\chi^*}$ is Gâteaux differentiable at each $f \neq 0, \in \chi^*$ with Gâteaux derivative denoted by $SGN_f^* \in \chi^{**}$.

- **The proposed median test** : It is defined by the test statistic

$$\mathrm{MED}_{\mathrm{fct}} = \frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{SGN}^{*}_{\frac{1}{m}\sum_{j=1}^{m}\mathbf{SGN}_{\{Y_i - X_j\}}}\right).$$

  - Under certain conditions, the test statistic proposed $\mathrm{MED}_{\mathrm{fct}}$ is an asymptotic unbiased estimator of $\mathrm{SGN}^{*}_{E(\mathbf{SGN}_{\{Y-X\}})}$ which is in the univariate case the direction of the median of the $(Y-X)$'s distribution from the origin. This result is obtained using the strong law of large numbers in such spaces.
  - We decided to reject the null hypothesis for large values of $\|\mathrm{MED}_{\mathrm{fct}}\|_{\chi^{**}}$.

  Remark: In particular case, when the space χ is assumed to be an Hilbert one, the proposed median test statistic can be rewrited as

$$\mathrm{MED}_{\mathrm{fct}} = \frac{1}{n}\sum_{i=1}^{n}\frac{\sum_{j=1}^{m}\frac{Y_i - X_j}{\|Y_i - X_j\|_{\chi}}}{\|\sum_{j=1}^{m}\frac{Y_i - X_j}{\|Y_i - X_j\|}\|_{\chi}}.$$

- **Rule of decision** : To derive the *p-value* of the tests $\mathrm{MED}_{\mathrm{fct}}$, we decided to use random permutations such as proposed by Chakraborty and Chaudhuri (2015) for the Wilcoxon-Mann-Whitney test.

2.3 Simulation study

In this section, we compare the proposed median statistic $\mathrm{MED}_{\mathrm{fct}}$ in the previous section to the Wilcoxon-Mann-Whitney statistic T_{WMW}.
We set $\chi = L^2[0,1]$ and we consider

$$X = \sum_{k=1}^{\infty} Z_k e_k,$$

where for all $k \geqslant 0$, $e_k = \sqrt{2}sin(t/\sigma_k)$ is an orthonormal basis of χ, $\sigma_k = ((k-0.5)\pi)^{-1}$ and Z_k's are independant random variables which correspond to the projection of X on the Karhunen-Loève basis. Then, we consider 5 cases : $Z_k/\sigma_k \backsim N(0,1)$, $Z_k/\sigma_k \backsim t(5)$, $Z_k/\sigma_k \backsim \mathcal{C}(0,1)$, $Z_k/\sigma_k \backsim \mathcal{D}exp(0,1)$ and $Z_k/\sigma_k \backsim \mathcal{L}(0,1)$.

We suppose that Y is distributed as $X + \Delta$ and, under the alternative hypotheses $H_1 : \Delta \neq 0$, we consider the case where $\Delta(t) = c$, $c > 0$ for all $t \in [0,1]$.
The power of the statistics is estimated using n_{sim} random simulations of (X, Y). Based on n_{perm} random permutations, the hypothesis H_0 is rejected if $p_{value} < \alpha$, where α is the significance level which is chosen equal to 0.05. We obtain the following power results:

- For $\Delta(t) = c$, $n_{\text{perm}} = 999$, $n_{\text{sim}} = 1000$ and $n = m = 10$:

Power	$N(0,1)$	$t(5)$	$\mathcal{C}(0,1)$	$\mathcal{D}exp(0,1)$	$\mathcal{L}(0,1)$
$c = 0.25$					
Power-WMW	0.127	0.113	0.065	0.098	0.094
Power-MED_{fct}	**0.134**	**0.119**	**0.067**	0.097	**0.098**
$c = 0.5$					
Power-WMW	0.469	0.364	**0.11**	0.319	0.184
Power-MED_{fct}	**0.469**	0.36	0.098	0.315	**0.184**
$c = 0.75$					
Power-WMW	0.862	0.729	0.04	0.68	0.388
Power-MED_{fct}	**0.864**	0.722	**0.046**	0.662	**0.391**

Table 1: The power results of MED_{fct} and WMW when $m = n = 10$.

As expected, the power of both tests increases when the difference between X and Y, the parameter c, increases. These two nonparametric tests behave very similarly, even if the median test is slightly more powerful for heavy-tailed distributions such as Laplace.

2.4 Application to real data

In this section, we have used three datasets to compare our test with the Wilcoxon-Mann-Whitney one. These datasets are those utilized by Chaudhuri and Chakraborty (2015) (for more details, see Ramsay and Silverman (2005) and Ferraty and Vieu (2006)).
The first one, named the coffee data, is available from `http://www.cs.ucr.edu/~eamonn/time_series_data/`. It contains the spectroscopy values for 14 samples of two differents types of coffee beans (Arabica and Robusta) taken at 286 wavelengths.
The second one is the Berkeley growth and is available in the R package "fda". It contains the heights of 39 boys and 54 girls measured at 31 time points from age 1 to 18.
The third one is named the spectrometry data and it can be found at `http://www.math.univ-toulouse.fr/staph/npfda`. It contains the spectrometric curves, recorded on 215 pieces of finely chopped meat, which corresponds to the absorbance measured at 100 wavelengths between 850 nm and 1050 nm. Moreover, we know whether the fat content of each meat unit is $\leqslant 20\%$ or $> 20\%$ thanks to an analytical chemical process.
For the coffee data, the *p-values*, based on the random permutations, of our test and the Wilcoxon-Mann-Whitney test allow us to reject the null hypothesis. However, in the article of Chaudhuri and Chakraborty (2015) the *p-value*, based on the asymptotic distribution, of the Wilcoxon-Mann-Whitney test is 0.072 which fails to reject H_0.
We suppose that the small size of this dataset ($n = m = 14$) doesn't allow the asymptotic results to be relevant in that case.
The *p-values* of the two tests for both of the two other datasets are 0 upto two decimal

places and it's exactly like the *p-values* obtained in Chaudhuri and Chakraborty (2015).

Bibliographie

Capéraà, Ph. and Cutsem, B.V. (1988). *Méthodes et modèles en statistiques non paramétrique. Exposé fondamental*. Presses de l'université Laval.

Chakraborty, A. and Chaudhuri, P. (2015). A Wilcoxon-Mann-Whitney type test for infinite-dimensional data. *Biometrika.***102,** 1, 239–246.

Ferraty, F. and Vieu, Ph. (2006). *Nonparametric Functional Data Analysis (Theory and practice).* Springer-Verlag, New York.

Hàjek, J., Šidák, Z. and Sen, K. (1999). *Theory of Rank Tests (Second edition).* Academic Press, United States of America.

Horváth, L., Kokoszka, P., and Reeder, R. (2013). Estimation of the mean of function time series and a two-sample problem. *Journal of the Royal Statistical Society. Series B.* **75,** Part 1, 103–122.

Lehmann, E.L. (1986). *Testing Statistical Hypotheses (Second edition).* Springer-Verlag, New York.

Lehmann, E.L and Romano, J.P. (2005). *Testing Statistical Hypotheses (Third edition).* Springer-Verlag, New York.

Mann, H.B., Whitney D.R. (1947). On a test of whether one of two random variables is stochastically larger than the order. *Ann. Math. Statist.* **18,** 50–60.

Mood, A.M. (1950). *Introduction to the Theory of Statistics.* McGraw-Hi1l series in probability and statistics, New York.

Mood, A.M. (1954). On the asymptotic efficiency of certain nonparametric two-Sample tests. *Ann. Math. Statist* **25,** 514–522.

Ramsay, J.O. and Silverman, B.W. (2005). *Functional Data Analysis (Second edition).* Springer-Verlag New York.

Wilcoxon, F. (1945). Individual comparaisons by ranking methods. *Biometrics.*, **1,** 80–83.