Contribution à l'apprentissage statistique à base de modèles génératifs pour données complexes.

Julien Jacques

Université Lille 1

Mémoire

présenté par

Julien JACQUES

pour l'obtention de

l'Habilitation à Diriger des Recherches

Spécialité : Mathématiques Appliquées

Contribution à l'apprentissage statistique à base de modèles génératifs pour données complexes

mémoire soutenu publiquement le 28 novembre 2012

JURY

Christophe	BIERNACKI	Professeur, Université Lille 1	Examinateur
Gilles	CELEUX	Directeur de Recherche, Inria	Président
Ali	GANNOUN	Professeur, Université Montpellier II	Rapporteur
Brendan	MURPHY	Professeur, University College Dublin, Ireland	Rapporteur
Jean-Michel	POGGI	Professeur, Université Paris-Sud	Rapporteur
Cristian	PREDA	Professeur, Université Lille 1	Examinateur
Jérôme	SARACCO	Professeur, Université Bordeaux 1	Examinateur

Habilitation préparée au sein
du laboratoire Paul Painlevé, UMR CNRS 8524, Université Lille 1
et de l'équipe MΘDAL, Inria Lille Nord-Europe

REMERCIEMENTS

Je tiens à remercier Ali Gannoun, Jean-Michel Poggi ainsi que Brendan Murphy pour avoir accepté d'examiner mes travaux d'habilitation à diriger des recherches et d'en être les rapporteurs. Je remercie également Gilles Celeux et Jérôme Saracco qui m'ont fait l'honneur de participer à mon jury, ainsi que Cristian Preda, également membre de mon jury et avec qui j'ai plaisir à travailler depuis plusieurs années.

Je tiens à adresser un remerciement tout particulier à Christophe Biernacki, que je connais depuis maintenant plus de dix années, qui m'a initié à la recherche et avec qui il est toujours très agréable et fructueux de travailler.

Je remercie également les équipes m'ayant accueilli depuis 2001, ainsi que l'ensemble des personnes que j'y ai rencontré : les projets IS2 et MISTIS d'Inria Rhône-Alpes, le Lab-SAD de l'Université Pierre Mendès France de Grenoble, le laboratoire de Mathématiques de Besançon, le laboratoire Paul Painlevé de l'Université Lille 1 ainsi que la récente équipe MΘDAL d'Inria Lille-Nord Europe. Un grand merci à mes collègues de l'équipe *proba-stat* du laboratoire Paul Painlevé et du département GIS de Polytech'Lille, qui m'ont permis de travailler depuis maintenant six années dans des conditions particulièrement agréables.

Enfin, j'adresse mes remerciements les plus sincères à mon épouse, Carole, pour son soutien tout au long de ces années de recherche.

TABLE DES MATIÈRES

1

INTRODUCTION

Ce mémoire synthétise les activités de recherche que j'ai menées depuis ma thèse de doctorat, soutenue en décembre 2005 à l'Université Joseph Fourier, et mon intégration au sein du laboratoire Paul Painlevé de l'Université Lille 1 en septembre 2006. Ces activités s'orientent autour de la thématique de l'apprentissage statistique des données complexes, abordée par le biais de modèles probabilistes paramétriques génératifs. Ces modèles génératifs, en plus d'être relativement compétitifs en terme de prédiction, disposent de plusieurs avantages : ils sont généralement riches en interprétabilité grâce à leurs paramètres, ils permettent d'évaluer un risque associé à une prédiction et ils permettent également d'utiliser un même modèle pour des problèmes d'apprentissage supervisé, semi-supervisé ou non supervisé.

Les différents types de données abordés, que je qualifierai de complexes, sont les suivants :

- les données **issues** de **populations différentes**,
- les données de **rang** et les données **ordinales**,
- les données **fonctionnelles**,
- les données de **grande dimension**.

Chaque type de données fera l'objet d'un chapitre de ce manuscrit.

L'apprentissage statistique des données issues de populations différentes, qualifié dans la suite d'apprentissage adaptatif (chapitre 2), s'intéresse au cas où la population sur laquelle on veut utiliser le modèle d'apprentissage statistique, à des fins de prédiction par exemple, est différente de celle ayant servi à construire et estimer ce modèle. Les différentes approches que nous avons développées pour répondre à cette question ont en commun l'introduction d'une modélisation paramétrique du lien entre les deux populations, permettant de transférer l'information disponible d'une population vers l'autre.

Les données de rang, définissant un classement d'objets selon un ordre de préférence (ex : classement d'activité sportives par ordre de préférence) et les données ordinales, qui sont des données qualitatives ayant des modalités ordonnées (ex : mentions au baccalauréat), sont étudiées dans le chapitre 3. Les modèles probabilistes proposés ont été construits en modélisant le processus de génération des données, considéré être un algorithme de tri pour les données de rang et un algorithme de recherche dans une table ordonnée pour les données ordinales. Ces modèles probabilistes ont ensuite été utilisés en classification de données de rang ou ordinales multivariées, sous la forme de modèle de mélange avec hypothèse d'indépendance conditionnelle. Précisons que seule la classification non supervisée est abordée, car le cadre supervisé s'en déduit simplement.

Pour les données fonctionnelles, où l'observation statistique consiste en une ou plusieurs courbes (ex : courbes de croissance d'enfants), le modèle a été construit à partir d'une troncature de la décomposition Karhunen-Loeve du processus observé (chapitre 4). Fort de ce modèle, nous avons pu proposer différents algorithmes de classification automa-

tique, dans le cas de courbes uniques mais également multivariées.
Les données de grande dimension, présentes lorsque le nombre de variables du problème dépasse celui des observations, ont été étudiées dans un cadre de régression (chapitre 5). Cette situation est de nos jours très fréquente, comme en témoigne le nombre de travaux actuels abordant le sujet. En effet, du fait des avancées technologiques dans des domaines comme la biologie, cette thématique de recherche est particulièrement d'actualité et nourrit bon nombre de recherches. Les travaux réalisés dans ce domaine sont le fruit de deux thèses de doctorat que je co-encadre : celle de Loïc Yengo, qui développe un modèle de régression regroupant ensemble les variables ayant un effet similaire, et celle de Julie Hamon, effectuant une sélection des variables pertinentes à l'aide d'algorithmes d'optimisation combinatoire. Ces deux thèses ont des objectifs applicatifs en génétique, humaine pour Loïc Yengo et animale pour Julie Hamon.

L'organisation de ce manuscrit est telle que chacun des chapitres que nous venons d'introduire, pouvant être lu de façon indépendante, comporte : une liste des publications associées, une introduction générale permettant de situer nos contributions, un résumé synthétique de celles-ci ainsi qu'une illustration applicative. Un chapitre supplémentaire (chapitre 6) est consacré aux différentes applications concrètes que j'ai pu traiter dans le cadre de contrats de collaboration avec des entreprises privées (PGxIS, Gènes Diffusion, Maïa Eolis) et des laboratoires de recherche d'autres disciplines (biologie, chimie). Enfin, les perspectives de recherches futures sont abordées dans la discussion finale concluant ce mémoire.

2

APPRENTISSAGE STATISTIQUE ADAPTATIF

Sommaire

PUBLICATIONS ASSOCIÉES À CE CHAPITRE

Revues avec comité de lecture

[R1] Bouveyron, C. and Jacques, J. *Adaptive mixtures of regressions : Improving predictive inference when population has changed*, Communications in Statistics – Simulation and Computation, en révision.

[R2] Bouveyron, C., Gaubert, P. and Jacques, J., *Adaptive models in regression for modeling and understanding evolving populations*, Case Studies in Business, Industry and Government Statistics, 4 (2), 83–92, 2011.

[R3] Bouveyron, C. and Jacques, J. *Adaptive linear models for regression : improving prediction when population has changed*, Pattern Recognition Letters, 31 (14), 2237–2247, 2010.

[R4] Jacques, J. and Biernacki, C. *Extension of model-based classification for binary data when training and test populations differ*, Journal of Applied Statistics, 37 (5), 749–766, 2010.

[R5] Jacques, J. and Biernacki, C. *Classement de données binaires lorsque les populations d'apprentissage et de test sont différentes*, Revue des Nouvelles Technologies de l'Information, RNTI-A-1 Data Mining et apprentissage, 794, 41–52, 2007.

Chapitre de livre

[L1] Beninel, F., Biernacki, C., Bouveyron, C., Jacques, J., Lourme, A. *Knowledge Transfer : Practices, Types and Challenges*, chapter Parametric link models for knowledge transfer in statistical learning, Nova Publishers, 2012.

Conférences internationales avec comité de lecture

[CI1] Jacques, J. and Biernacki, C. *Generalized discriminant rule for binary data when training and test populations differ on their descriptive parameters*. 17th International Conference on Computational Statistics (COMPSTAT'06), Rome, Italie, août 2006.

Conférences nationales

[CN1] Bouveyron C. and Jacques. J. *Adaptive mixtures of regressions : improving predictive inference when population has changed*. 4th conference on Computational Methods for Modelling and Learning in Social and Human Sciences (MASHS'10), Lille, France, juin 2010.

[CN2] Jacques, J. and Bouveyron C. *Modèles adaptatifs pour les mélanges de régressions*. 41èmes Journées de Statistique organisée par la Société Française de Statistique, Bordeaux, mai 2009.

[CN3] Jacques, J. *Apprentissage adaptatif en classification et régression*. Colloquim Statistique pour le traitement d'Images (STATIM08), Paris, France, janvier 2009. *Conférence invité*.

[CN4] Bouveyron C. and Jacques, J. *Adaptive linear models in regression for the modeling of housing market in different U.S. cities*. Computational Methods for Modelling and Learning in Social and Human Sciences (MASHS'08), Créteil, France, juin 2008.

[CN5] Bouveyron C. and Jacques, J. *Adaptive linear models for regression*. First joint meeting of the Statistical Society of Canada and the Société Française de Statistique, Ottawa, Canada, mai 2008.

[CN6] Jacques, J. and Biernacki, C. *Analyse discriminante généralisée : cas des données binaires avec modèles des classes latentes*. Première Rencontre des Jeunes Statisticiens, Aussois, août-septembre 2005.

[CN7] Jacques, J. and Biernacki, C. *Analyse discriminante généralisée : cas des données binaires avec modèles des classes latentes*. Colloque Data Mining et Apprentissage Statistique : applications en assurance, banque et marketing, Niort, mai 2005.

2.1 PROBLÉMATIQUE

L'apprentissage statistique [56] regroupe un ensemble de techniques permettant d'expliquer et de prédire un phénomène à partir d'observations de ce dernier. Ces techniques, fournissant généralement des outils d'aide à la décision, sont de plus en plus utilisées dans un grand nombre de domaines. En voici quelques exemples :

- aide au diagnostic en médecine : prédire un risque de récidive d'un cancer de la peau, en fonction du type de traitement utilisé ainsi que d'autres variables cliniques,
- ciblage client en marketing : définir les typologies de clients susceptibles de souscrire à un nouveau produit à partir d'un historique d'achat et de critères sociodémographiques,
- recherche de prédispositions génétiques : déterminer les facteurs génétiques responsables d'une intolérance à un traitement médical et ceux qui au contraire entraînent une meilleure efficacité du traitement.

Dans un problème typique d'apprentissage statistique, une variable réponse $y \in \mathcal{Y}$ doit être expliquée ou prédite à partir d'un ensemble de variables explicatives ou covariables $\boldsymbol{x} = (x_1, \ldots, x_p)' \in \mathcal{X}$. Les espaces $\mathcal{X}$ et $\mathcal{Y}$ peuvent être quantitatifs ou qualitatifs. Lorsque $\mathcal{Y}$ est un espace quantitatif, $\mathbb{R}$ par exemple, on parle généralement de problème de *régression*, tandis qu'on parle de *classification* lorsqu'il est qualitatif. Si dans une problématique de régression, il est nécessaire pour estimer un modèle liant la variable réponse y aux variables explicatives $\boldsymbol{x}$ de disposer d'observations conjointes de y et $\boldsymbol{x}$, ce n'est pas toujours le cas en classification. On différencie en effet la *classification supervisée*, encore appelée *analyse discriminante*, qui a pour objectif de déterminer une règle de classification à partir d'observations d'échantillons de chaque classe, de la *classification non supervisée*, appelée également *classification automatique* ou encore *clustering*, qui détermine des classes d'observations uniquement à partir de l'observation $\boldsymbol{x}$.

Que ce soit en régression ou en classification, une hypothèse primordiale qui est faite lorsque l'apprentissage statistique est utilisé dans un but de prédiction, est que la population statistique étudiée n'a pas évolué entre la phase de collecte des données et d'estimation du modèle et la phase de prédiction. On dit alors que la population *source*, dont est extraite la base d'apprentissage, et la population *cible*, pour laquelle on veut réaliser des prédictions, sont identiques. Malheureusement, cette hypothèse n'est pas toujours vérifiée, comme on peut le voir dans les exemples suivants :

- en médecine, le modèle de prédiction du risque de récidive peut être utilisé sur des patients atteints d'un carcinome épidermoïde (type de cancer de la peau assez rare), alors qu'il a été *appris* à partir de patients atteints de carcinomes basocellulaires (type plus fréquent),
- en marketing, la typologie des clients peut être réalisée à partir d'une base de clients alors que l'on souhaite l'utiliser pour démarcher des personnes non encore clientes,
- en biologie, des prédispositions génétiques peuvent avoir été identifiées sur une population de caucasiens alors que l'on cherche à distribuer le médicament sur un nouveau marché en Asie.

Une solution simple à ce problème serait alors d'oublier la population source, de recollecter des données sur la population cible, et de réapprendre un nouveau modèle. On comprend aisément que cette approche n'est pas à privilégier, surtout lorsque la collecte de nouvelles données coûte cher. Des techniques permettant de réutiliser l'information acquise au préalable sur la population source seraient alors les bienvenues. Ce besoin a été exprimé et défini pour la première fois lors du *workshop NIPS-95* intitulé *Learning to Learn*. Depuis, beaucoup de travaux ont été entrepris pour apporter des solutions à une problématique encore plus générale qui est celle de l'apprentissage statistique faisant intervenir des populations différentes. On trouve dans [85] une revue de ces techniques, classifiées en fonction des similitudes entre les populations source et cible, de la disponibilité de la variable réponse dans chacune de ces populations et de l'objectif de l'apprentissage (classification, régression ...). Il existe deux grands courants : les techniques d'apprentissage multitâches, dont l'objectif est d'estimer simultanément plusieurs modèles sur des populations différentes, et les techniques de transfert d'apprentissage, où l'objectif est de transférer l'information disponible sur une population source vers une population cible, ceci dans le but d'utiliser cette information pour améliorer l'apprentissage d'un modèle statistique sur cette population cible. C'est ce second courant qui nous intéresse dans ce chapitre. Les hypothèses suivantes précisent le cadre d'application des travaux qui vont être présentés dans la suite :

- la variable réponse ainsi que les variables explicatives sont identiques dans les deux populations, mais leurs distributions de probabilité peuvent différer,
- l'échantillon disponible dans la population source est de taille suffisante, tandis qu'il n'est que de taille réduite dans la population cible,
- dans un cadre de classification, la variable réponse pourra ne jamais avoir été observée dans la population cible.

Cette thématique de recherche est répertoriée par [85] sous le nom *transductive transfer learning*, terme introduit par [5] que nous traduirons (abusivement) dans ce document par *apprentissage statistique adaptatif*. Elle recoupe des travaux en classification autour du mot-clé *domain adaptation*, avec par exemple des travaux basés sur un modèle de mélange dont les deux composantes correspondent aux deux populations source et cible [33], ou encore d'autres s'appuyant sur des approches bayésiennes utilisant la population cible pour définir les lois a priori à utiliser [6, 68]. Toujours en classification, le mot-clé *sample selection biais* regroupe un ensemble de méthodes exploitant des techniques d'échantillonnage d'importance [41, 61, 108]. Ces techniques, comme celles en régression connues sous le mot-clé *covariate shift* [93, 95, 96, 97, 98], supposent que la distribution de la variable réponse y conditionnellement aux variables explicatives $\boldsymbol{x}$ reste inchangée.

Les travaux que nous présentons dans ce chapitre s'affranchissent de cette dernière hypothèse, et utilisent une approche différente, basée sur la définition d'un lien paramétrique entre les deux populations. L'utilisation d'un tel lien paramétrique est possible car nous concentrons nos études sur l'utilisation de modèles d'apprentissage statistique génératifs paramétriques. L'approche par modélisation paramétrique du lien entre population que nous développons a plusieurs avantages, dont celui d'être utilisable même lorsque la taille d'échantillon disponible pour la population cible est faible. De plus, l'interpréta-

tion des paramètres du lien fournit une caractérisation du lien entre les deux populations. Ces travaux, réalisés dans un cadre de classification et de régression, ont fait l'objet de quatre articles méthodologiques dont trois sont publiés [R3,R4,R5] et un actuellement en révision [R1], un article appliqué [R2], et un chapitre d'ouvrage [L1] sur le transfert de connaissance.

La suite de ce chapitre est organisée de la façon suivante. La section 2.2 s'intéresse au cas de la classification supervisée de données binaires, tandis que la section 2.3 est consacrée à la régression. Chaque section sera illustrée par des applications numériques montrant l'utilité et illustrant l'intérêt des modèles développés.

2.2 APPRENTISSAGE ADAPTATIF EN CLASSIFICATION

2.2.1 *Classification de données binaires à base de mélange de Bernoulli*

Nous nous intéressons dans cette section à la classification supervisée de données binaires : nous disposons, pour un ensemble de n individus, de l'observation de p variables explicatives $\boldsymbol{x}$ et d'une variable réponse y telle que $y = k$ indique que l'individu décrit par $\boldsymbol{x}$ appartient à la $k$ème classe ($k = 1, \ldots, K$). D'un point de vue probabiliste, le couple $(\boldsymbol{x}, y)$ est supposé être une réalisation du vecteur aléatoire $(\boldsymbol{X}, Y)$ où $\boldsymbol{X} = (X_1, \ldots, X_p)'$. Les n observations constituent un n-échantillon $\mathcal{S} = (\boldsymbol{x}_i, y_i)_{1 \leq i \leq n}$ de réalisations indépendantes et identiquement distribuées de $(\boldsymbol{X}, Y)$. L'objectif de la classification supervisée est alors de définir une règle de classification permettant de prédire y en se basant uniquement sur les covariables $\boldsymbol{x}$.

Si dans le cas de données continues l'hypothèse gaussienne est la plus souvent utilisée par les méthodes génératives de classification supervisée [78], les variables binaires sont habituellement supposées être des réalisations de variables aléatoires de loi de Bernoulli [78]. Conditionnellement à l'appartenance à la classe k ($Y = k$), chaque covariable X_j est supposée suivre une loi de Bernoulli de paramètre α_{kj} ($0 < \alpha_{kj} < 1$) :

$$X_j | Y = k \sim \mathcal{B}(\alpha_{kj}) \quad (j = 1, \ldots, p). \tag{1}$$

En utilisant l'hypothèse classique, pour les données qualitatives, des classes latentes [26, 40] supposant que conditionnellement à la classe d'appartenance, les p variables explicatives binaires sont indépendantes, la densité de probabilité de $\boldsymbol{X}$ conditionnellement à Y est :

$$f_k(\boldsymbol{x}; \boldsymbol{\alpha}_k) = \prod_{j=1}^{p} {\alpha_{kj}}^{x_j} (1 - \alpha_{kj})^{1 - x_j}, \tag{2}$$

où $\boldsymbol{\alpha}_k = (\alpha_{k1}, \ldots, \alpha_{kp})'$. La distribution marginale de $\boldsymbol{X}$ est alors un mélange de loi de Bernoulli

$$\boldsymbol{X} \sim f(\cdot; \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k f_k(\cdot; \boldsymbol{\alpha}_k),$$

où $(\pi_1, \ldots, \pi_K)$ sont les proportions du mélange vérifiant $\pi_k > 0$ et $\sum_{k=1}^{K} \pi_k = 1$, et où $\boldsymbol{\theta} = \{(\pi_k, \boldsymbol{\alpha}_k),\ k = 1, \ldots, K\}$ désigne l'ensemble des paramètres du modèle. Lorsque les coûts de mauvais classements sont supposés symétriques, la règle du *maximum a posteriori* (MAP) consiste à classer un nouvel individu $\boldsymbol{x}$ dans le groupe $\hat{y}$ maximisant la probabilité conditionnelle $t_{\hat{y}}(\boldsymbol{x}; \boldsymbol{\theta})$ d'appartenance au groupe :

$$\hat{y} = \underset{k \in \{1,\ldots,K\}}{\operatorname{argmax}}\, t_k(\boldsymbol{x}; \boldsymbol{\theta}), \tag{3}$$

où

$$t_k(\boldsymbol{x}; \boldsymbol{\theta}) = P(Y = k | \boldsymbol{X} = \boldsymbol{x}; \boldsymbol{\theta}) = \frac{\pi_k f_k(\boldsymbol{x}; \boldsymbol{\alpha}_k)}{f(\boldsymbol{x}; \boldsymbol{\theta})}.$$

L'estimation de cette règle de classement est généralement réalisée par maximum de vraisemblance, qui dans ce cas revient simplement à estimer les paramètres α_{kj} par les fréquences empiriques relatives :

$$\hat{\alpha}_{kj} = \frac{\operatorname{card}\{i : y_i = k, x_{ij} = 1\}}{n_k},$$

où $n_k = \operatorname{card}\{i : y_i = k\}$ est le nombre d'individus de l'échantillon d'apprentissage $\mathcal{S}$ appartenant à la classe k. Les proportions du mélange s'estiment également simplement par

$$\hat{\pi}_k = \frac{n_k}{n}.$$

2.2.2 *Modèles adaptatifs paramétriques pour la classification de données binaires*

Supposons désormais que les données consistent en deux échantillons : un premier échantillon étiqueté $\mathcal{S} = (\mathbf{x}, \mathbf{y})$, de taille n, provenant de la population source Ω, et un second $\mathcal{S}^* = (\mathbf{x}^*, \mathbf{y}^*)$, non étiqueté (i.e. $\mathbf{y}^*$ non observé), provenant de la population cible Ω^*. Insistons sur le fait que ce sont les mêmes variables qui sont mesurées dans chacune des deux populations, mais leurs distributions peuvent différer. Les modèles d'apprentissage statistique adaptatif que nous présentons ici, publiés dans [R4,R5], vont permettre de définir une règle de classification pour la population cible en adaptant celle de la population source, et ce bien qu'aucune information sur les étiquettes des observations de la population cible ne soit disponible. Une extension au cas où l'échantillon de la population cible est partiellement étiqueté est discutée dans [R4], mais ne sera pas présentée ici.

Les variables explicatives dans la population cible Ω^* sont supposées avoir la même distribution que dans (1) mais avec des paramètres potentiellement différents α^*_{kj} :

$$X^*_j | Y^* = k \sim \mathcal{B}(\alpha^*_{kj}).$$

DÉFINITION D'UNE FONCTION DE TRANSFERT. Afin d'atteindre l'objectif de l'apprentissage statistique adaptatif, nous avons défini un lien entre les distributions des variables

explicatives des populations source et cible. Si dans le cas gaussien un lien linéaire est non seulement intuitif mais également justifié [13], ce n'est pas le cas pour des variables binaires. L'idée que nous avons développée fut de voir les variables binaires comme une discrétisation de variables continues, gaussiennes, et d'utiliser le lien linéaire entre ces dernières pour définir le lien suivant entre les paramètres α^*_{kj} de Ω^* et α_{kj} de Ω :

$$\alpha^*_{kj} = \Phi\Big(\delta_{kj}\ \Phi^{-1}(\alpha_{kj}) + \lambda_j \gamma_{kj}\Big), \tag{4}$$

où Φ est la fonction de répartition de la loi $\mathcal{N}(0,1)$, $\delta_{kj} \in \mathbb{R}^+ \setminus \{0\}$, $\lambda_j \in \{-1,1\}$ et $\gamma_{kj} \in \mathbb{R}$. Cette transformation correspond à une transformation linéaire entre les fonctions *probit* des paramètres α_{kj} et α^*_{kj}.

Conditionnellement au fait que les paramètres α_{kj} soient connus (ils seront estimés en pratique), l'estimation des Kd paramètres continus α^*_{kj} est obtenue à partir des estimations des paramètres de lien δ_{kj}, γ_{kj} et λ_j entre Ω et Ω^* par *plug-in*. Or, le nombre de paramètres dont dépend la fonction de lien ($2Kp$) étant plus grand que le nombre de paramètres du modèle complet ($Kp-1$), le modèle est sur-paramétré. Il est donc nécessaire de réduire ce nombre de paramètres, et pour ce faire nous avons proposé des modèles contraints correspondant à des contraintes naturelles et interprétables sur le lien entre les populations Ω et Ω^*.

MODÈLES CONTRAINTS. Les paramètres de lien δ_{kj} ($1 \leq k \leq K$ et $1 \leq j \leq p$) vont successivement être contraints à être égaux à 1 (le modèle sera noté 1 par la suite), être indépendants de la classe et de la dimension (δ), être indépendants de la classe seulement (δ_k) ou de la dimension seulement (δ_j). De la même façon, γ_{kj} peuvent être contraints à être égaux à 0, γ, γ_k ou γ_j. Ainsi, 16 modèles sont définis et indexés suivant la notation données dans la table 1.

	0	γ	γ_j	γ_k
1	10	1γ	$1\gamma_j$	$1\gamma_k$
δ	10	$\delta\,\gamma$	$\delta\,\gamma_j$	$\delta\,\gamma_k$
δ_j	$\delta_j 0$	$\delta_j\,\gamma$	$\delta_j\,\gamma_j$	$\delta_j\,\gamma_k$
δ_k	$\delta_k 0$	$\delta_k\,\gamma$	$\delta_k\,\gamma_j$	$\delta_k\,\gamma_k$

TAB. 1 – . Modèles adaptatifs paramétriques contraints pour la classification de données binaires.

Pour chacun de ces 16 modèles, il faut ajouter une hypothèse supplémentaire sur la conservation des proportions entre Ω et Ω^* : les modèles avec proportions conservées seront par exemple notés $\delta_k\,\gamma_k$ tandis que le modèle équivalent avec proportion libre sera noté $\pi\,\delta_k\,\gamma_k$. Le nombre de modèles contraints croît ainsi à 32. Nous avons montré dans [R4] que, bien que certains modèles ne soient pas identifiables, ces situations singulières de non-identifiabilité ne sont généralement pas rencontrées en pratique.

ESTIMATION ET CHOIX DE MODÈLE. L'objectif est d'estimer la règle de classement sur la population cible Ω^* :

$$\hat{y}^* = \underset{k\in\{1,\dots,K\}}{\operatorname{argmax}}\, t_k(\boldsymbol{x}^*;\boldsymbol{\theta}^*), \tag{5}$$

qui dépend des paramètres $\boldsymbol{\theta}^* = \{(\pi_k^*, \boldsymbol{\alpha}_k^*), k = 1,\dots,K\}$ où $\boldsymbol{\alpha}_k^* = (\alpha_{k1}^*,\dots,\alpha_{kp}^*)'$. Sous l'hypothèse du lien (4) entre les populations source et cible, contraint par les modèles définis au paragraphe précédent, l'estimation de la règle de classement revient à estimer les paramètres $\boldsymbol{\theta}$ du mélange de Bernoulli au sein de la population source ainsi que les paramètres de lien $\boldsymbol{\zeta} = \{(\delta_{kj}, \gamma_{kj}, \lambda_j), k = 1,\dots,K, 1 \leq j \leq p\}$. Deux approches sont alors possibles. La première consiste à estimer séquentiellement $\boldsymbol{\theta}$ puis $\boldsymbol{\zeta}$ et en déduire $\boldsymbol{\theta}^*$ par plug-in. La seconde consiste à estimer conjointement les deux types de paramètres. Lourme et Biernacki [72] ont montré que les deux approches étaient équivalentes lorsque la taille n de l'échantillon disponible pour la population source n'était pas trop petite.
Nous donnons quelques précisions sur la première approche, qui sera utilisée pour les expérimentations numériques présentées dans le paragraphe suivant. L'estimation de $\boldsymbol{\theta}$ par maximum de vraisemblance a été décrite dans la section 2.2.1. L'estimation des paramètres de lien est quant à elle moins directe étant donné que l'échantillon $\mathcal{S}^*$ est non supervisé. Pour surmonter ceci, on a recours à l'algorithme EM [37], connu pour sa simplicité et ses bonnes propriétés théoriques (voir [75] pour une revue détaillée). L'algorithme EM, partant du principe qu'il serait plus facile de maximiser la vraisemblance du modèle si l'échantillon $\mathcal{S}^*$ était supervisé, consiste à maximiser ce qu'on appelle communément la log-vraisemblance complétée, qui s'écrit ici :

$$l_c(\boldsymbol{\zeta};\mathbf{x}^*,\mathbf{y}^*) = \sum_{i=1}^{n^*}\sum_{k=1}^{K} \mathbb{1}_{y_i^*=k} \log\left(\pi_k^* \prod_{j=1}^{p} {\alpha_{kj}^*}^{x_{ij}^*}(1-\alpha_{kj}^*)^{(1-x_{ij}^*)}\right).$$

où α_{kj}^* est relié à $\boldsymbol{\zeta}$ par (4), et où $\mathbb{1}_{y_i^*=k}$ est la fonction indicatrice valant 1 si $y_i^* = k$ et 0 sinon. Les étiquettes $\mathbf{y}^*$ n'étant pas observées, l'algorithme EM itère successivement les étapes E et M suivantes jusqu'à convergence de la vraisemblance :

- étape E : calcul de l'espérance de la log-vraisemblance complétée conditionnellement aux observations

$$\mathcal{Q}(\boldsymbol{\zeta};\boldsymbol{\zeta}^{(q)}) = E_{\boldsymbol{\zeta}^{(q)}}[l_c(\boldsymbol{\zeta};\mathbf{x}^*,\mathbf{y}^*)|\mathbf{x}^*] = \sum_{i=1}^{n^*}\sum_{k=1}^{K} t_{ik}^q \log\left(\pi_k^* \prod_{j=1}^{p} {\alpha_{kj}^*}^{x_{ij}^*}(1-\alpha_{kj}^*)^{(1-x_{ij}^*)}\right)$$

où

$$t_{ik}^q = P(y_i^* = k|\mathbf{x}^*) = \frac{p_k^{*(q)} \prod_{j=1}^{d} \left(\alpha_{kj}^{*(q)}\right)^{x_{ij}^*}(1-\alpha_{kj}^{*(q)})^{(1-x_{ij}^*)}}{\sum_{\kappa=1}^{K} p_\kappa^{*(q)} \prod_{j=1}^{d} \left(\alpha_{\kappa j}^{*(q)}\right)^{x_{ij}^*}(1-\alpha_{\kappa j}^{*(q)})^{(1-x_{ij}^*)}},$$

– étape M : calcul des paramètres $\zeta^{(q+1)}$ maximisant $\mathcal{Q}$

$$\zeta^{(q+1)} = \underset{\zeta \in \Theta}{\operatorname{argmax}}\, \mathcal{Q}(\zeta; \zeta^{(q)}) \tag{6}$$

où Θ est l'espace des paramètres spécifique au modèle contraint considéré.
Une fois les 32 modèles contraints estimés, il est nécessaire de fournir un critère permettant de choisir parmi ces modèles. Nous utiliserons classiquement le critère BIC [91]) :

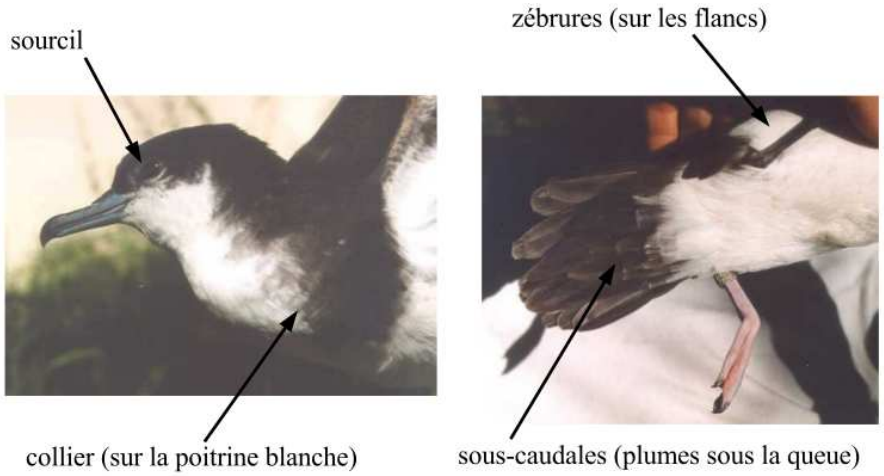
$$\text{BIC} = -2 \ln \ell + \nu \ln n^* \tag{7}$$

où ℓ est la valeur du maximum de vraisemblance et ν est le nombre de paramètres continus estimés.

2.2.3 *Une application en biologie*

Afin d'illustrer l'utilité de ces modèles adaptatifs paramétriques pour la classification de données binaires, nous considérons un exemple en biologie consistant à prédire le sexe d'oiseaux à partir de variable morphologiques. En effet, suivant les espèces, le sexe n'est pas toujours identifiable aisément, et nécessite parfois même une dissection de l'animal.
L'espèce d'oiseaux de mer *puffins* étudiée est composée de plusieurs sous-espèces vivant dans des régions différentes du globe, que nous regroupons ici en deux groupes :

– les oiseaux vivant dans les îles du Pacifique, appartenant aux sous-espèces suivantes : *subalaris* (îles Galapagos), *polynesial*, *dichrous* (îles Enderbury et Palau) et *gunax*. Pour ces oiseaux du Pacifique, nous disposons d'un échantillon de $n = 171$ individus, dont cinq caractéristiques morphologiques (dont le sexe) ont été mesurées à partir de squelettes présents dans différents musées [22] : absence ou présence de collier, de zébrures, de liseret et coloration unie ou non de la partie sous-caudale de l'oiseau. Ces caractéristiques (mis à part le liseret) sont illustrées par les photographies ci-dessous :



Cette population d'oiseaux, pour laquelle nous disposons d'un échantillon de taille confortable, sera utilisée comme population source, et servira à prédire le sexe des oiseaux du groupe suivant.

– les oiseaux vivant dans les îles de l'Atlantique : *boydi* (îles du Cap Vert), dont nous disposons un échantillon de $n = 19$ oiseaux pour lesquels les mêmes caractéristiques que précédemment ont été mesurées. Le sexe de ces oiseaux est également connu, et nous utiliserons cette information pour valider la qualité de nos prédictions.

Suivant le biologiste qui nous a fourni les données, ces variables morphologiques devraient être assez peu discriminantes du sexe des oiseaux, et il s'attend à des taux d'erreurs de l'ordre de $40 - 45\%$. Les 32 modèles contraints d'apprentissage adaptatif sont utilisés, ainsi que l'analyse discriminante standard (correspondant au modèle 10, et qui consiste à appliquer la règle de classification apprise sur les oiseaux du Pacifique directement sur les oiseaux des îles de l'Atlantique) et la classification automatique (l'information sur les oiseaux du Pacifique n'est alors pas utilisée). L'analyse discriminante et la classification automatique sont réalisées sur la base du modèle de mélange de Bernoulli.

model	10	1γ	$1\gamma_k$	$1\gamma_j$	$\delta 0$	$\delta\,\gamma$	$\delta\,\gamma_k$	$\delta\,\gamma_j$
error	50.94	**43.39**	45.28	43.39	50.94	**43.39**	45.28	45.28
BIC	212	**209**	216	224	212	**209**	216	224
model	$\delta_k 0$	$\delta_k\,\gamma$	$\delta_k\,\gamma_k$	$\delta_k\,\gamma_j$	$\delta_j 0$	$\delta_j\,\gamma$	$\delta_j\,\gamma_k$	$\delta_j\,\gamma_j$
error	45.28	45.28	52.83	45.28	45.28	52.83	50.94	50.94
BIC	210	210	215	226	225	224	227	239
model	$\pi 10$	$\pi 1\gamma$	$\pi 1\gamma_k$	$\pi 1\gamma_j$	$\pi\,\delta 0$	$\pi\,\delta\,\gamma$	$\pi\,\delta\,\gamma_k$	$\pi\,\delta\,\gamma_j$
error	45.28	50.94	50.94	45.28	45.28	50.94	50.94	45.28
BIC	213	213	220	228	213	213	220	228
model	$\pi\,\delta_k 0$	$\pi\,\delta_k\,\gamma$	$\pi\,\delta_k\,\gamma_k$	$\pi\,\delta_k\,\gamma_j$	$\pi\,\delta_j 0$	$\pi\,\delta_j\,\gamma$	$\pi\,\delta_j\,\gamma_k$	$\pi\,\delta_j\,\gamma_j$
error	45.28	45.28	47.16	45.28	45.28	52.83	45.28	52.83
BIC	214	213	213	229	228	227	224	243

Tab. 2 – Taux d'erreur de classification (%) et valeur du critère BIC pour la population cible d'oiseaux de l'Atlantique avec comme population source les oiseaux du Pacifique.

Le meilleur modèle contraint d'apprentissage adaptatif, qui est d'ailleurs celui sélectionné par le critère BIC, donne un taux d'erreur (43%) meilleur que l'analyse discriminante standard (50.94%) ou la classification automatique (49.05%) pour classer les oiseaux des îles du Cap Vert en fonction de leur sexe. Même si la qualité des classifications est assez pauvre, en accord avec les attentes du biologiste, cette application illustre l'apport de l'utilisation de l'échantillon des oiseaux du Pacifique pour classer ceux de l'Atlantique.

2.3 apprentissage adaptatif en régression

Nous venons de voir, dans le cas de la classification de données binaires, l'intérêt de l'apprentissage adaptatif pour transférer l'information d'une population vers une autre. Dans cette section nous présentons des outils développés dans un but similaire, mais

pour un objectif de régression. Les cas de la régression classique [R3] puis des mélanges de régressions [R1] seront abordés.

2.3.1 *Modèles adaptatifs paramétriques pour la régression linéaire*

Nous considérons le modèle de régression linéaire sur une base de fonctions, qui suppose que la variable réponse $Y \in \mathbb{R}$ peut être liée aux variables explicatives $\boldsymbol{x} \in \mathbb{R}^p$ par la relation :

$$Y = \sum_{j=0}^{d} \beta_j \psi_j(\boldsymbol{x}) + \epsilon,$$

où les résidus $\epsilon \sim \mathcal{N}(0, \sigma^2)$ sont supposés indépendants, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_d)' \in \mathbb{R}^{d+1}$ sont les paramètres de régression, $\psi_0(\boldsymbol{x}) = 1$ et $(\psi_j)_{1 \leq j \leq d} : \mathbb{R}^p \to \mathbb{R}$ est une base de fonctions (polynomiale, splines, trigonométriques...). Le cas de la régression linéaire classique correspond à $p = d$ et $\psi_j(\boldsymbol{x}) = x_j$ pour $j = 1, \ldots, d$.
Ce modèle revient à supposer que la distribution conditionnelle de Y est :

$$Y|\boldsymbol{X} = \boldsymbol{x} \sim \mathcal{N}(g(\boldsymbol{x}, \boldsymbol{\beta}), \sigma^2),$$

où la fonction de régression $g(\boldsymbol{x}, \boldsymbol{\beta}) = \boldsymbol{\beta}' \Psi(\boldsymbol{x}^*)$, avec $\psi_j(\boldsymbol{x}) = (1, \psi_1(\boldsymbol{x}), \ldots, \psi_d(\boldsymbol{x}))$, est définie comme l'espérance conditionnelle $E[Y|\boldsymbol{x}]$.

L'estimation des paramètres peut être faite par maximum de vraisemblance classique :

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{\Psi}^t \boldsymbol{\Psi})^{-1} \boldsymbol{\Psi}^t \mathbf{y} \qquad \text{et} \qquad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \boldsymbol{\beta}^t \Psi(\boldsymbol{x}_i) \right)^2,$$

où $\boldsymbol{\Psi}$ est la matrice de taille $(n) \times (d+1)$ contenant les lignes $\psi_j(\boldsymbol{x}_i)$ $(1 \leq i \leq n)$.

Modèles paramétriques de lien entre population

Dans le cadre de l'apprentissage statistique adaptatif, nous supposons que le modèle de régression a été estimé sur la population source Ω à partir d'un échantillon $\mathcal{S}$. L'objectif est d'adapter ce modèle à la population cible Ω^*, au sein duquel le modèle de régression s'écrit

$$Y^*|\boldsymbol{X}^* = \boldsymbol{x}^* \sim \mathcal{N}(\boldsymbol{\beta}^{*\prime} \Psi(\boldsymbol{x}^*), \sigma^{*2}). \tag{8}$$

Nous rappelons que suivant les hypothèses que nous avons présentées dans la problématique de ce chapitre, les variables $(\boldsymbol{X}, Y)$ et $(\boldsymbol{X}^*, Y^*)$ ont la même signification mais peuvent avoir des distributions de probabilités différentes, et l'échantillon $\mathcal{S}^* = (\mathbf{x}^*, \mathbf{y}^*)$ dont nous disposons pour Ω^* est de taille petite (en particulier $n^* << n$).

Nous avons proposé dans [R3] un modèle supposant que la transformation entre les populations Ω et Ω^* s'exprime au travers d'un lien entre les paramètres de régression :

$$\boldsymbol{\beta}^* = \boldsymbol{\Lambda} \boldsymbol{\beta},$$

où $\mathbf{\Lambda}$ est une matrice $(p+1) \times (p+1)$. Le modèle ainsi défini étant sur-paramétré, nous devons contraindre ce modèle en supposant la matrice $\mathbf{\Lambda}$ diagonale, ce qui conduit au modèle de transformation suivant

$$\beta_j^* = \lambda_j \beta_j \qquad \forall j = 0, \ldots, d. \tag{9}$$

Afin de proposer des modèles encore plus parcimonieux, correspondant à des situations concrètes, nous avons proposé des contraintes supplémentaires sur les paramètres λ_j, résumées dans le tableau 3. Par exemple, le modèle M_5 suppose que l'intercepte du modèle de régression de Ω^* est identique à celui de Ω tandis que les autres paramètres de régression sont proportionnels à ceux du modèle de la population source.

modèle	M_0	M_1	M_2	M_3	M_4	M_5	M_6
β_0^* est supposé être	$\lambda_0\beta_0$	β_0	$\lambda_0\beta_0$	$\lambda\beta_0$	β_0	$\lambda_0\beta_0$	β_0
β_i^* est supposé être	$\lambda_i\beta_i$	$\lambda_i\beta_i$	$\lambda\beta_i$	$\lambda\beta_i$	$\lambda\beta_i$	β_i	β_i
Nb. de paramètres	d+1	d	2	1	1	1	0

TAB. 3 – Une famille de modèles paramétriques de lien pour la régression linéaire sur base de fonctions.

D'autres modèles de lien peuvent être définis en utilisant une information a priori sur le phénomène modélisé pour contraindre la matrice de transformation $\mathbf{\Lambda}$. Ainsi, des modèles de transformation propres au phénomène étudié peuvent être proposés.

Estimation des paramètres du modèle de lien

Nous donnons ici une expression de l'estimateur des paramètres du modèle de lien commune à chaque modèle. Pour cela, notons γ_j les indices associés aux q ($q \leq d+1$) paramètres de régression qui ont changé entre les deux populations (i.e. $\beta^*_{\gamma_j} = \lambda_{\gamma_j}\beta_{\gamma_j}$, avec $1 \leq \gamma_j \leq d+1$ et $1 \leq j \leq q$), et $\bar{\gamma}_j$ ceux associés aux paramètres n'ayant pas changé ($1 \leq j \leq d+1-q$). Avec ces notations le modèle (8) peut s'écrire

$$Y = \boldsymbol{Q}\mathbf{\Lambda}_q + \bar{\boldsymbol{Q}}\mathbf{1}_{p-q} + \epsilon,$$

où

$$\mathbf{\Lambda}_q = \left(\lambda_{\gamma_1}, \ldots, \lambda_{\gamma_q}\right)^t, \qquad \mathbf{1}_{p-q} \text{ est le vecteur unité de taille } p-q,$$

$$\boldsymbol{Q} = \begin{pmatrix} \beta_{\gamma_1}\psi_{\gamma_1}(x_1) & \cdots & \beta_{\gamma_q}\psi_{\gamma_q}(x_1) \\ \vdots & & \vdots \\ \beta_{\gamma_1}\psi_{\gamma_1}(x_n) & \cdots & \beta_{\gamma_q}\psi_{\gamma_q}(x_n) \end{pmatrix}, \qquad \bar{\boldsymbol{Q}} = \begin{pmatrix} \beta_{\bar{\gamma}_1}\psi_{\bar{\gamma}_1}(x_1) & \cdots & \beta_{\bar{\gamma}_q}\psi_{\bar{\gamma}_q}(x_1) \\ \vdots & & \vdots \\ \beta_{\bar{\gamma}_1}\psi_{\bar{\gamma}_1}(x_n) & \cdots & \beta_{\bar{\gamma}_q}\psi_{\bar{\gamma}_q}(x_n) \end{pmatrix}.$$

L'estimation par maximum de vraisemblance (*Ordinary Least Square*, OLS) de $\mathbf{\Lambda}_q$ est alors

$$\hat{\mathbf{\Lambda}}_q^{OLS} = \left(\boldsymbol{Q}^t\boldsymbol{Q}\right)^{-1}\boldsymbol{Q}^t\left(\mathbf{y} - \bar{\boldsymbol{Q}}\mathbf{1}_{p-q}\right).$$

Choix du modèle de lien

Afin de choisir le modèle de lien le plus adapté, nous proposons d'utiliser trois critères classiques. Le premier critère est le critère PRESS [4], qui estime l'erreur quadratique moyenne par validation croisée *leave-one-out* :

$$PRESS = \frac{1}{n^*} \sum_{i=1}^{n^*} ||y_i^* - \widehat{y_i^*}^{-i}||^2$$

où $\widehat{y_i^*}^{-i}$ est la prédiction de y_i^* obtenue sans utiliser la i-ème observation de l'échantillon $\mathcal{S}^*$. Ce critère est certainement le plus utilisé en sélection de modèle pour la régression, et nous conseillons son utilisation lorsque cela est numériquement réalisable. En effet l'estimation par validation croisée peut s'avérer être lourde lorsque l'estimation des paramètres du modèle n'est pas très rapide. Les deux autres critères de type vraisemblance pénalisée, AIC[3] et BIC[91], ne souffrent pas de ce problème :

$$BIC = -2 \ln \ell + \nu \ln n^* \qquad \text{et} \qquad AIC = -2 \ln \ell + 2\nu,$$

où ℓ est la valeur du maximum de vraisemblance et ν le nombre de paramètres continus estimés (donné dans le tableau 3).

2.3.2 *Modèles adaptatifs paramétriques pour les mélanges de régressions*

Dans un certain nombre de domaines comme en économie, il arrive que les phénomènes étudiés aient plusieurs états et qu'un modèle de régression classique (linéaire ou non), ne soit pas adapté. Dans l'exemple économique traité à la fin de ce chapitre, où le lien entre les émissions de CO_2 et le PIB (rapportés au nombre d'habitants) est étudié pour 111 pays, deux états sont présents correspondant à deux stratégies économiques différentes : l'une payant la croissance économique par des émissions de CO_2 plus importantes que l'autre. La figure 1 illustre cette situation à l'aide d'une simulation basique d'un modèle de régression à deux états.

Les modèles de mélange de régressions [48, 58], connus également sous le nom de *switching regression* apportent une solution, en supposant que plusieurs régressions linéaires sont *cachées* au sein des données. Le modèle considère que $Y \in \mathcal{Y} = \mathbb{R}$ est relié aux covariables $\boldsymbol{x}$ par un des K modèles de régression suivants :

$$Y = \boldsymbol{\beta}_k' \Psi(\boldsymbol{x}) + \epsilon, \ k = 1, ..., K$$

où $\epsilon \sim \mathcal{N}(0, \sigma_k^2)$, $\boldsymbol{\beta}_k = (\beta_{k0}, ..., \beta_{kd})' \in \{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K\}$ est le vecteur des paramètres de régression ($\boldsymbol{\beta}_k \in \mathbb{R}^{d+1}$) et $\sigma_k^2 \in \{\sigma_1^2, \ldots, \sigma_K^2\}$ est la variance résiduelle. La distribution conditionnelle de Y sachant $\boldsymbol{x}$ est :

$$f(y|\boldsymbol{x}; \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k f_k(y; \boldsymbol{\beta}_k^t \Psi(\boldsymbol{x}), \sigma_k^2),$$

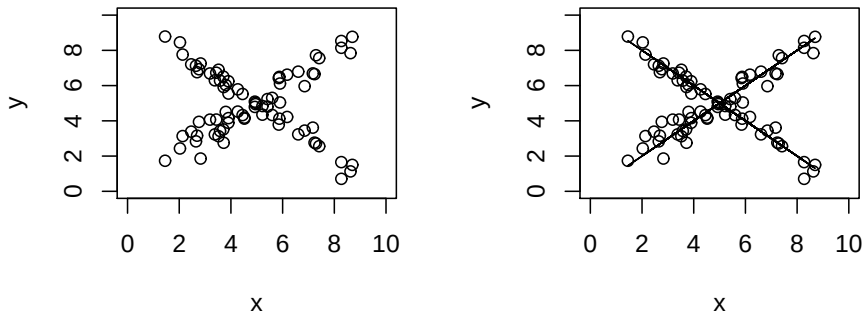


FIG. 1 – Illustration d'un phénomène à deux états pour lequel un mélange de deux régressions semble bien adapté.

où $\pi_1, \ldots, \pi_K$ sont les proportions du mélange et $f_k(\cdot)$ est la densité gaussienne.
L'estimation de ce modèle peut être réalisée par maximum de vraisemblance, mais pas directement. En effet, comme les appartenances des observations à l'une ou l'autre des régressions ne sont pas connues, une solution efficace est d'avoir recours à l'algorithme EM. La prédiction quant à elle s'effectue en deux étapes : pour un nouvel $\boldsymbol{x}$ observé, on estime tout d'abord à quelle régression il appartient, puis dans un second temps on utilise cette régression pour prédire le y correspondant.

Comme pour la régression simple, l'idée est d'adapter un modèle de mélange de régressions d'une population source Ω vers une population cible Ω^*. Pour ce faire, nous proposons dans [R1] une famille de modèles paramétriques parcimonieux, supposant

$$\boldsymbol{\beta}_k^* = \Lambda_k \boldsymbol{\beta}_k, \text{ avec } \Lambda_k = \text{diag}(\lambda_{k0}, \lambda_{k1}, \ldots, \lambda_{kd}) \quad \text{et} \quad \sigma_k^* \text{ libre.}$$

Comme pour la régression classique, nous proposons de définir des modèles parcimonieux en introduisant des contraintes sur les matrices Λ_k : 12 modèles ont ainsi été définis et présentés en détail dans [R1]. Ces 12 modèles sont regroupés en 5 familles. Les modèles MM_1 qui ne supposent aucune transformation entre les deux populations, tandis que les modèles MM_5 considèrent qu'elles sont indépendantes. Les modèles MM_2 supposent que le lien entre les modèles de régression est indépendant des covariables et des composantes du mélange. Les modèles MM_3 supposent que le lien est indépendant des composantes mais dépendant des covariables, tandis que les modèles MM_4 font l'inverse. Le tableau 4 présente la complexité de ces modèles en nombre de paramètres.

Modèle	MM_1	MM_{2a-c}	MM_{2d}	MM_{3a-c}	MM_{3d}	MM_{4a}	MM_{4b}	MM_5
Nb. de paramètres	0	1	2	K	$2K$	$d+K$	$d+K+1$	$K(d+2)$

TAB. 4 – Nombre de paramètres à estimer pour les modèles de lien dans le cas du mélange de régressions.

Comme dans le cas de mélange de régressions classique, l'estimation par maximum de vraisemblance des paramètres des modèles parcimonieux de lien peut être réalisée à l'aide d'un algorithme EM. Le détail des calculs à réaliser au cours des étapes E et M est donné dans [R1].

2.3.3 *Modèles adaptatifs bayésiens pour les mélanges de régressions*

Nous avons également développé dans [R1] une approche bayésienne permettant d'établir un lien entre les modèles de mélange de régressions des populations source et cible. Considérant une modélisation bayésienne du mélange de régressions pour Ω^*, l'approche développée consiste à faire dépendre les lois a priori des paramètres du mélange de régressions sur Ω^* de la population source Ω.

Modélisation bayésienne du mélange de régressions

Une approche bayésienne des mélanges de régressions [62, 90] est une alternative intéressante à l'estimation par maximum de vraisemblance via l'algorithme EM. Afin de lier les populations source et cible, nous supposons que la distribution a priori de $\boldsymbol{\beta}_k^*$, $k = 1, ..., K$, est une loi normale centrée en $\boldsymbol{\beta}_k$:

$$\boldsymbol{\beta}_k^* \sim \mathcal{N}(\boldsymbol{\beta}_k, \sigma_k^{*2}\boldsymbol{A}_k),$$

où $\boldsymbol{A}_k$ est une matrice de variance de taille $(p+1) \times (p+1)$. Le terme de variance $\sigma_k^{*2}\boldsymbol{A}_k$ contrôle alors la similitude entre les coefficients de régression des deux populations source et cible. La distribution a priori des proportions du mélange $\boldsymbol{\pi}^* = \{\pi_1^*, ..., \pi_K^*\}$ est supposée être une loi de Dirichlet centrée sur les proportions de Ω :

$$\boldsymbol{\pi}^* \sim \mathcal{D}(\pi_1, ..., \pi_K).$$

La loi a priori de σ_k^{*2}, $k = 1, ..., K$, est supposée être une Inverse-Gamma $\mathcal{IG}(\gamma_k, \nu_k)$. Classiquement, nous supposons les paramètres $\boldsymbol{\beta}_k^*$, $\boldsymbol{\pi}_k^*$ et σ_k^{*2} indépendants conditionnellement au groupe d'appartenance k. Dans ce travail, nous ajoutons l'hypothèse supplémentaire d'indépendance globale des couples $(\boldsymbol{\beta}_k^*, \sigma_k^{*2})$. Enfin, les paramètres $\boldsymbol{A}_k$, γ_k et ν_k ont été fixés arbitrairement.
L'inférence de ce modèle bayésien de mélange de régressions peut être réalisée à l'aide d'une approche Monte-Carlo par chaîne de Markov (MCMC, [90]), connue sous le nom d'échantillonneur de Gibbs, que nous présentons ci-après.

Estimation par échantillonneur de Gibbs

Nous considérons l'échantillonneur de Gibbs afin d'inférer le mélange de régressions de la population Ω^*. L'application de cette technique dans le cas des modèles de mélange est décrite dans [39]. Pour cela, considérons la variable latente $Z^* \in \{0, 1\}^K$ représentant l'appartenance des observations à l'une des K composantes du mélange. L'algorithme de Gibbs échantillonne alors, à l'itération q, des valeurs des paramètres β_k^*, π_k^* et σ_k^{*2} selon les distributions a priori suivantes :

– la distribution a priori conditionnelle de $\boldsymbol{Z}^*$ est une loi multinomiale :

$$\boldsymbol{z}_i^* | Y^*, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\pi}}, \boldsymbol{\beta}^*, \boldsymbol{\pi}^*, {\sigma^*}^2 \sim \mathcal{M}(1, t_{i1}, ..., t_{iK}),$$

où $t_{ik} = \pi_k^* \phi(y_i^* | \boldsymbol{x}_i^{*t} \boldsymbol{\beta}_k^*, \sigma_k^{*2}) / \sum_{\ell=1}^K \pi_\ell^* \phi(y_i^* | \boldsymbol{x}_i^{*t} \boldsymbol{\beta}_\ell^*, \sigma_\ell^{*2})$ avec ϕ la densité gaussienne,
– la distribution a priori conditionnelle de $\boldsymbol{\pi}^*$est une loi de Dirichlet :

$$\boldsymbol{\pi}^* | Y^*, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\pi}}, \boldsymbol{Z}^*, \boldsymbol{\beta}^*, {\sigma^*}^2 \sim \mathcal{D}(\hat{\pi}_1 + n_1^*, ..., \hat{\pi}_K + n_K^*),$$

avec $n_k^* = \sum_{i=1}^n z_{ik}^*$,
– une fois les appartenances aux composantes du mélange estimées, les observations d'une même composante k peuvent être regroupées au sein des matrices notées $\mathbf{x}_k^*$ et $\mathbf{Y}_k^*$, $k = 1, ..., K$. Avec ces notations, la distribution a priori conditionnelle de σ_k^{*2} est une Inverse Gamma :

$$\sigma_k^{*2} | \mathbf{Y}_k^*, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\pi}}, \boldsymbol{Z}^*, \pi^*, \boldsymbol{\beta}_k^* \sim \mathcal{IG}(\gamma_k + n_k/2, \nu_k + S_k/2),$$

où $S_k = (\mathbf{Y}_k^* - \mathbf{x}_k^{*t} \boldsymbol{\beta}_k^*)^t (\mathbf{Y}_k^* - \mathbf{x}_k^{*t} \boldsymbol{\beta}_k^*) + (\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^*)^t (\boldsymbol{A}_k + (\mathbf{x}_k^{*t} \mathbf{x}_k^*)^{-1})^{-1} (\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^*)$,
– finalement, la distribution a priori conditionnelle de $\boldsymbol{\beta}_k^*$ est une loi normale :

$$\boldsymbol{\beta}_k^* | \mathbf{Y}_k^*, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\pi}}, \boldsymbol{Z}^*, \boldsymbol{\pi}^*, \sigma_k^{*2} \sim \mathcal{N}(\boldsymbol{m}_k, \boldsymbol{\Delta}_k),$$

avec

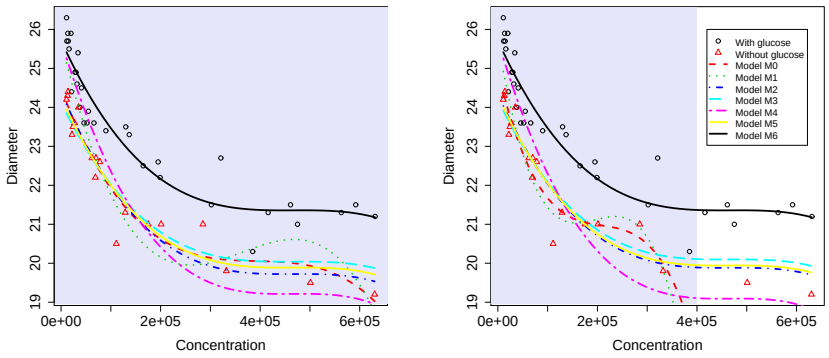
$$\begin{aligned} \boldsymbol{m}_k &= (\boldsymbol{A}_k^{-1} + \mathbf{x}_k^{*t} \mathbf{x}_k^*)^{-1} (\mathbf{x}_k^{*t} \mathbf{Y}_k^* + \boldsymbol{A}_k^{-1} \hat{\boldsymbol{\beta}}_k), \\ \boldsymbol{\Delta}_k &= \sigma_k^{*2} (\mathbf{x}_k^{*t} \mathbf{x}_k^* + \boldsymbol{A}_k^{-1})^{-1}. \end{aligned}$$

Des estimateurs consistants des paramètres π_k^*, β_k^* et ${\sigma^*}_k^2$ sont obtenus en moyennant les paramètres simulés sur les $Q - q_0$ dernières itérations, où q_0 définit le nombre d'itérations de la phase de *chauffe* de l'échantillonneur de Gibbs. Remarquons qu'avec ce type d'échantillonnage, des problèmes peuvent arriver lorsque les numéros des classes sont inter-changés au cours des itérations, ce qui peut arriver étant donné que la numérotation des classes est arbitraire. Pour palier à ce problème, connu sous le nom de *label switching*, nous utilisons la solution proposée dans [28] qui consiste à classer les paramètres simulés à l'aide d'un algorithme des k-means.

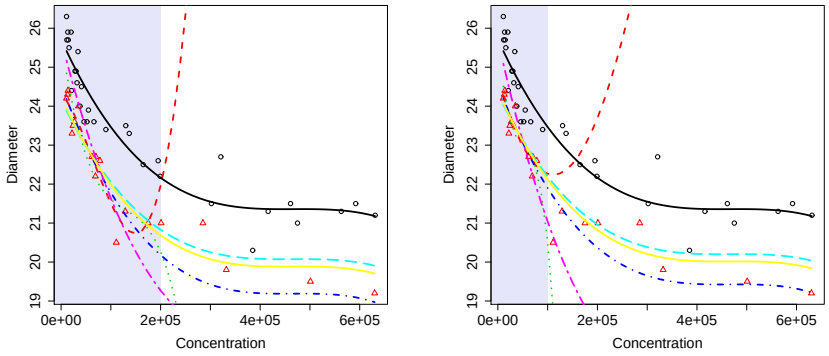
2.3.4 *Expérimentations numériques*

Afin d'illustrer l'intérêt des modèles adaptatifs pour la régression linéaire, nous avons conduit dans [R3] une étude sur données biologiques, dans laquelle nous avons appliqué une censure sur les données disponibles pour Ω^*. Ainsi, la connaissance et l'utilisation d'un modèle de régression sur Ω doit permettre, grâce aux modèles adaptatifs proposés, de régulariser l'estimation du modèle sur Ω^*.
Le jeu de données utilisé, *hellung*, disponible dans le paquet ISwR du logiciel **R**, reporte l'évolution de cellules Tetrahymena dans deux conditions de cultures : avec ou sans glucose. Dans les deux cas, le diamètre moyen ainsi que la concentration du milieu en nombre

de cellules ont été enregistrés au cours du temps. La population des cellules avec glucose ($n = 32$ observations) a été considérée comme population source, tandis que les cellules sans glucose forment la population cible. Cette dernière population a été utilisée avec toutes les données disponibles ($n^* = 19$ observations) puis avec trois modèles de censures, supprimant les données au delà d'un certain seuil. La base de fonction $(\psi_j)_{1\leq j\leq d}$ utilisée est composée de polynômes d'ordre 3.
La figure 2 présente les modèles de régression ainsi estimés sur Ω^* à l'aide des modèles adaptatifs et par moindres carrés ordinaires. On constate que la censure sur Ω^* a un effet très marqué sur l'estimateur des moindres carrés (modèle M_0) tandis que les modèles adaptatifs utilisant l'information disponible dans Ω fournissent des estimateurs très stables et ce même avec une très forte censure.



Pas de censure $\quad X \leq 4 \times 10^5$



$X \leq 2 \times 10^5 \quad X \leq 1 \times 10^5$

FIG. 2 – Effet de la censure sur Ω^* sur les modèles adaptatifs de régression pour les données *hellung*. La zone bleue correspond à la partie des observations de Ω^* utilisée pour l'estimation des modèles.

Le même type d'étude a été conduit dans [R1] dans le cadre des mélanges de régressions. Pour cela, nous avons simulé des données selon un modèle de mélange à deux composantes avec une base de fonctions polynomiales d'ordre 2. La figure 3 représente ces données simulées, avec des cercles gris pour Ω et des triangles rouges pour Ω^* (graphiques du haut), ainsi que l'estimation des modèles de mélange de régressions sur Ω^*

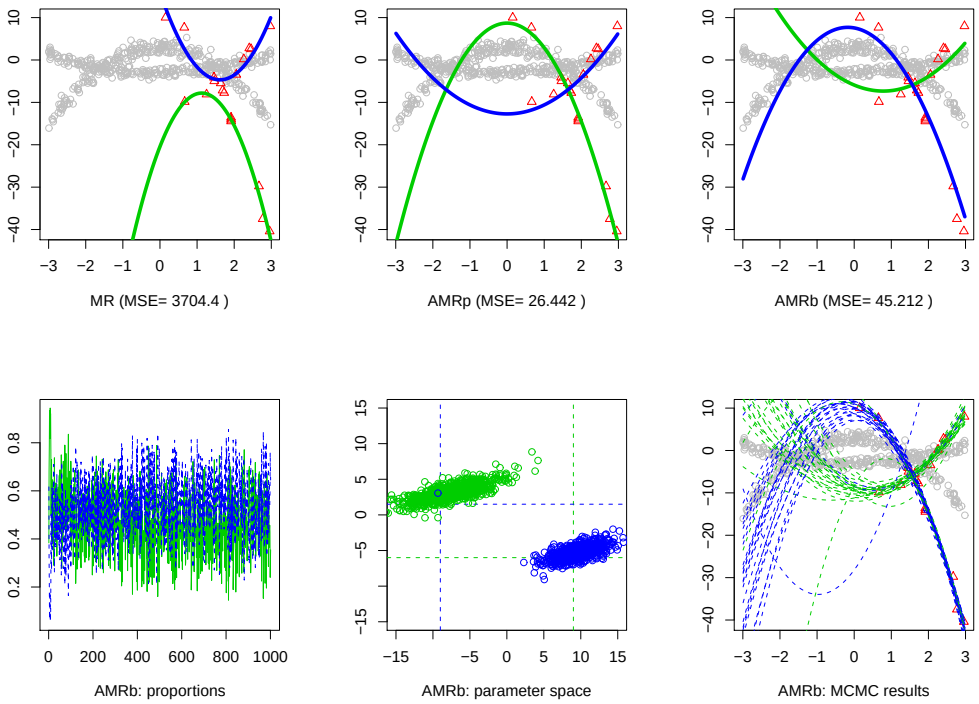


FIG. 3 – En haut : modèles de régression estimés pour un jeu de données simulées avec les méthodes MR, AMRp et AMRb. En bas et de gauche à droite : proportions π_k^* estimées au cours des itérations de l'échantillonneur de Gibbs, β_k simulés et modèles de régression associés.

par moindres carrés (à gauche) et à l'aide des modèles adaptatifs paramétriques (AMRp, graphique du milieu) et bayésiens (AMRb, graphique de droite). Les graphiques du bas illustrent l'échantillonnage de Gibbs des paramètres $\boldsymbol{\pi}_k^*$ et $\boldsymbol{\beta}_k^*$ ainsi que quelques modèles de régression associés.
Comme dans le cas de la régression classique, étant donnée la faible taille d'échantillon disponible pour Ω^*, l'estimation par moindres carrés ordinaires sur-apprend les données et fournit une estimation du modèle de régression éloignée de la réalité. Les modèles d'apprentissage adaptatifs paramétriques et bayésiens fournissent des estimations nettement plus satisfaisantes, avec des erreurs quadratiques moyennes respectives de 26.4 et 45.3 contre 3704 pour les moindres carrés ordinaires.

Les modèles adaptatifs pour les mélanges de régressions ont aussi été appliqués sur des données socio-économiques dans [R1,R2]. Ces données consistent en les émissions de CO2 et le PIB par habitant de 111 pays, mesurés en 1980 et en 1999 (figure 4). L'objectif était d'utiliser l'année 1980 comme population source et l'année 1999 comme population cible. Des bases de fonctions polynomiales d'ordre 2 ont été utilisées, et les modèles adaptatifs paramétriques et bayésiens ont été comparés à la régression simple (UR) ainsi qu'aux mélanges de régressions classique (MR) sur Ω^*.
Le tableau 5 donne les erreurs quadratiques moyennes en fonction de la taille de $\mathcal{S}^*$ pour

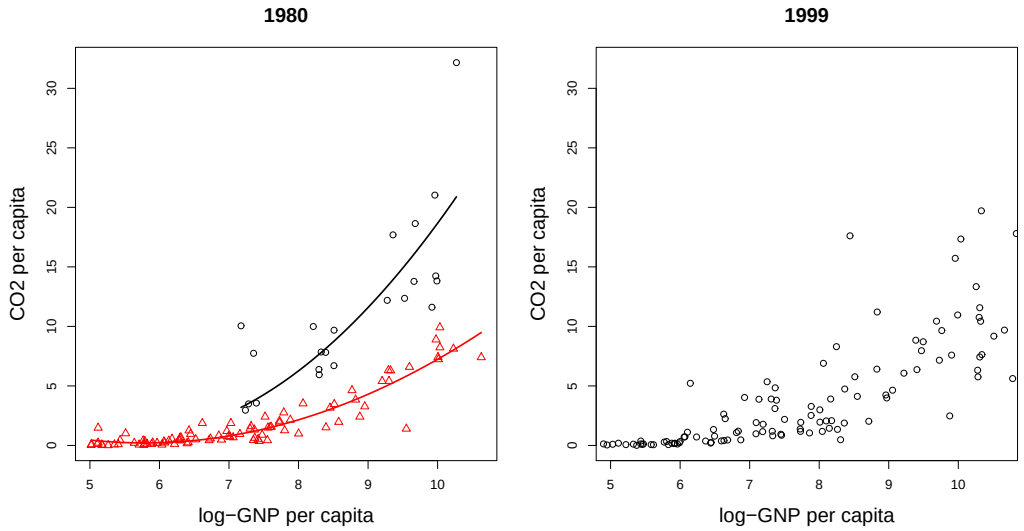


FIG. 4 – Émission de CO_2 par habitant *versus* PIB par habitant en 1980 (gauche) et 1999 (droite).

les quatre méthodes étudiées. Le choix parmi les différents modèles adaptatifs paramétriques est réalisé à l'aide du critère BIC. Les modèles adaptatifs, paramétriques notamment, montrent encore de très bonnes performances lorsque la taille de $\mathcal{S}^*$ est restreinte. De plus, le choix du modèle adaptatif permet d'apporter un éclairage sur l'évolution des données de 1980 à 1999 : le choix d'un modèle de type MM_2 indique que chaque composante du mélange a évolué de façon similaire au cours du temps.

Modèle	30% de $\mathcal{S}^*$	50% de $\mathcal{S}^*$	70% de $\mathcal{S}^*$	100% de $\mathcal{S}^*$
AMRp	**3.86**	**3.44**	3.53	3.47
AMRb	5.99	5.66	5.99	5.66
UR	7.66	7.21	7.10	6.99
MR	5.11	4.77	**3.33**	**2.89**

TAB. 5 – Erreurs quadratiques moyennes en fonction de la taille de $\mathcal{S}^*$ pour les modèles adaptatifs paramétriques (AMRp) et bayésiens (AMRb), la régression simple (UR) ainsi qu'un mélange de régressions classique (MR).

La figure 5 illustre les mélanges de régressions estimés, par maximum de vraisemblance pour les données de 1980 et en utilisant les modèles AMRp pour les données de 1999.

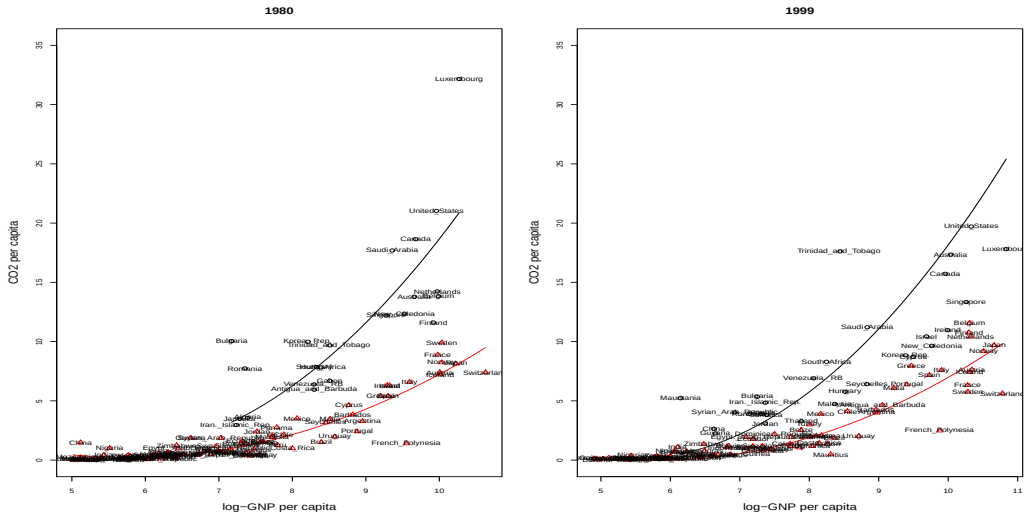


FIG. 5 – Émission de CO_2 par habitant *versus* logarithme du PIB par habitant en 1980 (gauche) et 1999 (droite) et mélange de régressions estimés (utilisant un modèle AMRp pour 1999).

3

MODÈLES GÉNÉRATIFS POUR DONNÉES DE RANG ET ORDINALES

Sommaire

PUBLICATIONS ASSOCIÉES À CE CHAPITRE

Revues avec comité de lecture

[R6] Biernacki, C. and Jacques, J. *Binary Search Ordinal: a new model for ordinal data clustering*, en préparation.

[R7] Biernacki, C. and Jacques, J. *Model-based clustering for multivariate partial ranking data*, soumis.

[R8] Grimonprez, Q., Biernacki, C. and Jacques, J. *Rankclust: an R package for clustering multivariate partial ranking data*, en préparation.

[R9] Biernacki, C. and Jacques, J. *A generative model for rank data based on sorting algorithm*, Computational Statistics and Data Analysis, 2012, DOI 10.1016/j.csda.2012.08.008.

Conférences internationales avec comité de lecture

[CI2] Jacques, J. and Biernacki, C. *A generative model for rank data based on sorting algorithm*. 19th International Conference on Computational Statistics (COMPSTAT'10), Paris, France, août 2010.

Conférences nationales

[CN8] Biernacki, C. and Jacques, J. *Modèle génératif pour données ordinales*. 44èmes Journées de Statistique organisée par la Société Française de Statistique, Bruxelles, mai 2012.

[CN9] Jacques, J. and Biernacki, C. *Model-based clustering for rank data based on an insertion sorting algorithm*. 17ème Rencontres de la Société Francophone de Classification, La Réunion, juin 2010.

[CN10] Biernacki, C. and Jacques, J. *Modèles génératifs de rangs relatifs à un algorithme de tri par insertion*. 42èmes Journées de Statistique organisée par la Société Française de Statistique, Marseille, mai 2010.

3.1 PROBLÉMATIQUE

Dans les travaux de recherche actuels en statistique, une place prédominante est accordée aux données continues ($\mathcal{X} = \mathbb{R}^p$ typiquement), alors que d'autres types de données (catégorielles, ordinales, par intervalle, de rang...) sont également très répandues. Par exemple, dans le domaine des enquêtes et sondages, les données sont généralement catégorielles ou ordinales, voir de rang lorsqu'on demande de classer plusieurs items par ordre de préférence. Le manque de modèles propres à ce type de données conduit souvent l'utilisateur à transformer les données, avec toutes les pertes d'information et biais que cela peut induire. Ainsi, les données ordinales sont souvent traitées soit comme des données qualitatives, et dans ce cas l'information d'ordre est perdue, soit comme des données quantitatives, ce qui induit un biais dû au choix des distances entre modalités.
Fort de ce constat, nous avons proposé dans [R9] et [R6] deux modèles probabilistes pour les données de rang et les données ordinales. Ces modèles sont tout deux construits sur un même principe, qui consiste à modéliser le processus de génération des données. Ainsi, les données de rang sont interprétées comme le résultat d'un algorithme de tri, basé sur des comparaisons élémentaires de paires d'objets. Le modèle probabiliste est alors défini en introduisant de l'aléatoire dans la prise de décision lors de la comparaison d'une paire d'objets, et en propageant l'impact de cette prise de décision aléatoire sur la probabilité finale des rangs. De même, les données ordinales sont vues comme le résultat d'un algorithme de recherche dans une table ordonnée de modalités.
L'organisation de ce chapitre est la suivante. Pour chacun des deux types de données, de rang et ordinales, après avoir établi un tour d'horizon des modèles existants nous présentons le modèle probabiliste développé. Nous montrons ensuite comment ces modèles peuvent être utilisés, sous la forme de modèle de mélange, pour définir des algorithmes de classifications automatiques. Des applications sur données réelles illustrent l'intérêt de chaque modèle.

3.2 ANALYSE ET MODÉLISATION DES DONNÉES DE RANG

Une donnée de rang est le résultat d'un classement par un juge de m objets $\mathcal{O}_1, \ldots, \mathcal{O}_m$ suivant un certain ordre (de préférence par exemple). Une façon de représenter cette donnée est la notation *ordering* que nous utilisons dans la suite de ce document : $x = (x_1, \ldots, x_m)$ signifiant que l'objet $\mathcal{O}_{x_i}$ est rangé à la $i$ème place. Le rang x ainsi noté est un élément de l'ensemble $\mathcal{X} = \mathcal{P}_m$ des permutations des m premiers entiers. Nous supposons dans cette section que les rangs sont observés *complets*, ce qui signifie que tous les objets ont été classés.

Marden [74] fournit une synthèse des modèles existants pour de telles données, que l'on peut résumer en trois classes :

– les modèles de comparaison par paire [21, 66], dont le plus célèbre est le modèle Φ de Mallows :

$$P(x; \mu, \lambda) = \mathcal{C}(\lambda)^{-1} \exp^{-\lambda d_K(x,\mu)},$$

où μ est le rang de référence, exprimant une notion de tendance centrale de la distribution, λ est un paramètre de précision et d_K la distance de Kendall [30, 65], définie comme le nombre minimum de permutations d'éléments adjacents nécessaires à transformer un rang en l'autre. La constante de normalisation $C(\lambda)$ peut être calculée de façon explicite ([43] et [R9]). Ce modèle est l'un des plus étudiés ces dernières années, avec notamment des extensions au cas de rangs partiels (lorsque tous les objets ne sont pas classés, [69]) ou de rangs multiples (les mêmes rangs sont observés plusieurs fois sous différentes conditions, [70]). Murphy et Martin [82] utilisent également ce modèle pour définir un algorithme de classification automatique de données de rang.

- les modèles multi-étapes [44, 73, 86] considèrent que le juge choisit au fur et à mesure dans l'ensemble d'objet celui qu'il préfère. Ces modèles ont également été récemment étendus [9, 50, 51] et utilisés en classification automatique [52, 53].
- les modèles de Thursthone [16, 101] considèrent l'existence d'une variable latente gaussienne associée à chaque objet, et interprètent une donnée de rang comme un classement des réalisations de ces variables aléatoires. Ces modèles sont assez peu utilisés en pratique en raison des problèmes d'intégrations numériques qu'ils induisent.

Nous expliquons dans [R9] que les deux premières classes de modèles peuvent être interprétées comme la modélisation d'algorithmes de tri des objets, non optimaux d'un point de vue du nombre de comparaisons de paires effectuées. Le modèle ISR que nous avons publié dans [R9] considère l'algorithme de tri par insertion, qui s'avère être optimal pour un petit nombre d'objets à classer ($m < 10$). Néanmoins, ISR peut être utilisé pour un nombre d'objets à classer plus important.

3.2.1 ISR : *un modèle probabiliste pour données de rang*

Le modèle

Le modèle que nous proposons, basé sur une modélisation d'une version stochastique de l'algorithme de tri par insertion, est le suivant :

$$P(x;\mu,p) = \frac{1}{m!}\sum_{\sigma\in\mathcal{P}_m} \underbrace{p^{G(x,\sigma,\mu)}(1-p)^{A(x,\sigma)-G(x,\sigma,\mu)}}_{P(x|\sigma;\mu,p)}, \tag{10}$$

où μ est le mode de la distribution et p est la probabilité pour le juge d'effectuer une *bonne* comparaison de paires (*i.e.* en accord avec μ). $G(x,\sigma,\mu)$ est le nombre de bonnes comparaisons de paires effectués lors du processus de tri par insertion ayant retourné x comme résultat, avec comme ordre de présentation initial des objets l'ordre σ et pour un rang de référence μ. $A(x,\sigma)$ est quant à lui le nombre total de comparaisons de paires effectuées. Les expressions de A et G sont données dans [R9].

Nous avons montré que ce modèle admettait plusieurs propriétés très intéressantes d'un point de vue interprétabilité : la distribution est uniforme lorsque $p = \frac{1}{2}$; μ est le mode de la distribution, d'autant plus prononcé que p est proche de 1 ; l'identifiabilité est assurée

pour $p > \frac{1}{2}$; le rang défini par $\bar{\mu} = \mu \circ \bar{e}$, où $\bar{e} = (m, \ldots, 1)$ est la permutation d'inversion totale, est l'anti-mode de la distribution, c'est-à-dire le rang de probabilité minimale.
En outre, ce modèle est le premier modèle pour données de rang prenant en compte l'ordre de présentation initial des objets, dont on se convainc aisément de l'importance dès lors que la donnée de rang est vue comme le résultat d'un algorithme de tri.

Estimation

Nous présentons dans [R9] plusieurs algorithmes de maximisation de la log-vraisemblance du modèle ISR :

$$l(\mu, p; \boldsymbol{x}) = \sum_{i=1}^{n} \ln \left(\frac{1}{m!} \sum_{\sigma \in \mathcal{P}_m} P(x_i | \sigma; \mu, p) \right)$$

où $\boldsymbol{x} = (x_1, \ldots, x_n)'$ est un échantillon i.i.d. de rangs provenant du modèle ISR. En assimilant les ordres de présentation initiaux σ des objets à des variables manquantes, une façon efficace de maximiser cette vraisemblance est d'utiliser l'algorithme EM. Mais lorsque m dépasse 10, la taille de l'espace $\mathcal{P}_m$ rend impossible tout parcours exhaustif des $m!$ éléments qu'il contient, et l'algorithme EM n'est plus utilisable. Il faut alors avoir recours à un algorithme de type SEM que nous détaillons ci-après.

ALGORITHME EM POUR $m < 10$. La log-vraisemblance complétée que maximise l'algorithme EM a pour expression

$$l_c(\mu, p; \boldsymbol{x}) = \sum_{i=1}^{n} \sum_{\sigma \in \mathcal{P}_m} \mathbf{1}\{\sigma = \sigma_i\} \ln \left(\frac{1}{m!} P(x_i | \sigma; \mu, p) \right).$$

La $(q+1)$ème itération de l'étape E calcule l'espérance conditionnelle $\mathcal{Q}$ de l_c :

$$\mathcal{Q}((\mu, p), (\mu, p)^{\{q\}}; \boldsymbol{x}) = \sum_{i=1}^{n} \sum_{\sigma \in \mathcal{P}_m} t_{i\sigma}^{\{q\}} \ln \left(\frac{1}{m!} P(x_i | \sigma; \mu, p) \right),$$

où $t_{i\sigma}^{\{q\}}$ est la probabilité conditionnelle que $\sigma_i = \sigma$, donnée par

$$t_{i\sigma}^{\{q\}} = \frac{P(x_i | \sigma; (\mu, p)^{\{q\}})}{\sum_{\tau \in \mathcal{P}_m} P(x_i | \tau; (\mu, p)^{\{q\}})},$$

tandis que l'étape M maximise $\mathcal{Q}((\mu, p), (\mu, p)^{\{q\}}; \boldsymbol{x})$ en fonction de $(\mu, p) \in \mathcal{P}_m \times [\frac{1}{2}, 1]$. Dans le cas où $m < 10$, la maximisation suivant μ est réalisée en testant tous les éléments de $\mathcal{P}_m$ (nous verrons plus loin une stratégie permettant de restreindre cet espace). L'expression du p maximisant $\mathcal{Q}$ est quant à elle explicite :

$$p^{\{q+1\}} = \frac{\sum_{i=1}^{n} \sum_{\sigma \in \mathcal{P}_m} t_{i\sigma}^{\{q\}} G(x_i, \sigma, \mu^{\{q\}})}{\sum_{i=1}^{n} \sum_{\sigma \in \mathcal{P}_m} t_{i\sigma}^{\{q\}} A(x_i, \sigma)}.$$

ALGORITHME SEM-GIBBS POUR $m \geq 10$. Lorsque $m \geq 10$, le calcul exhaustif des $t_{i\sigma}^{\{q\}}$ n'est plus possible. Nous avons proposé dans [R9] une solution alternative faisant appel à un algorithme SEM-Gibbs. Alors que l'algorithme SEM classique [46, 25] consiste au sein de l'étape SE à simuler les variables manquantes, *i.e.* les ordres de présentation σ_i, à partir des probabilités $t_{i\sigma}^{\{q\}}$, l'algorithme SEM-Gibbs que nous avons proposé réalise cette simulation sans avoir recours au $t_{i\sigma}^{\{q\}}$, grâce à un échantillonneur de Gibbs [90]. L'étape SE-Gibbs correspondante prend alors la forme suivante : partant d'un échantillon arbitraire $\sigma_i^{\{q,0\}}$, on génère $r \in \{1, \ldots, R\}$ séquences $\sigma_i^{\{q,r\}}$ (R fixé) où

$$(\sigma_{ij}^{\{q,r+1\}}, \cdot) \sim p\left(\sigma_{ij}, \sigma_{ij+1} | \sigma_1^{\{q,r+1\}}, \ldots, \sigma_{ij-1}^{\{q,r+1\}}, \sigma_{ij+2}^{\{q,r\}}, \ldots, \sigma_{im}^{\{q,r\}}, \boldsymbol{x}; (\mu, p)^{\{q\}}\right)$$

pour $j \in \{1, \ldots, m-2\}$ et où

$$(\sigma_{im-1}^{\{q,r+1\}}, \sigma_{im}^{\{q,r+1\}}) \sim p\left(\sigma_{im-1}, \sigma_{im} | \sigma_{i1}^{\{q,r+1\}}, \ldots, \sigma_{im-2}^{\{q,r+1\}}, \boldsymbol{x}; (\mu, p)^{\{q\}}\right).$$

La maximisation de $l_c(\mu, p; \boldsymbol{x}, \sigma_1^{\{q\}}, \ldots, \sigma_n^{\{q\}})$ au sein de l'étape M consiste à parcourir tout $\mathcal{P}_m$ pour μ, ce qui est impossible dans le cas $m \geq 10$, mais une stratégie spécifique permet de surmonter ce problème, tandis que pour p nous avons :

$$p^{\{q+1\}} = \frac{\sum_{i=1}^{n} G(x_i, \sigma_i^{\{q\}}, \mu^{\{q\}})}{\sum_{i=1}^{n} A(x_i, \sigma_i^{\{q\}})},$$

qui ne souffre d'aucune difficulté combinatoire.
Après avoir itéré un certain nombre de fois ces étapes SE-Gibbs et M, et après avoir supprimé les premières itérations correspondant à une période de chauffe de l'algorithme, l'estimation des paramètres s'obtient en moyennant les valeurs de p pour chaque valeur différente de μ rencontrée, et en conservant le couple (μ, p) maximisant la vraisemblance. Cette dernière souffrant également de complexité combinatoire, une approximation pourra être utilisée en remplaçant $\frac{1}{m!} \sum_{\sigma \in \mathcal{P}_m} P(x_i | \sigma; \mu, p)$ dans (11) par $\frac{1}{S} \sum_{s=1}^{S} P(x_i | \sigma_s; \mu, p)$ où $(\sigma_s)_{s=1,S}$ est un échantillon de simulations des ordres de présentation obtenu comme dans l'étape SE-Gibbs.

RÉDUCTION DU NOMBRE DE MODES POTENTIELS. Enfin, nous proposons dans [R9] une stratégie permettant de réduire considérablement le nombre de candidats pour μ. Cette stratégie est basée sur une estimation par bootstrap paramétrique du nombre de rangs potentiellement plus fréquents que μ dans un échantillon de taille similaire à celui observé.

3.2.2 *Expérimentations numériques*

Nous présentons ici une comparaison des modèles ISR et Φ de Mallows sur 4 jeux de données :

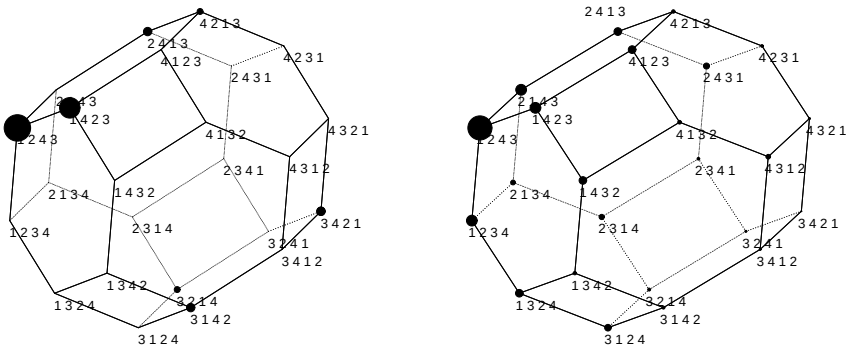


FIG. 6 – Distributions empirique (gauche) et estimée (droite) pour le quizz Football.

- *Football* [R9]. Ce quiz de culture général, proposé à 40 étudiants, a pour objectif de classer 4 équipes nationales en fonction du nombre croissant de victoires lors de la coupe du monde de football : $\mathcal{O}_1 =$ France, $\mathcal{O}_2 =$ Allemagne, $\mathcal{O}_3 =$ Brésil, $\mathcal{O}_4 =$ Italie. La réponse correcte est $\mu^* = (1, 2, 4, 3)$. La figure 6 (gauche) présente ces données sous la forme d'un polytope, sur lequel la taille des points est proportionnelle aux effectifs observés.
- *Words* [43]. Il a été demandé à 98 étudiants de classer 5 mots en fonction de l'intensité de leur association avec le mot « idée » : $\mathcal{O}_1 =$ pensée, $\mathcal{O}_2 =$ jeu, $\mathcal{O}_3 =$ théorie, $\mathcal{O}_4 =$ rève et $\mathcal{O}_5 =$ attention.
- *Sports* [74]. 130 étudiants de l'Université de l'Illinois ont classé 7 sports par ordre de préférence : $\mathcal{O}_1 =$ baseball, $\mathcal{O}_2 =$ football, $\mathcal{O}_3 =$ basketball, $\mathcal{O}_4 =$ tennis, $\mathcal{O}_5 =$ cyclisme, $\mathcal{O}_6 =$ natation, $\mathcal{O}_7 =$ jogging.
- *Election* [51]. Le système d'élection au parlement Irlandais est basé sur le classement des candidats par ordre de préférence. Ce jeu de données consiste en les 2490 rangs complets enregistrés lors des élections de 2002, représentant 4% de l'ensemble des votes.

Les résultats d'estimation sont présentés dans la table 6. On y voit que le modèle ISR s'avère être une alternative très intéressante au modèle Φ de Mallows. Par ailleurs l'interprétation des μ estimés par le modèle ISR apporte des informations intéressantes : par exemple, sur les données Sports, on constate que les étudiants de l'Université de l'Illinois préfèrent tout d'abord les sports collectifs, puis le tennis, et enfin les sports individuels. Le tennis se trouve tout à fait à sa place entre les sports collectifs et individuels, ce qui n'est pas le cas du μ estimé par le Φ de Mallows. Concernant les données des élections irlandaises, on remarque dans le rang modal, que tous les candidats, à partir de la troisième place, sont classés selon l'ordre de présentation. Heureusement, dans ce système les électeurs n'ont pas l'obligation de classer tous les candidats (seulement 4% d'entre eux le font). Dans le cas contraire, l'ordre de présentation des candidats aurait un rôle très important sur le résultat des élections.

données	modèle	$\hat{\mu}$	$\hat{p}$ ou $\hat{\lambda}$	l
Football	ISR	$(1,2,4,3)$	0.834	**-88.53**
	Mallows	$(1,2,4,3)$	1.106	-89.17
Words	ISR	$(2,5,4,3,1)$	0.879	-275.43
	Mallows	$(2,5,4,3,1)$	1.431	**-251.27**
Sports	ISR	$(1,3,2,4,5,7,6)$	0.564	**-1102.12**
	Mallows	$(1,3,4,2,5,6,7)$	0.083	-1102.84
Election	ISR	$(13,4,1,2,3,5,6,7,8,9,10,11,12,14)$	0.682	**-48329.76**
	Mallows	$(4,13,2,5,1,14,7,6,10,8,9,12,3,11)$	0.164	-60157.38

TAB. 6 – Résultats d'estimation des modèles ISR et Φ de Mallows : estimation des paramètres $\hat{\mu}$, $\hat{p}$ (ISR) et $\hat{\lambda}$ (Mallows), maximum de la log-vraisemblance l.

3.2.3 *Classification automatique de données de rang multivariées*

Le bon comportement du modèle ISR nous a encouragé à proposer dans [R7] une extension permettant de traiter :

- la présence d'*hétérogénéité* au sein de la population de rangs observée. En effet, dans beaucoup d'applications, l'étude des données de rang révèle une hétérogénéité au sein de la population étudiée, due par exemple à des différences d'opinions politiques, de stratégies économiques ou de sensibilités personnelles. Murphy et Martin [82] proposent ainsi un mélange de Φ de Mallows pour la classification automatique de données de rang complètes univariées.
- les rangs *multivariés*, présents lorsque plusieurs données de rang sont disponibles pour un même individu. Ce type de données se rencontre souvent dans les études marketing lorsqu'on interroge des clients sur leurs préférences dans différents domaines. Par exemple, une agence de voyage pourra être intéressée de connaître pour de futurs clients un classement de leurs lieux de vacances préférés ainsi que du moyen de transport pour s'y rendre et du type d'hébergement. Rarement abordées dans la littérature (Böckenholt propose une extension du modèle de Thursthone [15] peu applicable en raison de problèmes d'intégration numériques), les données de rang multivariées sont néanmoins très fréquentes mais rarement traitées comme telles.
- la présence éventuelle de rangs *partiels*, lorsque les individus n'ont pas classés tous les objets qui leur ont été proposés. Par exemple, le jeux de données *Election* étudié dans la section précédente comporte près de 96% de rangs partiels, pour lesquels les électeurs n'ont pas classé l'ensemble des 14 candidats. Cette thématique des rangs partiels a été plus étudiée ces dernières années (cf. [69, 70] par exemple).

Le modèle

Soit $\boldsymbol{x} = (x_1, \ldots, x_p)'$ un ensemble de p rangs, appelés rangs multivariés, où chaque rang univarié $x_j = (x_{j1}, \ldots, x_{jm_j})$ correspond à un classement de m_j objets ($1 \leq j \leq p$).
Afin de prendre en compte les rangs partiels, notons $\check{I}_j \subset \{1, \ldots, m_j\}$ l'ensemble des indices correspondants à des positions dans le rang pour lesquelles un objet a été classé.

Ainsi, pour tout $h \in \check{I}_j$, la valeur x_{jh} de x_j est connue. De façon symétrique, nous notons $\tilde{I}_j \subset \{1, \ldots, m_j\}$ l'ensemble des indices des positions non observées : pour $h \in \tilde{I}_j$, x_{jh} n'est pas observée.
La population de rangs observée est supposée être composée de K groupes, présents en proportion $\pi_k \in [0,1]$ ($\sum_{k=1}^{K} \pi_k = 1$). Au sein de chaque groupe, les p données de rang sont modélisées par un modèle ISR de paramètres $\mu_{jk} \in \mathcal{P}_{m_j}$ et $p_{jk} \in [\frac{1}{2}, 1]$. Sous l'hypothèse d'indépendance conditionnelle, la probabilité d'un rang $\boldsymbol{x}$ est

$$\mathrm{p}(\boldsymbol{x}; \boldsymbol{\theta}) = \sum_{k=1}^{K} \frac{\alpha_k}{m_j!} \prod_{j=1}^{p} \sum_{\sigma \in \mathcal{P}_{m_j}} \mathrm{p}(x_j | \sigma; \mu_{jk}, p_{jk}) \tag{11}$$

où $\boldsymbol{\theta} = (p_{jk}, \mu_{jk}, \pi_k)_{j=1,\ldots,p,k=1,\ldots,K}$ et $\mathrm{p}(x_j|\sigma; \mu_{jk}, p_{jk})$ est donnée par (10).
Le mélange d'ISR multivarié ainsi défini permet des applications directes en classification supervisée et non supervisée de données de rang, qu'elles soient univariées ou multivariées, complètes ou partielles. Nous décrivons ci-après l'estimation de paramètre $\boldsymbol{\theta}$ dans le cas non-supervisé.

Estimation

L'estimation par maximum de vraisemblance du modèle (11) doit tenir compte de la présence de plusieurs types de données manquantes : tout d'abord les ordres de présentation σ des objets à classer, comme dans le cas univarié homogène, mais également les classes d'appartenance des rangs ainsi que les éventuels classements manquants (x_{jh} pour $h \in \tilde{I}_j$) en cas de présence de rangs partiels.
L'utilisation d'un algorithme EM pour maximiser la vraisemblance est très vite impraticable du fait de la complexité combinatoire induite. En effet, nous avons vu que dans le cas univarié homogène, l'algorithme EM était envisageable pour $m < 10$, mais l'introduction de deux types de variables manquantes supplémentaires réduit encore ce seuil. Nous proposons donc dans [R7] un algorithme d'estimation similaire à celui proposé dans [R9] pour le cas $m \geq 10$, basé sur un algorithme de type SEM-Gibbs.
Cet algorithme, consiste à itérer successivement une étape SE-Gibbs dans laquelle des réalisations des variables aléatoires manquantes sont générées, et une étape M dans laquelle le paramètre $\boldsymbol{\theta}$ maximisant la log-vraisemblance complétée (complétée ici par les réalisations générées dans l'étape SE-Gibbs) est calculé.
L'étape SE-Gibbs se décompose de la façon suivante :

- conditionnellement aux appartenances aux classes et aux rangs complétés de l'itération précédente, les ordres de présentation sont générés séquentiellement (de façon similaire au cas ISR univarié homogène),
- conditionnellement à ces ordres de présentation, aux rangs complétés de l'itération précédente ainsi qu'à la valeur courante de $\boldsymbol{\theta}$, les appartenances aux classes sont simulées suivant une loi multinomiale,
- conditionnellement aux ordres de présentation ainsi qu'aux appartenances aux classes ainsi simulés, et à la valeur courante de $\boldsymbol{\theta}$, les classements manquants x_{jh} ($h \in \tilde{I}_j$) sont simulés de manière séquentielle analogue à celle utilisée pour les ordres de présentation.

L'étape M quant à elle revient à calculer :

- les proportions des classes π_{jk} de façon classique utilisant les appartenances aux classes simulées,
- les couples (p_{jk}, μ_{jk}) maximisant la log-vraisemblance complétée. Pour ce faire, nous générons des valeurs possibles pour ces couples, en alternant une simulation séquentielle de μ_{jk} conditionnellement à p_{jk} et le calcul de p_{jk} maximisant la log-vraisemblance complétée conditionnellement à μ_{jk}. Après quelques itérations, le couple retenu sera celui maximisant la log-vraisemblance complétée.

Après avoir itéré un certain nombre de fois ces étapes SE-Gibbs et M, et après avoir supprimé la période de chauffe de l'algorithme, l'estimation de $\boldsymbol{\theta}$ s'obtient en moyennant les valeurs de p_{jk} et π_{jk} pour chaque valeur différente de $\{\mu_{jk}; 1 \leq j \leq p, 1 \leq k \leq K\}$ rencontrée, et en conservant l'ensemble de paramètres maximisant (une approximation de) la vraisemblance.

Expérimentations numériques

Le modèle ainsi défini est le seul modèle de la littérature permettant de réaliser une classification automatique de données de rangs multivariées potentiellement partielles. Nous illustrons ici l'intérêt de ce modèle sur les données du concours de l'Eurovision. Ce concours, qui consiste à élire la meilleure chanson présentée par les pays candidats, fonctionne sous la forme de vote de préférence. Lors de la finale de ce concours, chaque pays participant au vote (une quarantaine, la liste évoluant d'une année sur l'autre) émet un vote en classant par ordre de préférence 10 chansons, et donc pays, parmi les 25 participant à la finale. En raison de l'organisation de demi-finale avant le concours final, la liste des pays présents en finale évolue chaque année. Nous avons sélectionné les 8 pays ayant participé à toutes les finales du concours de 2007 à 2012 [1]. Les rangs étudiés sont donc de taille $m = 8$, multivariés puisque nous prenons en compte simultanément les $p = 6$ années de concours, et souvent incomplets car il est rare que les pays votant aient classés les 8 pays en question parmi leur 10 favoris. Enfin, la taille d'échantillon est $n = 34$, correspondant au nombre de pays ayant participé au vote lors des six concours de 2007 à 2012.

Nous avons réalisé sur ces données une classification automatique à l'aide du modèle ISR, en faisant varier le nombre de classes de 1 à 6. Le critère BIC a sélectionné 5 classes. Le tableau 7 fournit les résultats d'estimation pour 5 classes.

La classification obtenue est illustrée par la figure 7. On y constate certaines proximités géographiques entre les individus d'un même groupe, ce qui prône en faveur des votes d'alliances entre pays voisins souvent sujet à discussion lors de ce concours. En effet, on peut distinguer un groupe de pays d'Europe de l'Ouest (rouge), un groupe de pays de l'Europe de l'Est (gris), un groupe de pays méditerranéens (jaune) et deux groupes plus mélangés mais contenant néanmoins principalement des pays du Nord de l'Europe (bleu et vert).

[1] données disponibles sur `wikipedia`

k	π_k
1	0.353
2	0.088
3	0.235
4	0.176
5	0.147

	2007		2008		2009	
k	μ_{1k}	p_{1k}	μ_{2k}	p_{2k}	μ_{3k}	p_{3k}
1	(3,7,5,2,4,6,8,1)	0.831	(3,5,7,6,2,4,8,1)	0.874	(3,1,8,2,4,7,6,5)	0.845
2	(5,7,3,2,1,8,4,6)	0.915	(5,1,7,3,2,4,6,8)	0.889	(1,5,3,2,6,7,4,8)	0.886
3	(5,7,3,4,6,2,8,1)	0.888	(7,5,3,6,4,8,1,2)	0.886	(5,7,8,1,4,3,2,6)	0.747
4	(7,5,3,6,4,2,8,1)	0.921	(5,7,1,3,4,6,8,2)	0.852	(8,1,4,2,6,3,5,7)	0.892
5	(7,5,4,6,3,2,8,1)	0.911	(5,1,7,4,3,2,8,6)	0.921	(5,1,8,3,7,6,2,4)	0.949
	2010		2011		2012	
k	μ_{4k}	p_{4k}	μ_{5k}	p_{5k}	μ_{6k}	p_{6k}
1	(3,7,2,1,6,4,5,8)	0.838	(3,6,7,1,2,4,8,5)	0.763	(6,5,2,4,3,8,7,1)	0.863
2	(2,5,4,3,7,1,8,6)	0.875	(2,8,5,3,6,7,4,1)	0.967	(2,5,8,6,7,1,4,3)	0.881
3	(4,3,2,1,5,7,6,8)	0.855	(7,8,1,2,5,4,3,6)	0.789	(5,4,7,2,6,8,3,1)	0.825
4	(2,4,1,8,5,7,6,3)	0.972	(2,8,4,1,7,6,3,5)	0.889	(5,2,4,7,3,1,6,8)	0.909
5	(2,7,5,6,4,1,3,8)	0.869	(5,7,3,8,2,4,6,1)	0.803	(5,7,3,1,4,8,2,6)	0.703

TAB. 7 – Résultats d'estimation du modèle ISR multivarié hétérogène avec rangs partiels sur les données eurovision.

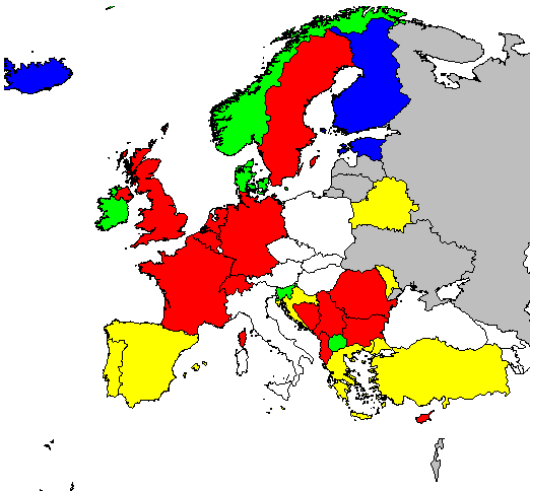


FIG. 7 – Classification estimée des pays européens en fonction de la similitude de leur vote au concours de l'Eurovision.

3.3 ANALYSE ET MODÉLISATION DES DONNÉES ORDINALES

Les données ordinales sont des données très répandues mais rarement traitées comme telles. Effectivement, en raison du peu de méthodes et modèles statistiques dédiés, une approximation nominale ou quantitative est souvent faite. S'il existe un certain nombre de travaux en régression [2], principalement basés sur des modèles log-linéaires contraints, les travaux en classification (supervisée ou non) sont beaucoup moins nombreux : Gouget [54] définit des modèles basés sur l'hypothèse que les données ordinales sont une discrétisation de données gaussiennes latentes ainsi que des modèles basés sur une contrainte du modèle multinomial imposant une certaine forme de décroissance des probabilités autour du mode de la distribution ; Giordand et Diana [47] définissent un modèle non paramétrique basé sur un seuillage des effectifs empiriques ; D'Elia et Piccolo [35] défi-

nissent le modèle CUB comme un mélange de distributions binomiale, uniforme et de Dirac, de sorte à obtenir un modèle ayant des propriétés intéressantes d'un point de vue interprétatif.

Notre travail vise à définir un modèle spécifique aux données ordinales sans avoir à adapter artificiellement un modèle issu du monde nominal ou du monde quantitatif. Pour ce faire, comme dans le cas des données de rang, nous modélisons le processus de génération d'une donnée ordinale, que nous considérons être un algorithme de recherche dichotomique dans une table ordonnée auquel on aurait adjoint une erreur aléatoire de décision à chaque itération. Le modèle résultant [R6] possède des propriétés remarquables et appréciables comme la présence d'un mode et d'une décroissance autour de ce mode, ou encore l'uniformité de la distribution pour certaines valeurs des paramètres. Ce modèle une fois défini, nous avons pu l'utiliser pour la classification automatique de données ordinales multivariées en ayant recours aux modèles de mélange ainsi qu'à l'hypothèse d'indépendance conditionnelle.

3.3.1 *Un modèle probabiliste pour données ordinales*

Une variable ordinale μ est une variable nominale à m modalités $\{1, \ldots, m\}$ vérifiant la relation d'ordre total $1 < \ldots < m$. On peut ainsi interpréter μ comme la réalisation d'un algorithme de recherche dans une liste ordonnée $(1, \ldots, m)$. Un algorithme de recherche s'appuie cependant sur des comparaisons successives et si des erreurs de comparaisons se produisent, une valeur x différente de μ peut être obtenue. Dans ce cas, la loi de x dépend alors de l'algorithme de recherche et aussi de la loi régissant le type d'erreur de comparaison. Nous faisons dans [R6] l'hypothèse que x est obtenue par un algorithme dichotomique, l'un des algorithmes de recherche les plus performants [67]. Trois niveaux d'aléatoire sont introduits à chaque itération : au niveau du point de comparaison, de la qualité de la comparaison et de l'intervalle retenu contenant potentiellement la valeur μ recherchée.
L'algorithme dichotomique retenu itère en m itérations pour obtenir une réalisation x. Partant d'un intervalle $e_{j-1} = \{b^-_{j-1}, \ldots, b^+_{j-1}\} \subset \{1, \ldots, m\}$, l'itération j ($2 \leq j \leq m$) se décompose elle-même en trois étapes aléatoires :

1. choix d'un point de comparaison $c_j \in e_j$;
2. choix de la qualité de comparaison $z_j \in \{0, 1\}$ pour comparer c_j avec μ : 0 correspond au cas où la comparaison est faite de façon aléatoire, 1 au cas où elle est parfaite ;
3. choix d'un nouvel intervalle $e_{j+1} \in \{e^-_j, e^=_j, e^+_j\}$, avec $e^-_j = \{b^-_j, \ldots, c_j - 1\}$, $e^=_j = \{c_j\}$, $e^+_j = \{c_j + 1, \ldots, b^+_j\}$, choix dépendant de c_j et de z_j.

On peut ainsi schématiser comme suit la séquence des e_j, c_j, z_j obtenus :

$$e_1 = \{1, \ldots, m\} \to c_1 \to z_1 \to e_2 \to \ldots \to c_{m-1} \to z_{m-1} \to e_m = \{x\}.$$

Un algorithme dichotomique standard (déterministe) impose toujours $z_j = 1$, c'est-à-dire que toute comparaison entre c_j et μ se fait sans erreur possible. L'originalité de l'algorithme proposé réside donc dans la possibilité de rendre l'issue de cette comparaison stochastique.

LE MODÈLE DSO. La loi de probabilité sur $x \in \{1, \dots, m\}$ est obtenue en associant un modèle probabiliste à chacune des trois étapes précédentes. Partant de $P(e_1) = 1$, on pose pour chacune des trois étapes précédentes, à l'itération j :

1. choix de c_j uniformément dans e_j : $P(c_j|e_j) = \frac{1}{|e_j|}\mathbb{1}_{\{c_j \in e_j\}}$;
2. choix de z_j selon une Bernoulli $\mathcal{B}(p)$ ($p \in [0,1]$) ;
3. le choix de e_{j+1} dépendra de la qualité de comparaison :
   - si la comparaison est aléatoire, l'intervalle e_{j+1} est choisi avec une probabilité proportionnelle à chacun des trois intervalles autorisés :

$$P(e_{j+1}|c_j, e_j, z_j = 0) = \frac{|e_{j+1}|}{|e_j|}\mathbb{1}_{\{e_{j+1} \in \{e_j^-, e_j^=, e_j^+\}\}};$$

   - si la comparaison est parfaite, l'intervalle contenant μ est retenu de façon certaine

$$P(e_{j+1}|c_j, e_j, z_j = 1; \mu) = \mathbb{1}_{\{e_{j+1} = \arg\min_{\{e_j^-, e_j^=, e_j^+\}} \delta(\cdot, \mu)\}}\mathbb{1}_{\{e_{j+1} \in \{e_j^-, e_j^=, e_j^+\}\}},$$

   où δ mesure l'écart suivant entre μ et un intervalle $e = \{b^-, \dots, b^+\}$

$$\delta(e, \mu) = \min(|\mu - b^-|, |\mu - b^+|).$$

À partir de ces éléments, on obtient la loi sur x en marginalisant sur les z_j

$$P(e_{j+1}|e_j, c_j; \mu, p) = pP(e_{j+1}|c_j, e_j, z_j = 1; \mu) + (1-p)P(e_{j+1}|c_j, e_j, z_j = 0)$$

puis sur les c_j

$$P(e_{j+1}|e_j; \mu, p) = \sum_{c_j \in e_j} P(e_{j+1}|e_j, c_j; \mu, p)P(c_j|e_j).$$

Utilisant la propriété markovienne des e_j on obtient

$$P(x; \mu, p) = \underbrace{\sum_{e_{m-1}} \left\{ P(e_m|e_{m-1}; \mu, p) \underbrace{\sum_{e_{m-2}} \left\{ P(e_{m-1}|e_{m-2}; \mu, p) \dots \underbrace{\sum_{e_1} \{P(e_2|e_1; \mu, p)P(e_1)\}}_{P(e_2;\mu,p)} \right\}}_{P(e_{m-1};\mu,p)} \right\}}_{P(e_m;\mu,p)}.$$

Le modèle ainsi formé est noté DSO pour *Dichotomic Search Ordinal* dans [R6].

FORME DE LA DISTRIBUTION. Le modèle proposé est identifiable et ses paramètres impactent sa forme de façon remarquable. La figure 8 illustre la forme de la distribution pour $m = 5$ et différentes valeur de μ et de p. On remarque que μ joue le rôle de pa-

ramètre de *position* tandis que p est un paramètre de *dispersion*. Ainsi, μ est le mode de la distribution, avec décroissance de façon monotone autour de μ. Ce mode est d'autant plus prononcé que p augmente. Les cas extrêmes $p = 0$ et $p = 1$ correspondent respectivement à une loi uniforme sur $\{1, \ldots, m\}$ et une loi totalement concentrée en μ (situation non représentée sur la figure). A noter que nous retrouvons ainsi naturellement toutes les propriétés qui avaient été introduites de façon artificielle dans le modèle CUB [35].

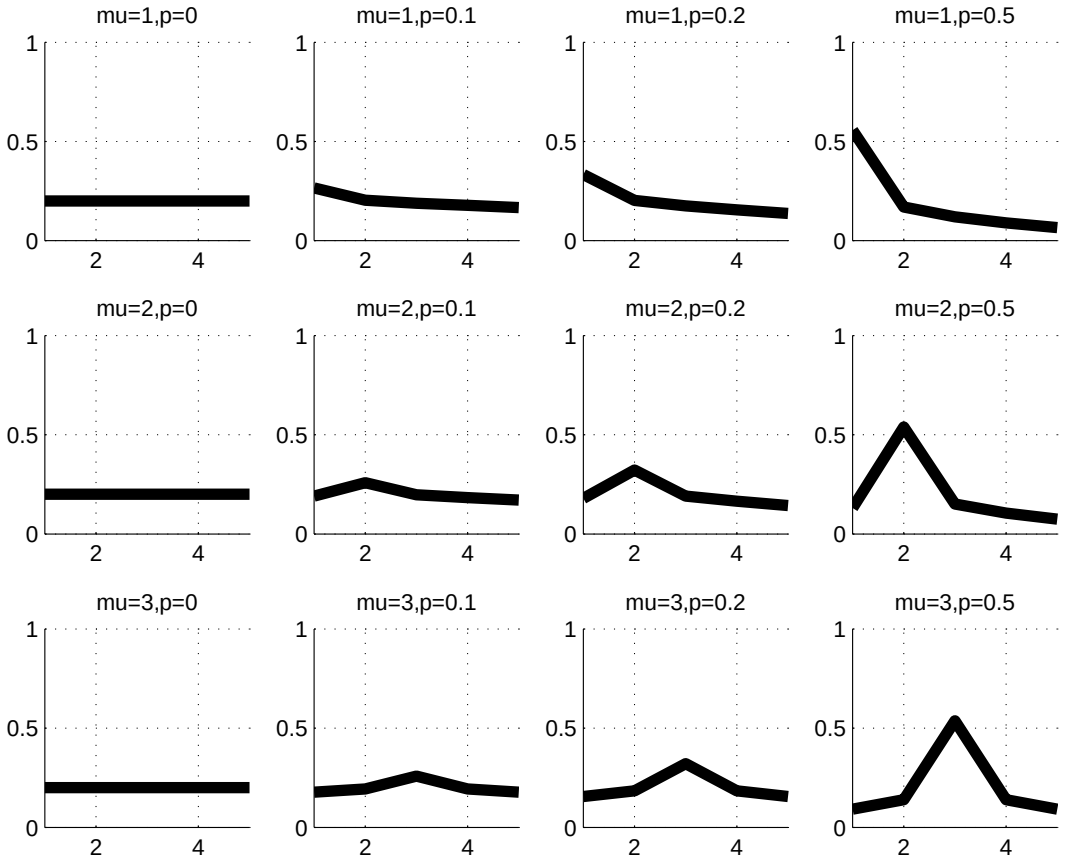


FIG. 8 – Illustration de la distribution pour différentes valeurs des paramètres μ et p.

ESTIMATION. Considérons un échantillon i.i.d. $(x_1, \ldots, x_n)$ provenant du modèle ordinal précédent. Puisqu'il s'agit d'un modèle à variables cachées, un algorithme EM [37] peut être utilisé pour estimer les paramètres par maximum de vraisemblance. On note $s_i = \{e_{ij}, c_{ij}, z_{ij}\}_{j=1,\ldots,m}$ les variables cachées (e_j, c_j, z_j) associées à l'individu x_i et S_i l'espace associé. À l'itération q, les deux étapes de l'algorithme s'écrivent alors comme suit :

– Étape E : pour tous les $s_i \in S_i$ $(i = 1, \ldots, n)$, on calcule des probabilités conditionnelles

$$P(s_i | x_i; \mu^{\{q\}}, p^{\{q\}}) = P(s_i, x_i; \mu^{\{q\}}, p^{\{q\}}) / P(x_i; \mu^{\{q\}}, p^{\{q\}}).$$

– Étape M : on cherche $\mu^{\{q+1\}} \in \{1, \ldots, m\}$ maximisant l'espérance conditionnelle de la log-vraisemblance complétée $\sum_{i=1}^{n} \sum_{s_i \in S_i} P(s_i|x_i; \mu^{\{q\}}, p^{\{q\}}) \ln P(x_i, s_i; \mu^{\{q+1\}}, p^{\{q+1\}})$ où

$$p^{\{q+1\}} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m-1} P(z_{ij} = 1|x_i; \mu^{\{q\}}, p^{\{q\}})}{n(m-1)}.$$

3.3.2 *Classification automatique de données ordinales multivariées*

Nous considérons désormais des variables aléatoires ordinales p-variées $\boldsymbol{x} = (x_1, \ldots, x_p)'$, provenant de K distributions DSO, telle que conditionnellement à la classe d'appartenance les p variables ordinales soient indépendantes (hypothèse d'indépendance conditionnelle) :

$$P(\boldsymbol{x}|y = k; \boldsymbol{\mu}_k, \boldsymbol{p}_k) = \prod_{j=1}^{p} P(x_j; \mu_{kj}, p_{kj})$$

où y est la variable indiquant le groupe d'appartenance, $\boldsymbol{\mu}_k = \{\mu_{k1}, \ldots, \mu_{kp})$et $\boldsymbol{p}_k = (p_{k1}, \ldots, p_{kp})'$. La distribution marginale de $\boldsymbol{x}$ est

$$P(\boldsymbol{x}; \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k P(\boldsymbol{x}|y = k; \boldsymbol{\mu}_k, \boldsymbol{p}_k)$$

où π_k est la proportion de la classe k et $\boldsymbol{\theta} = (\pi_k, \mu_{kj}, p_{kj})_{1 \leq j \leq p, 1 \leq k \leq K)}$.
L'estimation de ce modèle peut également être réalisée par maximum de vraisemblance à l'aide de l'algorithme EM :

– Étape E : calcul des probabilités conditionnelles $P(y_i|\boldsymbol{x}_i; \boldsymbol{\pi}^{\{q\}}, \boldsymbol{\mu}^{\{q\}}, \boldsymbol{p}^{\{q\}})$
– Étape M :
  – $p_k^{\{q+1\}} \propto \sum_{i=1}^{n} P(y_i = k|\boldsymbol{x}_i; \boldsymbol{\pi}^{\{q\}}, \boldsymbol{\mu}^{\{q\}}, \boldsymbol{p}^{\{q\}})$
  – pour tout (k, j), $(\mu_{kj}, p_{kj})^{\{q+1\}}$ sont estimés par un algorithme EM similaire à celui utilisé dans le cas DSO univarié sans mélange.

3.3.3 *Expérimentations numériques*

Afin d'illustrer le bon comportement du modèle DSO pour la classification de données ordinales multivariées, nous analysons les données de l'évaluation des Licences universitaires de mars 2011 par l'AÉRES (Agence d'Évaluation de la Recherche et de l'Enseignement Supérieur). Au cours de cette évaluation, 23 universités françaises des académies de Bordeaux, Toulouse, Lyon, Montpellier et Grenoble ont été évaluées selon 4 critères : pilotage (PT), projet pédagogique (EP), dispositifs d'aide à la réussite (SS) et insertion professionnelle et poursuite d'étude choisie (EFS). Chaque critère est évalué par une note ordinale {A+, A, B, C}. La table 8 donne les résultats de cette évaluation.

Une classification en 1 à 6 classes a été réalisée sur ces données, à l'aide du modèle DSO, mais également du modèle multinomial (approximation nominale des variables) et du

Université	PT	EP	SS	EFS
Bordeaux 1	A	A	A	B
Bordeaux 2	A+	A	A+	A
Bordeaux 3	B	A	B	B
Bordeaux 4	B	A	A+	A
Pau	C	B	B	C
Toulouse 1	B	B	B	B
Toulouse 2	B	B	A	B
Toulouse 3	A	A	A+	A
Champollion	A	B	B	B
Lyon 1	A	A+	A	A
Lyon 2	B	A	B	B
Lyon 3	B	A+	B	B
St Etienne	A	B	A	B
Montpellier 1	B	A	A	B
Montpellier 2	A	A	A	B
Montpellier 3	B	B	A	B
Nîmes	C	B	C	C
Perpignan	B	B	B	B
Grenoble 1	B	B	A+	A
Grenoble 2	A	A	B	B
Grenoble 3	C	B	B	C
Savoie	A	A	A	B

TAB. 8 – Notes de l'évaluation des Licences universitaires par l'AÉRES (mars 2011).

modèle gaussien diagonal (approximation normale). Les valeurs du critère BIC, données dans la table 9, indiquent que le meilleur des modèles est le modèle DSO avec 4 classes, tandis que le modèle gaussien et le modèle multinomial donnent tous deux une unique classe.

Model	$g = 1$	$g = 2$	$g = 3$	$g = 4$	$g = 5$	$g = 6$
Ordinal	-111.90	-109.14	-107.80	**-104.25**	-108.49	-114.28
Nominal	-108.37	-111.50	-120.08	-135.01	-151.84	-169.75
Gaussian	-105.76	-109.31	NaN	NaN	NaN	NaN

TAB. 9 – Valeurs du critère BIC pour la classification automatique des données AÉRES en 1 à 6 classes, pour les modèles DSO, multinomial et gaussien diagonal.

La répartition fournie par le modèle DSO à 4 classes est illustrée par la figure 9, représentant les projections des points sur le premier plan factoriel de l'analyse des correspondances multiples (ACM). Les 4 classes obtenues ont chacune une interprétation cohérente : la classe 1 ($\hat{\mu}_1 = (A, A, A, B)$) contient des universités ayant des scores élevés homogènes sur les 4 critères, la classe 2 ($\hat{\mu}_2 = (B, A, A+, C)$) a des scores contrastés, la classe 3 ($\hat{\mu}_3 = (B, B, B, B)$) a des scores moyens homogènes, tandis que la classe 4 a des scores faibles ($\hat{\mu}_4 = (C, B, B, C)$).

Cette application illustre l'intérêt du modèle DSO vis-à-vis des modèles assimilant les données ordinales à des données nominales ou continues.

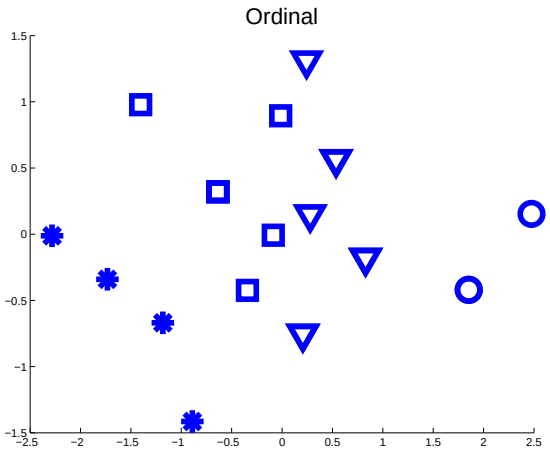


FIG. 9 – Répartition en 4 classes des données AÉRES fournie par le modèle DSO sur le premier plan factoriel de l'ACM.

4

MODÈLE DE MÉLANGE POUR DONNÉES FONCTIONNELLES ET CLASSIFICATION AUTOMATIQUE DE COURBES

Sommaire

Revues avec comité de lecture

[R10] Jacques, J. and Preda, C. *Funclust : a curves clustering method using functional random variable density approximation*, Pub. IRMA Lille Vol. 71-I, Preprint HAL n°00 628247, en révision pour Neurocomputing.

[R11] Jacques, J. and Preda, C. *Clustering of multivariate functional data*, Preprint HAL n°00 713334, en révision pour Computational Statistics and Data Analysis.

[R12] Bouveyron, C. and Jacques, J. *Model-based Clustering of Time Series in Group-specific Functional Subspaces*, Advances in Data Analysis and Classification, 5 (4), 281–300, 2011.

Conférences internationales avec comité de lecture

[CI3] Jacques, J. and Preda, C. *Clustering multivariate functional data*. 20th International Conference on Computational Statistics (COMPSTAT'12), Limassol, Chypre, août 2012.

[CI4] Jacques, J. and Preda, C. *Curves clustering with approximation of the density of functional random variables*. 20th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN'12), Bruges, avril 2012.

[CI5] Bouveyron C. and Jacques, J.*Model-based Clustering of Time Series in Group-specific Functional Subspaces*. 12th annual conference of the International Federation of Classification Societies, Frankfurt, Germany, 2011.

Conférences nationales

[CN11] Jacques, J. and Preda, C. *Functional data clustering using density approximation*. 44èmes Journées de Statistique organisée par la Société Française de Statistique, Bruxelles, mai 2012.

[CN12] Jacques, J. *Model-based clustering of functional data*. Astrostatistique en France, Grenoble, décembre 2011.

4.1 PROBLÉMATIQUE

Nous nous intéressons dans ce chapitre à l'analyse des données fonctionnelles [88], et plus particulièrement à la définition d'algorithmes de classification automatique pour variables aléatoires fonctionnelles univariées (l'unité statistique est une courbe $x \in \mathcal{X}$) et multivariées (l'unité statistique est un ensemble de p courbes $\boldsymbol{x} = (x_1, \ldots, x_p)' \in \mathcal{X}$). Une des principales difficultés inhérentes à ce type de données est la dimension infinie de l'espace $\mathcal{X}$. Ainsi, le principal obstacle à la définition d'algorithme de classification à base de modèles de mélange probabilistes est que la notion de densité de probabilité n'est en général pas définie sur un tel espace [34].

Face à un problème de classification non supervisée de courbes, l'approche la plus simple consiste à transformer le problème de dimension infinie en un problème de dimension finie. Plusieurs approches sont possibles : travailler directement avec une discrétisation temporelle des courbes, qui de toute façon sont toujours observées en un nombre fini d'instants, avec une approximation de celles-ci dans une base de fonctions de dimension finie, ou encore avec les composantes principales résultant d'une analyse en composantes principales fonctionnelles (ACPF ou ACP fonctionnelle). Les courbes étant ainsi décrites par un nombre fini d'objets, des algorithmes de classification automatique usuels peuvent être utilisés, tels les modèles de mélange gaussiens : des simples k-means [55] aux mélanges gaussiens parcimonieux [7, 27] en passant par les modèles spécifiques à la grande dimension [19, 76, 102]. Ainsi, Abraham *et al.* [1] appliquent la méthode des k-means sur l'approximation des courbes dans une base de B-splines, Tarpey et Kinateder [100] sur certains points particuliers des courbes, et Chiou et Li [29] sur les premières composantes de l'ACP fonctionnelle. Des paramétrisations plus spécifiques des courbes peuvent également être utilisées : James et Sugar [64] définissent une approche particulièrement efficace lorsque les courbes ne sont pas observées aux même instants, Ray et Mallick [89] proposent un modèle basé sur un mélange de processus de Dirichlet, et Frühwirth-Schnatter et Kaufmann [45] construisent un algorithme de classification automatique spécifique aux séries temporelles. L'avantage de la méthode des k-means, même si elle peut paraître très contraignante lorsqu'on l'utilise avec la distance euclidienne puisqu'elle correspond à un modèle de mélange gaussien imposant des matrices de variances sphériques identiques par groupes, est qu'elle peut être utilisée avec d'autres distances et notamment des distances spécifiques aux variables aléatoires fonctionnelles. La classification hiérarchique [106] a également le même avantage. Ainsi, Ferraty et Vieu [42] définissent un algorithme de classification hiérarchique descendante utilisant une semi-métrique basée sur les dérivées des fonctions, tandis que Ieva *et al.*[63] utilisent la méthodes des k-means dans le cas de courbes multivariées en utilisant ces mêmes semi-métriques.

D'un autre coté, des travaux plus théoriques s'intéressent à la notion de densité de probabilité pour une variable aléatoire fonctionnelle. Un certain nombre d'entre eux ont conduit à définir des estimateurs non paramétriques de la densité de probabilité sous l'hypothèse d'existence d'une mesure pour laquelle le processus étudié admette une densité [31, 32, 42]. Plus récemment, Delaigle et Hall [34] ont montré que la notion de densité pouvait être approchée par la densité de probabilité des composantes principales de l'ACP fonctionnelle. C'est ce dernier travail qui a motivé les approches que nous développons

dans ce chapitre. En effet, nous proposons dans [R10] un premier modèle de classification automatique de courbes basé sur une approximation de la notion de densité de probabilité de variables aléatoires fonctionnelles, reposant sur la densité des premières composantes principales. Des versions parcimonieuses de ce modèle ont également été publiées dans [R12]. Enfin, une extension au cas de variables aléatoires fonctionnelles multivariées à été proposée dans [R11]. Dans ce dernier article, nous avons également décrit en détail l'ACP fonctionnelle multivariée (ACPFM) ainsi qu'une version normalisée permettant de traiter des courbes de natures différentes (courbes de précipitations et de température par exemple). La suite de ce chapitre reprend ces travaux dans l'ordre dans lequel ils viennent d'être présentés.

4.2 CLASSIFICATION AUTOMATIQUE DE COURBES

Soit X un processus stochastique, ou variable aléatoire fonctionnelle, à valeur dans $L_2([0,T])$, $T > 0$. Nous supposons dans ce chapitre que X est un processus stochastique L_2-continu, $X = \{X(t),\ t \in [0,T]\}$. La décomposition Karhunen-Loeve de ce processus s'écrit :

$$X(t) = \mu(t) + \sum_{j=1}^{\infty} C_j f_j(t), \tag{12}$$

où μ est la fonction moyenne, $C_j = \int_0^T (X(t) - \mu(t)) f_j(t) dt$ sont les *composantes principales* et où les f_j forment un système orthonormé de *fonctions propres* de l'opérateur de covariance

$$\int_0^T Cov(X(t), X(s)) f_j(s) ds = \lambda_j f_j(t), \quad \forall t \in [0,T].$$

Les composantes principales C_j sont des variables aléatoires réelles non corrélées, d'espérance nulle et de variance λ_j ($\lambda_1 \geq \lambda_2 \geq \ldots$).

4.2.1 *Approximation de la densité de probabilité d'une variable aléatoire fonctionnelle*

On peut définir une *approximation $X^{(q)}$ d'ordre q* de X en tronquant (12) de sorte à ne garder que les q premiers termes de la somme :

$$X^{(q)}(t) = \mu(t) + \sum_{j=1}^{q} C_j f_j(t). \tag{13}$$

A partir de cette approximation, qui est la meilleure approximation en moyenne quadratique de X sous cette forme, Delaigle et Hall [34] montrent que la probabilité que X soit dans une petite boule de rayon h autour de $x \in L_2[0,T]$ s'écrit

$$\log P(\|X - x\| \leq h) = \sum_{j=1}^{q} \log f_{C_j}(c_j(x)) + \xi(h, q(h)) + o(q(h)), \tag{14}$$

où f_{C_j} est la densité de probabilité de C_j et $c_j(x)$ la valeur de la $j$ème composante principale pour $X = x$ donnée par $c_j(x) =< x, f_j >_{L_2}$. Les fonctions $q(h)$ et ξ sont telles que $q(h)$ tend vers l'infini lorsque h tend vers 0 et ξ est une constante ne dépendant que de h et $q(h)$. Étant donné que la notion de densité de probabilité peut être vue comme la limite de cette probabilité lorsque h tend vers 0, et sachant que toutes les variations de $\log P(\|X - x\| \leq h)$ avec x sont contenues dans le terme $\sum_{j=1}^{q} \log f_{C_j}(c_j(x))$, nous avons proposé d'utiliser la densité de probabilité suivante comme approximation de la notion de densité de X :

$$f_X^{(q)}(x) = \prod_{j=1}^{q} f_{C_j}(c_j(x)). \tag{15}$$

Dans la suite, nous ferons l'hypothèse supplémentaire que les f_{C_j} sont des densités gaussiennes, ce qui est notamment le cas lorsque X est un processus gaussien. Disposant d'une approximation de la densité de X nous pouvons définir un modèle de mélange qui nous servira pour déterminer un algorithme de classification automatique de données fonctionnelles.

4.2.2 *Funclust : un modèle de mélange pour la classification automatique de courbes*

Comme dans les chapitres précédents nous appelons Y la variable de groupe, $Y = k$ indiquant que l'individu décrit par la variable X appartient au $k$ème groupe. L'objectif de la classification automatique est de prédire la variable Y à partir de l'observation de X.

Le modèle

Nous supposons que conditionnellement à l'appartenance au $k$ème groupe, l'approximation d'ordre q_k de la densité de $X_{|Y=k}$ est :

$$f_{X_{|Y=k}}^{(q_k)}(x; \Sigma_k) = \prod_{j=1}^{q_k} f_{C_{j|Y=k}}(c_{jk}(x); \lambda_{jk})$$

où $c_{jk}(x)$ est la valeur de la $j$ème composante principale de $X_{|Y=k}$ pour $X = x$, $f_{C_{j|Y=k}}$ la densité de probabilité de la loi normale, d'espérance nulle et de variance λ_{jk}, et $\mathbf{\Sigma}_k = diag(\lambda_{1k}, \ldots, \lambda_{q_k k})$ la matrice diagonale de dimension $q_k \times q_k$ contenant les variances des

composantes principales. Une approximation de la densité marginale de X est fournie par la densité mélange suivante

$$f_X^{(q)}(x;\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{q_k} f_{C_{j|Y=k}}(c_{jk}(x);\lambda_{jk}) \tag{16}$$

où $\boldsymbol{\theta} = (\pi_k, \boldsymbol{\Sigma}_k)_{1\leq k\leq K}$ sont les paramètres du modèle à estimer. Par extrapolation du cas fini dimensionnel nous définissons dans [R10] la *pseudo-vraisemblance* suivante :

$$l^{(q)}(\boldsymbol{\theta};\boldsymbol{X}) = \prod_{i=1}^{n}\sum_{k=1}^{K} \pi_k \prod_{j=1}^{q_k} \frac{1}{\sqrt{2\pi\lambda_{jk}}} \exp -\frac{1}{2}\left(\frac{C_{ijk}}{\lambda_{jk}}\right)^2 \tag{17}$$

où $C_{ijk} = C_{jk}(X_i)$ est la valeur de la $j$ème composante principale de la courbe X_i appartenant au $k$ème groupe, et $\boldsymbol{X} = (X_1, \ldots, X_n)'$ un échantillon i.i.d.

Estimation

Nous proposons dans [R10] un algorithme de type EM pour maximiser (17). Si l'étape E est classique, l'étape M aura la spécificité d'intégrer le calcul des composantes principales en fonction des probabilités d'appartenances des observations aux classes estimées à l'étape E ainsi que la sélection des ordres d'approximation q_k.

Comme cela est décrit dans [88], le calcul des composantes principales de l'ACP fonctionnelle nécessite généralement des approximations, dont la plus courante consiste à supposer que les courbes admettent une décomposition dans une base de fonctions $\Phi = (\phi_1, \ldots, \phi_L)$:

$$X_i(t) \simeq \sum_{j=1}^{L} \gamma_{ij}\phi_j(t). \tag{18}$$

Outre cette nécessité calculatoire, cette hypothèse permet de reconstruire la nature fonctionnelle des données, qui dans la pratique sont toujours observées de façons discrétisées. Le choix de la base, tout comme celui du nombre de fonctions de base, est une problématique cruciale à laquelle nous ne nous sommes pas encore intéressés. Une façon simple d'aborder les choses est de voir ce choix comme un choix de modèles, ce que propose entre autres [80] dans un cadre de régression sur variables aléatoires fonctionnelles approchées par une base de splines. Dans les applications traitées dans ce document, le choix de la base sera fait empiriquement à partir de l'observation des données discrétisées.

Décrivons désormais l'algorithme d'estimation que nous proposons pour maximiser (17). Soit $\boldsymbol{\theta}^{(h)}$ l'estimation courante du paramètre $\boldsymbol{\theta}$, $h \geq 1$.

ÉTAPE E. Cette étape revient à calculer les probabilités t_{ik} pour que la $i$ème courbe X_i appartienne au groupe k conditionnellement à $C_{ijk}^{(h)} = c_{ijk}^{(h)}$:

$$t_{ik}^{(h)} \simeq \frac{\pi_k^{(h)} \prod_{j=1}^{q_k^{(h)}} f_{C_{j|Y_i=k}}(c_{ijk}^{(h)})}{\sum_{l=1}^{K} \pi_l^{(h)} \prod_{j=1}^{q_l^{(h)}} f_{C_{j|Y_i=l}}(c_{ijl}^{(h)})}.$$

ÉTAPE M : MISE À JOUR DES COMPOSANTES PRINCIPALES. Sous l'hypothèse que les courbes observées admettent une décomposition dans une base de fonctions (équation (18)), le calcul des composantes principales de chaque groupe est réalisé en pondérant les individus en fonction des probabilités $t_{ik}^{(h)}$. En notant Γ la matrice $n \times L$ des coefficients γ_{ij}, la première étape consiste donc à centrer les courbes au sein du groupe k en retranchant la courbe moyenne obtenue en utilisant les pondérations $t_{ik}^{(h)}$:

$$\Gamma_k^{(h)} = (\boldsymbol{I}_n - \mathbf{1}_n(t_{1k}^{(h)}, \ldots, t_{nk}^{(h)}))\Gamma,$$

où $\boldsymbol{I}_n$ et $\mathbf{1}_n$ sont respectivement la matrice identité $n \times n$ et le vecteur unité de taille n. La $j$ème composante principale $C_{jk}^{(h)}$ est alors le $j$ème vecteur propre de la matrice $\Gamma_k^{(h)} W {\Gamma_k^{(h)}}' T_k^{(h)}$ associé à la $j$ème valeur propre :

$$\Gamma_k^{(h)} W {\Gamma_k^{(h)}}' T_k^{(h)} C_{jk}^{(h)} = \lambda_{jk}^{(h)} C_{jk}^{(h)},$$

où W est la matrice des produits scalaires des fonctions de base deux à deux $w_{j\ell} = \int_0^T \phi_j(t)\phi_\ell(t)dt$ $(1 \leq j, \ell \leq L)$.

ÉTAPE M : SÉLECTION DES ORDRES D'APPROXIMATION $q_k^{(h)}$. La sélection des ordres d'approximation est toujours à l'heure actuelle un problème ouvert. Nous préconisons dans [R10] l'utilisation du test du coude de Cattell [24], qui a l'avantage de permettre de régler tous les ordres d'approximation avec un unique hyper-paramètre (le seuil du test de Cattell).

ÉTAPE M : MISE À JOUR DES PARAMÈTRES. Les variances $\lambda_{jk}^{(h+1)}$ sont celles calculées lors de l'étape de mise à jour des composantes principales, quant aux proportions, elles sont estimées comme d'habitude par $\pi_k^{(h+1)} = \frac{1}{n}\sum_{i=1}^{n} t_{ik}^{(h)}$.

CRITÈRE D'ARRÊT. Le fait de régler les ordres d'approximation au sein de l'algorithme d'estimation conduit à des variations dans la vraisemblance dues à une augmentation ou une diminution d'un ou plusieurs ordres d'approximation. L'algorithme utilisé n'assure donc pas nécessairement la croissance de la pseudo-vraisemblance. Nous avons donc proposé d'itérer l'algorithme pendant un nombre conséquent d'itérations, puis de retenir la solution $\theta^{(h)}$ de pseudo-vraisemblance maximale.

modèle	Nb. de param.	Nb. de param. $K = 4$, $d_k = 10$, $p = 100$
$[a_{kj}b_k]$	$\rho + \tau + 2K + D$	4231
$[a_{kj}b]$	$\rho + \tau + K + D + 1$	4228
$[a_k b_k]$	$\rho + \tau + 3K$	4195
$[ab_k]$	$\rho + \tau + 2K + 1$	4192
$[a_k b]$	$\rho + \tau + 2K + 1$	4192
$[ab]$	$\rho + \tau + K + 2$	4189

TAB. 10 – Liste des sous-modèles de FunHDDC ainsi que leur nombre de paramètres.

4.2.3 *FunHDDC : définition de modèles parcimonieux*

Nous avons proposé une approche différente dans [R12], en poussant l'ordre d'approximation aussi loin que possible et en proposant une modélisation parcimonieuse de la matrice Σ_k des variances des composantes principales. Nous avons vu précédemment que le calcul des composantes principales nécessitait l'approximation des courbes observées dans une base de fonction $\phi = (\phi_1, \ldots, \phi_L)$. Ainsi, chaque courbe est décrite par une série de L coefficients, et il ne sera donc possible d'obtenir qu'au maximum L composantes principales. Nous proposons donc dans [R12] d'utiliser ces L composantes dans l'approximation (16), puis de régulariser la matrice des variances $\mathbf{\Sigma}_k$ en s'inspirant de la méthode HDDC [19] :

$$\mathbf{\Sigma}_k = \begin{pmatrix} \boxed{\begin{matrix} a_{k1} & & 0 \\ & \ddots & \\ 0 & & a_{kd_k} \end{matrix}} & & \mathbf{0} \\ & \boxed{\begin{matrix} b_k & & 0 \\ & \ddots & \\ & & \ddots \\ 0 & & b_k \end{matrix}} & \\ \mathbf{0} & & \end{pmatrix} \begin{matrix} \left.\vphantom{\begin{matrix} a \\ a \\ a \end{matrix}}\right\} d_k \\ \left.\vphantom{\begin{matrix} a \\ a \\ a \end{matrix}}\right\} (L - d_k) \end{matrix} \tag{19}$$

avec $a_{kj} > b_k$ pour $j = 1, ..., d_k$. D'un point de vue pratique, les $a_{k1}, ..., a_{kd_k}$ modélisent la variance des composantes principales dans le $k$ème groupe, tandis que b_k modélise la variance du bruit dans ce groupe. Ce modèle est noté $[a_{kj}b_k]$ dans la suite. Suivant la stratégie définie dans [19], nous avons proposé un certain nombre de contraintes supplémentaires sur les paramètres a_{kj} et b_k, afin de définir des sous-modèles parcimonieux. Ils sont résumés dans la table 10 ; par exemple, le modèle $[a_k b]$ considère qu'au sein d'un même groupe les composantes principales ont la même variance, et que la variance du bruit est commune à tous les groupes.

Estimation

L'algorithme d'estimation des modèles FunHDDC est relativement similaire à celui utilisé par Funclust. Nous détaillons ici les principales différences, l'algorithme complet étant détaillé dans [R12]. L'étape E est quasiment identique, sauf qu'elle ne requiert l'utilisation que des d_k premières composantes principales. Ainsi, au sein de l'étape U, il est

inutile de calculer les composantes principales (ainsi que leur variance) au delà du rang d_k. L'étape S, ne consiste plus à estimer les ordres d'approximation q_k, mais les seuils d_k au dessus duquel les composantes sont supposées correspondre à du bruit. La technique du test de Cattell reste néanmoins utilisée. Enfin, concernant l'étape M, les variances sont estimées de façon spécifique au modèle utilisé : par exemple, pour le modèle $[a_{kj}b_k]$, les d_k premières variances sont estimées comme pour Funclust, tandis que les $L - d_k$ restantes sont estimées en utilisant l'estimation de la trace de $\mathbf{\Sigma}_k$.

Un avantage de ce modèle par rapport à Funclust, outre le fait de proposer un panel de modèles plus ou moins parcimonieux, est qu'il est possible d'utiliser des critères de choix de modèles comme BIC pour sélectionner à la fois le meilleur des sous-modèles ainsi que le nombre K de groupes.

4.3 Analyse des données fonctionnelles multivariées

Nous nous sommes également intéressés dans [R11] au cas des variables aléatoires fonctionnelles multivariées $\mathbf{X} = (X_1, \ldots, X_p)$, que l'on rencontre lorsqu'un individu est décrit par plusieurs courbes. Plusieurs exemples de ce type de données sont décrits dans [88] : les courbes de température et de précipitations des villes canadiennes (données *CanadianWeather* du package *fda* de **R**), ou encore les courbes d'angle de flexion du genou et de la hanche lors du cycle de la marche (données *gait* du package *fda*).

Afin de développer un algorithme de classification automatique pour de telles données sur la même base que ceux développés dans le cas de courbes univariées, nous avons dans un premier temps présenté l'analyse en composantes principales fonctionnelles multivariée puis défini un modèle de mélange basé sur une approximation de la densité des variables aléatoires fonctionnelles multivariées.

4.3.1 *Analyse en composantes principales fonctionnelles multivariée (ACPFM)*

L'analyse en composantes principales pour les données fonctionnelles multivariées a déjà été abordée dans la littérature [11, 88]. Dans [88], les auteurs proposent de concaténer les observations discrétisées des fonctions en un long vecteur (ou les coefficients dans une certaine base de fonctions), et de réaliser une ACP classique sur ce vecteur. Le principal point faible de cette méthode est que cela oblige l'utilisation de bases orthonormées puisque la métrique induite par la base de fonctions n'est pas prise en compte. Nous étendons cette technique dans [R11] au cas de bases non orthonormales et potentiellement différentes pour chaque composante de la variable aléatoire fonctionnelle multivariée.

Supposons que $\boldsymbol{X}$ est un processus continu à valeurs dans $L_2([0,T])^p$. On note

$$\mu = (\mu_1, \ldots \mu_p)' = \mathbb{E}[\mathbf{X}],$$

la fonction moyenne de $\mathbf{X}$. L'opérateur de covariance de $\mathbf{X}$ est défini comme un opérateur intégral C de noyau

$$C(t,s) = \mathbb{E}\left[(\mathbf{X}(t) - \mu(t)) \otimes (\mathbf{X}(s) - \mu(s))\right],$$

où $\otimes$ est le produit tensoriel sur $\mathbb{R}^p$. Ainsi, $C(t,s)$ est une matrice $p \times p$ contenant les éléments $C(t,s)[i,j] = Cov(X_i(t), X_j(s)), \quad i,j = 1,\dots p$.
L'analyse spectrale de C fournit un ensemble $\{\lambda_j\}_{j\geq 1}$ de valeurs propres positives ou nulles, associées à une base de fonctions propres orthonormales $\{\mathbf{f}_j\}_{j\geq 1}$, $\mathbf{f}_j = (f_{j1},\dots,f_{jp})$, appelés facteurs principaux :

$$C\mathbf{f}_j = \lambda_j \mathbf{f}_j, \tag{20}$$

avec $\lambda_1 \geq \lambda_2 \geq \dots$ et $\langle \mathbf{f}_i, \mathbf{f}_j \rangle_{L_2([0,T])^p} = \int_0^T \sum_{\ell=1}^{p} f_{i\ell}(t) f_{j\ell}(t) = \delta_{ij}$ où $\delta_{ij} = 1$ si $i = j$ et 0 sinon.
Les composantes principales C_j de $\mathbf{X}$ sont des variables aléatoires réelles d'espérance nulle, définies comme les projections de $\mathbf{X}$ sur les fonctions propres de C :

$$C_j = \int_0^T \langle \mathbf{X}(t) - \mu(t), \mathbf{f}_j(t) \rangle_{\mathbb{R}^p} dt = \int_0^T \sum_{\ell=1}^{P} (X_\ell(t) - \mu_\ell(t)) f_{j\ell}(t) dt.$$

Comme dans le cas univarié, le calcul des composantes C_j ainsi que des variances λ_j nécessite de supposer par exemple que chaque courbe $x_{i\ell}$ ($1 \leq \ell \leq p$) admet une décomposition dans une base de fonctions $\Phi_\ell = (\phi_{\ell 1},\dots,\phi_{\ell L_\ell})$:

$$x_{i\ell}(t) = \sum_{j=1}^{L_\ell} \gamma_{i\ell j} \phi_{\ell j}(t).$$

Nous montrons alors dans [R11] que les composantes principales C_j sont solutions du problème aux valeurs propres suivant :

$$\frac{1}{n-1} \mathbf{A} W \mathbf{A}' C_j = \lambda_j C_j,$$

où $\mathbf{A} = (I_n - \mathbb{1}_n(1/n,\dots,1/n))\tilde{\mathbf{A}}$ est la version centrée de la matrice $\tilde{\mathbf{A}}$ de taille $n \times \sum_{\ell=1}^{p} L_\ell$ contenant les coefficients de la décomposition des courbes p-variées dans les bases $(\Phi_1,\dots,\Phi_p)$, W étant la matrice $\sum_{\ell=1}^{p} L_\ell \times \sum_{\ell=1}^{p} L_\ell$ symétrique bloc-diagonale des produits scalaires entre les fonctions de bases.

ACPFM NORMÉE. Lorsque les composantes de la variable aléatoire fonctionnelle multivariée observée sont de natures différentes, il est utile de normaliser les données comme cela est fait dans le cas fini dimensionnel. Nous montrons dans [R11] que l'ACPFM normée d'une variable aléatoire $\mathbf{X}(t)$ revient à l'ACPFM de

$$\mathbf{Y}(t) = R(t,t)^{-1} \mathbf{X}(t),$$

où $R(t,t)$ est défini par :

$$C(t,t) = R(t,t) R(t,t)'.$$

En pratique, si $\mathbf{X}$ est observé à des temps discrets $t_1, \ldots, t_r$, $r > 1$, alors $\mathbf{Y}$ est défini à partir de $\mathbf{X}$ par

$$\mathbf{Y}(t_i) = R(t_i, t_i)^{-1}\mathbf{X}(t_i), \quad \forall i = 1, \ldots, r.$$

4.3.2 *MFunclust : Classification automatique de courbes multivariées*

Par extension du cas univarié, nous développons dans [R11] un algorithme de classification automatique de courbes multivariées basé sur la densité mélange approchée de $\mathbf{X}$ suivante :

$$f_{\mathbf{X}}^{(\boldsymbol{q})}(\boldsymbol{x}; \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{q_k} f_{C_{j|Y=k}}(c_{jk}(\boldsymbol{x}); \lambda_{jk}) \tag{21}$$

où $\boldsymbol{\theta} = (\pi_k, \lambda_{1k}, \ldots, \lambda_{q_g k})_{1 \leq k \leq K}$ et $\boldsymbol{q} = (q_1, \ldots, q_K)$. Un des intérêts de l'approche que nous proposons, basée sur l'analyse en composantes principales fonctionnelles multivariée, est qu'elle permet de prendre en compte une éventuelle dépendance entre les différentes composantes des courbes multivariées, ce qui n'aurait pas été possible en étendant le modèle proposé pour les courbes univariées à l'aide d'une hypothèse d'indépendance conditionnelle.
L'algorithme d'estimation est similaire à celui utilisé dans le cas univarié, en effectuant au sein de l'étape U le calcul des composantes principales ainsi que de leur variance selon la méthodologie présentée dans la section 4.3.1.

4.3.3 *Expérimentations numériques*

Dans cette section, nous comparons dans un premier temps les modèles proposés à ceux de la littérature, en se concentrant sur le cas univarié pour lequel le nombre de méthodes concurrentes est plus important. Nous avons néanmoins montré dans [R11], sur données simulées et réelles (non reportées ici), que notre algorithme de classification automatique de courbes multivariées (MFunclust) donnait des résultats supérieurs à ceux de la principale méthode concurrente [63]. Dans un second temps nous comparons les trois modèles que nous avons proposés sur le jeu de données *CanadianWeather*, en comparant les classifications obtenues par FunHDDC et Funclust sur les courbes de température uniquement avec celle obtenue par MFunclust en utilisant à la fois les courbes de température et de précipitations.

Benchmark sur courbes univariées

Nous considérons dans cette section quatre jeux de données : *Kneading*, *Growth*, *ECG* et *CBF*. Le jeu de données Kneading, provenant de Danone Vitapole Paris Research Center et déjà étudié par [71, 87], consiste en 115 courbes de dureté de pâte à biscuit au cours du processus de pétrissage, observées en 241 instants de temps équidistants. Chacune de ces pâtes a donné un biscuit après cuisson, et parmi les 115 biscuits, 50 ont été jugés par des experts de bonne qualité, 25 de qualité moyenne et 40 de mauvaise qualité. Le jeu de

données Growth provient de l'étude *Berkeley growth study* [103] et est disponible dans le package *fda* de **R**. Il consiste en les courbes de croissance de 54 filles et 39 garçons mesurés 31 fois entre 1 et 18 ans. Les jeux de données ECG et CBF sont issus du site *UCR Time Series Classification and Clustering*[1]. ECG consiste en 200 électrocardiogrammes mesurés sur 96 pas de temps pour deux groupes de patients, et CBF est constitué de 930 courbes observées à 128 pas de temps réparties en 3 groupes. Ces jeux de données sont présentés par la figure 10.

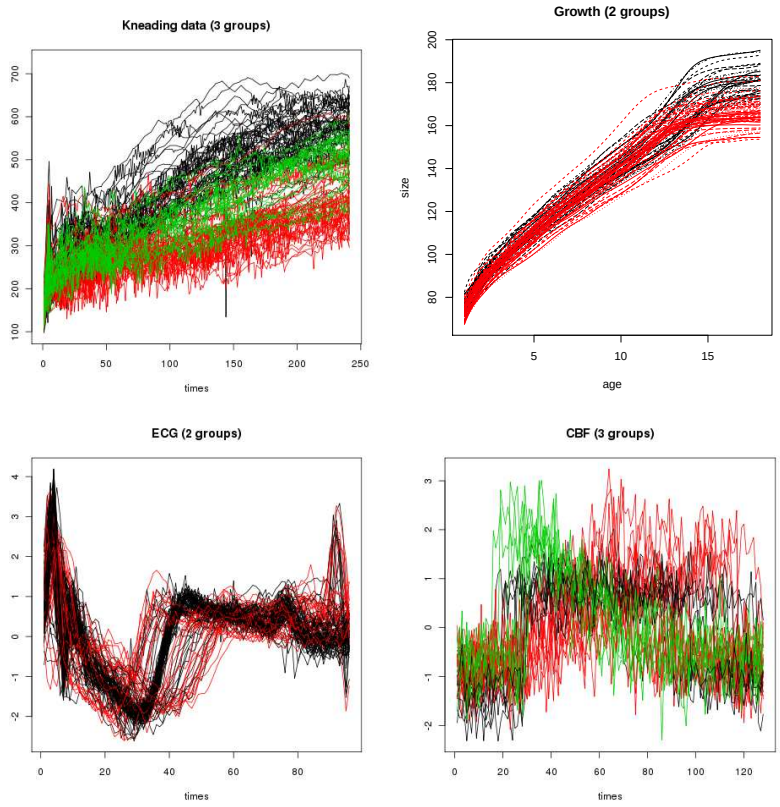


FIG. 10 – Jeux de données *Kneading*, *Growth*, *ECG* et *CBF*.

Les méthodes FunHDDC et Funclust sont comparées à un panel de techniques concurrentes, sur la base de la correspondance entre la classification fournie et celle existant dans les données (qualité du biscuit pour les données Kneading, sexe pour les données Growth...) : fclust [64], kCFC [29] qui sont utilisables directement sur les données fonctionnelles, ainsi que HDDC [19], MixtPPCA [102], mclust [7], k-means et hclust (classification hiérarchique) appliquées soit sur les courbes discrétisées, soit sur les coefficients des approximations des courbes par une base de splines, soit sur les composantes de l'ACP fonctionnelle. Les résultats figurent dans la table 11. A noter que pour la méthode kCFC, seuls les résultats sur le jeu de données Growth issus de l'article [29] sont disponibles.

Ces expériences mettent en avant plusieurs remarques importantes :

[1] *http ://www.cs.ucr.edu/∼eamonn/time_series_data/*

	Kneading functional	2-steps methods	Kneading discretized (241 instants)	Kneading spline coeff. (20 splines)	Kneading FPCA scores (4 components)
Funclust	**66.96**	HDDC	66.09	53.91	44.35
FunHDDC	62.61	MixtPPCA	65.22	64.35	62.61
fclust	64	mclust	63.48	50.43	60
kCFC	-	k-means	62.61	62.61	62.61
		hclust	63.48	63.48	63.48

	Growth functional	2-steps methods	Growth discretized (350 instants)	Growth spline coeff. (20 splines)	Growth FPCA scores (2 components)
Funclust	69.89	HDDC	56.99	50.51	**97.85**
FunHDDC	96.77	MixtPPCA	62.36	50.53	**97.85**
fclust	69.89	mclust	65.59	63.44	95.70
kCFC	93.55	k-means	65.59	66.67	64.52
		hclust	51.61	75.27	68.81

	ECG functional	2-steps methods	ECG discretized (96 instants)	ECG spline coeff. (20 splines)	ECG FPCA scores (19 components)
Funclust	**84**	HDDC	74.5	73.5	74.5
FunHDDC	75	MixtPPCA	74.5	73.5	74.5
fclust	74.5	mclust	81	80.5	81.5
kCFC	-	k-means	74.5	72.5	74.5
		hclust	73	76.5	64

	CBF functional	2-steps methods	CBF discretized (128 instants)	CBF spline coeff. (20 splines)	CBF FPCA scores (17 components)
Funclust	57.96	HDDC	68.60	51.18	68.17
FunHDDC	**70.65**	MixtPPCA	65.59	51.29	68.27
fclust	(†)	mclust	61.18	62.79	68.06
kCFC	-	k-means	64.95	54.09	64.84
		hclust	60.86	57.96	66.13

TAB. 11 – Taux de classification correcte (CCR) en pourcentage pour Funclust, FunHDDC (meilleur modèle selon BIC), fclust, kCFC et plusieurs méthodes non fonctionnelles classiques sur les jeux de données Kneading, Growth, ECG et CBF. (†)Le package 'fclust' ne supporte pas les jeux de données de cette taille.

- Funclust et FunHDDC sont toutes deux très compétitives. En effet Funclust s'avère donner la meilleure classification sur les données Kneading et ECG et FunHDDC sur les données CBF. Sur le jeu de données Growth, FunHDDC est tout près des meilleurs méthodes, qui sont HDDC et MixtPPCA sur les 2 premières composantes principales, tandis que Funclust est moins performante. Comme discuté dans [R10], cette contre performance de Funclust est due à la difficulté de régler les ordres d'approximation, car lorsqu'ils sont fixés à 2 pour toutes les classes (ce que fait HDDC et MixtPPCA), le taux de classification est alors le même que HDDC et MixtPPCA.
- La méthode fclust semble peu performante, mais il faut rappeler qu'elle a été conçue pour être performante sur des courbes observées à des instants différents avec potentiellement relativement peu d'observations. Quant à la méthode kCFC, elle donne des résultats intéressants sur le jeu de données Growth, qui est celui traité dans l'article proposant la méthode, mais malheureusement l'indisponibilité de code informatique pour cette méthode nous empêche de la tester sur les autres jeux de données.
- Enfin, les techniques de classification automatique pour données fini-dimensionnelles peuvent être performantes, mais souffrent du fait que l'on ne dispose d'aucun outil, dans le cas de la classification non supervisée, pour décider s'il faut appliquer ces méthodes sur les discrétisations des courbes, sur les composantes d'une ACPF ou sur les coefficients dans une base de fonctions. Or, la performance de ces méthodes est très sensible à ce choix, et il s'avère sur les jeux de données analysés ici qu'aucune technique pour transformer le problème de dimension infinie en un problème de dimension finie ne puisse être préférer à une autre.

Données CanadianWeather

Ce jeu de données, disponible dans le package *fda* de **R** et présenté en détail dans [88], consiste en les courbes de précipitations et température journalière de 35 villes du Canada, moyennée entre 1960 et 1994. L'objectif est de chercher à retrouver une classification en 4 groupes en lien avec la géographie du Canada : régions Arctique, Atlantique, Pacifique et Continental. Etant donné que les unités des courbes sont différentes (degrés Celsius et millimètres), les données sont normalisées (cf. section 4.3.1). La figure 11 représente les courbes initiales ainsi que celles normalisées.

Nous illustrons ici le comportement des méthodes Funclust et FunHDDC sur les courbes de température, ainsi que l'extension de Funclust au cas multivarié (section 4.3.2) sur à la fois les courbes de température et de précipitations (normalisées). Nous avons choisi de travailler avec les températures pour les méthodes univariées car c'est la caractéristique qui permet de retrouver le plus facilement la répartition géographique. La figure 12 illustre les classifications obtenues. Pour chaque méthode, une base de Fourier (65 noeuds) a été utilisée, et le seuil du test du coude de Cattell a été fixé à 0.2.

Les trois classifications obtenues par FunHDDC, Funclust et MFunclust sont relativement en concordance avec la répartition géographique des villes (figure 12). FunHDDC définit en effet quatre groupes : continental (bleu), pacifique (rouge), atlantique (noir) et arctique (vert). Funclust et MFunclust classent la ville de Resolute seule dans une classe, à cause de la situation particulière de cette ville : elle est de loin la ville la plus froide et

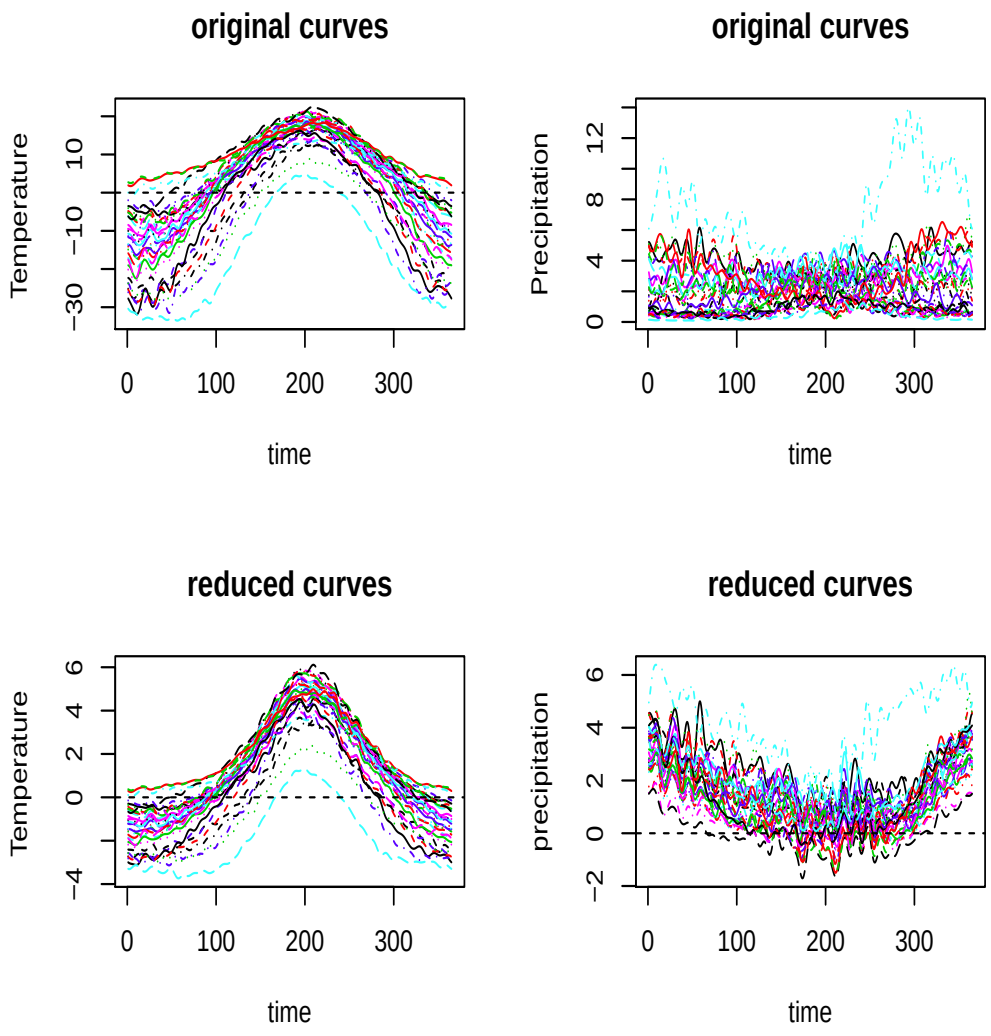


FIG. 11 – Courbes de température et de précipitations pour 35 villes du Canada, moyennée entre 1960 et 1994. Les figures du haut représentent les courbes originales tandis que les courbes réduites sont sur celles du bas.

où il pleut le moins. Cette ville occupant une classe, les villes continentales et Atlantiques sont alors mises dans une même classe (noire).
Ces deux expériences montrent clairement l'intérêt des modèles de classification automatique de courbes proposés.

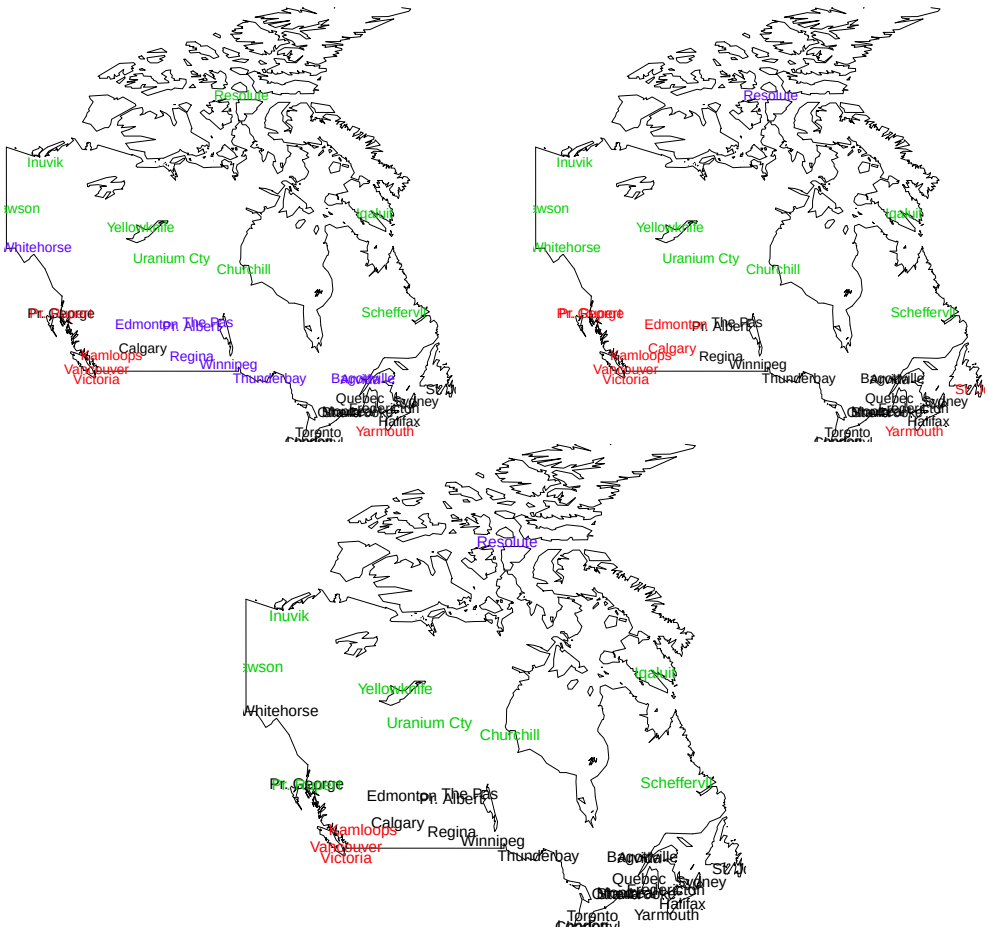


FIG. 12 – Classification des villes du Canada en 4 groupes, en utilisant uniquement les courbes de température (haut), *via* les méthodes FunHDDC (gauche) et Funclust (droite), et en utilisant à la fois les courbes de température et de précipitations (normalisées) *via* MFunclust.

5

RÉGRESSION EN GRANDE DIMENSION

Sommaire

Revues avec comité de lecture

[R13] Yengo L., Biernacki, C. and Jacques, J. *A block regression approach for simultaneous clustering and variables selection*, en préparation.

Conférences internationales avec comité de lecture

[CI6] Hamon, J., Dhaenens C., Even, G. and Jacques, J. *Feature selection combining combinatorial optimization and regression in high dimensional problems*, IEEE International Conference on Data Mining, Bruxelles, Belgique, décembre 2012.

[CI7] Hamon, J., Dhaenens C., Even, G. and Jacques, J. *Feature selection for high dimensional regression using local search and statistical criteria*, META'12, 4th International Conference on Metaheuristics and Nature Inspired Computing, Port El-Kantaoui, Tunisie, octobre 2012.

Conférences nationales

[CN13] Yengo L., Jacques, J. and Biernacki, C. *Classification et sélection de variables en régression*. 44èmes Journées de Statistique organisée par la Société Française de Statistique, Bruxelles, mai 2012.

[CN14] Hamon, J., Even, G., Jacques, J. and Dhaenens C. *Coopération entre optimisation combinatoire et statistique pour la sélection animale*, ROADEF'12, Angers, France, avril 2012.

[CN15] Hamon, J., Even, G., Jacques, J. and Dhaenens C. *Combining combinatorial optimization and statistic to mine high-throughput genotyping data*, JOBIM'11, Paris, France, juin 2011.

[CN16] Yengo L., Jacques, J. and Biernacki, C. *A block regression approach for simultaneous variables clustering and selection : Application to genetic data*, JOBIM'11, Paris, France, juin 2011.

La régression linéaire, développée bien avant l'émergence des ordinateurs, a été largement étudiée depuis de nombreuses années. Récemment, avec l'arrivée de grands ensembles de données comme on en rencontre en génétique par exemple, dans lesquels le nombre de variables explicatives dépasse largement le nombre d'observations, les statisticiens se trouvent confrontés à de nouveaux problèmes de *grandes dimensions*. L'estimation d'un modèle de régression n'est alors plus identifiable, et plusieurs solutions ont été développées (cf. Hastie *et al.*[56] pour une synthèse complète) :

- les techniques de recherche séquentielle du meilleur sous ensemble de variables explicatives (*forward*, *stepwise*...), qui consistent à intégrer une à une les variables explicatives dans le modèle de régression,
- les techniques de régression pénalisée (*ridge*, *lasso*, *elastic-net*...), qui pénalisent l'estimation des moindres carrés par diverses fonctions du vecteur des paramètres de régression : la somme des valeurs absolues des coefficients pour lasso (pénalité ℓ_1), qui tend à annuler un certain nombre de coefficients, la somme des carrés des coefficients pour ridge (pénalité ℓ_2) ou encore un mélange des deux pour elastic-net,
- les techniques de création de méta-variables, comme la régression sur composantes principales issues de l'analyse en composantes principales (ACP) ou la régression PLS (*Partial Least Square*).

Les travaux que nous présentons ici sont le fruit de deux thèses, débutées en 2010 et dont les soutenances sont prévues pour fin 2013 :

- la thèse CIFRE de Julie Hamon, financée par l'entreprise Gènes Diffusion spécialisée en sélection génétique animale, dont l'objectif est de définir un modèle de prédiction d'un trait quantitatif, comme la production de lait ou la qualité de la viande, à partir de marqueurs génétiques. L'idée développée dans cette thèse, que je co-encadre avec Clarisse Dhaenens (Univ. Lille 1, Professeure en Optimisation Combinatoire), est de combiner optimisation et statistique en utilisant les algorithmes d'optimisation combinatoire pour sélectionner les variables et les modèles statistiques pour évaluer la qualité de la sélection proposée.
- la thèse de Loïc Yengo, ingénieur de recherche CNRS à l'Institut de Biologie de Lille, a également pour objectif de définir un modèle de régression sur un ensemble important de marqueurs génétiques. L'approche développée dans cette thèse, que je co-encadre avec Christophe Biernacki (Univ. Lille 1, Professeur en Statistique), est la suivante : plutôt que de supprimer des variables explicatives, nous cherchons à regrouper ensemble les variables ayant un effet similaire. Ainsi, la dimension du problème sera réduite, sans pour autant supprimer des variables qui pourraient avoir un intérêt. Comme nous le verrons, l'ajout d'une hypothèse supplémentaire pourra néanmoins permettre de supprimer des variables qui n'auraient qu'un effet marginal.

Dans ces deux travaux, les marqueurs génétiques utilisés en tant que variables explicatives sont les polymorphismes de séquence (*Single Nucleotide Polymorphisme*, SNP). Un SNP est une variation (polymorphisme) d'une seule paire de bases du génome, entre individus d'une même espèce. Ces variations sont relativement fréquentes (environ une paire de

bases sur mille dans le génome humain), et responsables d'environ 90% de l'ensemble des variations génétiques humaines. La figure 13[1] illustre la notion de SNP, décrite plus en détail dans Vignal *et al.* [105] notamment. Un SNP peut prendre trois valeurs possibles, que l'on notera aa, Aa et AA, et qui sont traditionnellement codées $0, 1, 2$ ou $-1, 0, 1$.

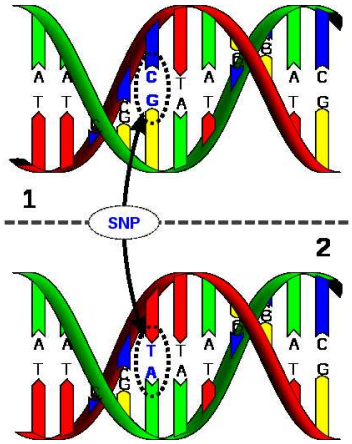


FIG. 13 – Polymorphismes de séquence ou *Single Nucleotide Polymorphisme*.

Notons $y \in \mathcal{Y} = \mathbb{R}$ le trait quantitatif que l'on cherche à prédire et $\boldsymbol{x} = (x_1, \ldots, x_p)' \in \mathcal{X}$ les SNPs prédicteurs. On dispose d'un échantillon $(\boldsymbol{x}_i, y_i)_{1 \leq i \leq n}$ d'observations. Les dimensions typiques sont de l'ordre de quelques centaines de milliers à quelques millions de SNP pour quelques milliers d'observations.

5.2 SÉLECTION DE VARIABLES PAR OPTIMISATION COMBINATOIRE

5.2.1 *Le modèle*

La thèse de Julie Hamon considère le modèle de régression suivant

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j \zeta_j x_{ij} + \epsilon_i, \tag{22}$$

où $\zeta_j \in \{0, 1\}$ est un paramètre binaire qui indique si la variable x_j est conservée dans le modèle de régression ou non, et avec $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ des résidus indépendants. Les paramètres de ce modèle sont $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)' \in \mathbb{R}^{p+1}$, $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_p)' \in \{0, 1\}^p$ et $\sigma^2 \in \mathbb{R}$.

[1] source *wikipedia*

5.2.2 *Estimation*

La maximisation de la vraisemblance du modèle (22) en fonction du paramètre discret ζ ne peut être réalisée en calculant la vraisemblance pour chaque valeur possible de $\boldsymbol{\zeta}$. Nous proposons alors dans [CI6,CI7] l'optimisation alternée suivante :

(i) estimation de $\boldsymbol{\zeta}$ conditionnellement à $(\boldsymbol{\beta}, \sigma)$,

(ii) estimation des $(\boldsymbol{\beta}, \sigma)$ conditionnellement à $\boldsymbol{\zeta}$.

Si l'étape (ii) peut être réalisée par maximum de vraisemblance classique, dès lors que $\sum_{j=1}^{p} \zeta_j \leq p$, l'étape (i) fait appel à un algorithme d'optimisation combinatoire de recherche locale itérée (*Iterated Local Search [99]*, ILS, illustré par la figure 14), qui consiste en une succession de recherches locales et de perturbations. La recherche locale est effectuée de la façon suivante : partant d'une sélection de variables donnée, on choisit aléatoirement une des p variables ; si elle est dans la sélection courante on teste la régression sans cette variable, et on conserve cette solution si la régression est de meilleure qualité (selon un critère qui sera détaillé plus loin) ; si la variable n'est pas dans la sélection courante, on l'ajoute et on teste également la qualité de la régression. Si on ne trouve pas de solution meilleure que la solution courante, on considère que l'on est sur un optimum local, et on réalise alors une perturbation en sélectionnant aléatoirement un certain nombre de variables (quelques unités), qu'on ajoute dans la sélection si elles n'y sont pas ou qu'on enlève de la sélection si elles y sont déjà. Cet algorithme est itéré jusqu'à ce que plusieurs répétitions successives (une centaine) des cycles perturbation / recherche locale soient réalisées sans améliorer l'optimum local.

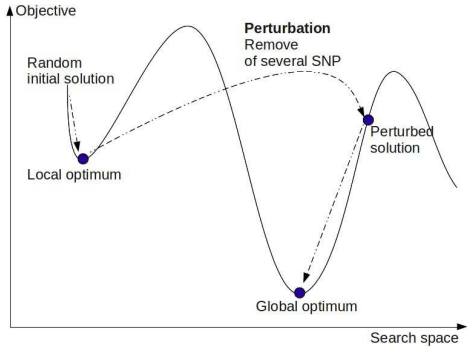


FIG. 14 – Illustration de l'algorithme ILS.

Différentes évaluations de la qualité de la régression peuvent être utilisées. Dans [CI6,CI7] les critères BIC, validation croisée *leave-one-out* et *3-fold* ont été testés, avec un avantage pour la validation croisée 3-fold comme nous le verrons dans les expériences suivantes.

5.2.3 *Expérimentations numériques*

Dans [CI6], nous présentons deux simulations permettant d'illustrer le comportement de la technique de sélection de variables proposée. Dans une première simulation, on

considère le modèle à $p = 1000$ variables dont seulement les dix premières ont une influence sur la réponse Y :

$$Y = x_1 + 2 * x_2 + \ldots + 10 * x_{10} + \epsilon$$

avec $\epsilon \sim \mathcal{N}(0, 0.25)$, $x_1, \ldots, x_{10}$ des réalisations indépendantes de loi uniforme sur $\{-1, 0, 1\}$, et une taille d'échantillon $n = 100$ dix fois plus petite que le nombre de variables. Le graphique de gauche de la figure 15 représente sous la forme de boxplot les erreurs quadratiques de prédiction obtenues, évaluées sur un échantillon test, pour 20 simulations différentes. La méthode proposée semble être une alternative très intéressante aux techniques classiques, bien que pour cette configuration de simulation la meilleure des méthodes soit la régression stepwise.

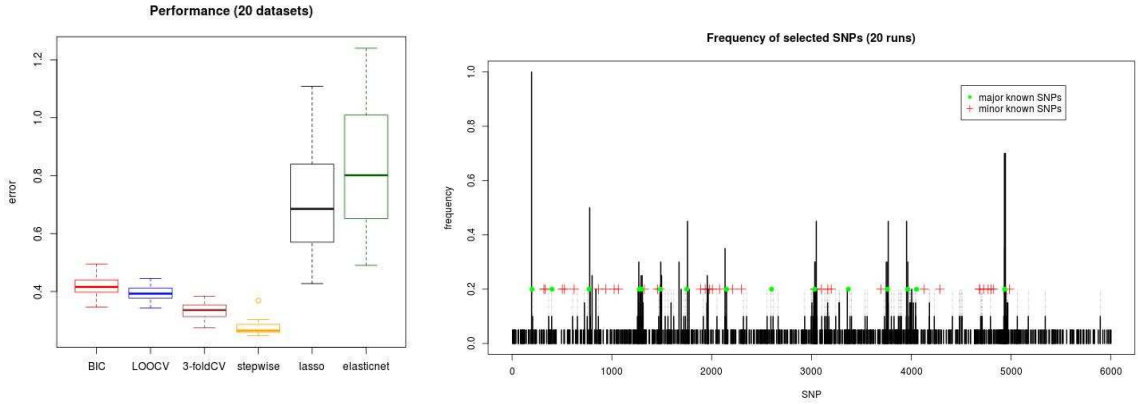


FIG. 15 – Comparaison avec les méthodes classiques de régression en grande dimension (gauche) et fréquences des SNPs sélectionnés sur le jeu de données QTLMAS (droite).

Le second jeu de données, QTLMAS, est une simulation provenant du *XII QTLMAS Workshop* censée être proche des problématiques réelles en génétique animale. Il comporte 6000 SNPs, dont 15 ont un effet majeur et 35 un effet mineur, pour 4665 individus. Pour ce jeu de données, nous avons exécuté 20 fois notre algorithme (avec validation croisée 3-fold). Le graphique de droite de la figure 15 présente les fréquences de sélection de chaque SNP (les points vers indiquent les emplacements des SNPs majeurs et les croix rouges ceux des SNPs mineurs). On constate avec satisfaction que les SNPs majeurs sont plutôt bien retrouvés. En revanche, les SNPs mineurs sont plus difficilement détectés.

5.2.4 *Spécificité de la génétique animale et suite de la thèse*

La principale spécificité de la génétique animale, qui la différencie notamment des études génétiques humaines, et qui n'a pas été prise en compte dans le modèle précédent, est la dépendance entre les observations. En effet, pour une sélection de SNPs donnée, les paramètres (β, σ) ont été estimés par maximum de vraisemblance classique, en supposant que les observations étaient i.i.d.. Pour cette raison, dans les études génétiques humaines,

on évite généralement de choisir des individus d'une même famille. En génétique animale, étant donné le passif de sélection animale (avant même l'avènement de la génétique), les animaux d'élevage ont pour beaucoup des liens de parenté. En effet, il n'est pas rare que pour une race donnée, seulement quelques reproducteurs mâles soient utilisés.
De ce fait, le modèle (22) n'est pas le plus adapté à ce type de données, et l'introduction de modèles mixtes permettant de prendre en compte les relations familiales est nécessaire [109]. L'avantage de la méthodologie développée précédemment, est qu'elle est indépendante du modèle statistique utilisé, et peut donc être adaptée pour la sélection de variables d'un modèle mixte.

5.3 CLASSIFICATION ET SÉLECTION DE VARIABLES EN RÉGRESSION

Dans la thèse de Loïc Yengo, nous supposons que les coefficients de régression β_j ne sont plus des paramètres mais des variables aléatoires, et qu'il existe une partition de ces coefficients de régression en K groupes latents. Ainsi, associant les coefficients de régression aux variables correspondantes, il existe une partition des variables explicatives, telle que les variables d'un même groupe aient des coefficients de régression provenant d'une même distribution, que nous précisons ci-après.
Récemment, cette idée de coupler classification et sélection de variables dans les modèles de régression a donné naissance à une riche littérature comprenant des approches de régression pénalisée comme OSCAR [17], qui mélange une pénalité ℓ_1 de type lasso avec une pénalité ℓ_∞, ou encore le *Clustered Lasso* [92] qui pénalise des contrastes sur l'espace des paramètres.

5.3.1 *Le modèle*

Le modèle que nous proposons est le suivant :

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i \tag{23}$$

avec $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ et où, conditionnellement à l'appartenance au $k$ème groupe,

$$\beta_j | Z_{jk} = 1 \sim \mathcal{N}(b_k, \gamma_k^2)$$

avec Z_{jk} la variable aléatoire égale à 1 si la $j$ème variable appartient au $k$ème groupe et 0 sinon, distribuée suivant une loi multinomiale :

$$Z_{jk} \sim \mathcal{M}(1, \pi_1, \ldots, \pi_K).$$

Les paramètres du modèle ainsi défini $(\pi_k, b_k, \gamma_k)_{k=1,K}$ sont en nombre réduit $(3K - 1)$ en regard de la complexité initiale du problème. En outre, il est possible avec ce modèle d'imposer un coefficient b_k nul, conduisant ainsi à réaliser une sélection de variables.

L'estimation des paramètres du modèle peut être réalisée à l'aide d'un algorithme EM-Gibbs dans lequel l'étape E, non calculable explicitement, est approchée à l'aide d'un algorithme de Gibbs. Le nombre K de groupes peut quant à lui être sélectionner classiquement en utilisant soit la validation croisée, soit le critère BIC.

5.3.2 *Expérimentations numériques*

Afin d'illustrer les performances de la méthode, nommée CLERE dans la suite pour CL*usterwise* E*ffect* RE*gression*, nous avons réalisé une étude de simulation basée sur le modèle de régression suivant

$$Y = \sum_{j=1}^{p} \beta_j x_j + \epsilon$$

avec $\epsilon \sim \mathcal{N}(0,1)$, les variables explicatives étant des réalisations d'une loi normale multivariée centrée telle que $Cov(X_j, X_{j'}) = \rho^{|j-j'|}$ (deux valeurs $\rho = 0$ et $\rho = 0.5$ seront considérées), et où les β_j sont simulés suivant un mélange de 5 lois normales, centrées respectivement en -10, -5, 0, 5 et 10 et de variance égale à 1. Les proportions du mélange sont telles que 20% puis 50% des β_j suivent une loi normale centrée, les 4 autres classes étant équiprobables. La taille d'échantillon est fixée à $n = 50$, tandis que le nombre de variables est fixé à $p = 100$. Nous comparons notre méthode CLERE, en fixant le nombre de classes à $K = 5$, aux méthodes classiques lasso, ridge, elastic-net et régression stepwise. Le critère de comparaison est l'erreur de prédiction, $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, évaluée sur un échantillon test.
La figure 16 illustre les résultats obtenus, qui s'avèrent très favorables à la méthode CLERE. En effet, les erreurs de prédiction obtenues sont bien meilleures que celles de toutes les autres méthodes, même dans le cas considérant 50% de β_j nuls pouvant être attendu comme favorable aux approches comme lasso. Des applications sur données réelles sont actuellement en cours, et devraient permettre de confirmer les bonnes dispositions de CLERE observées sur simulation.

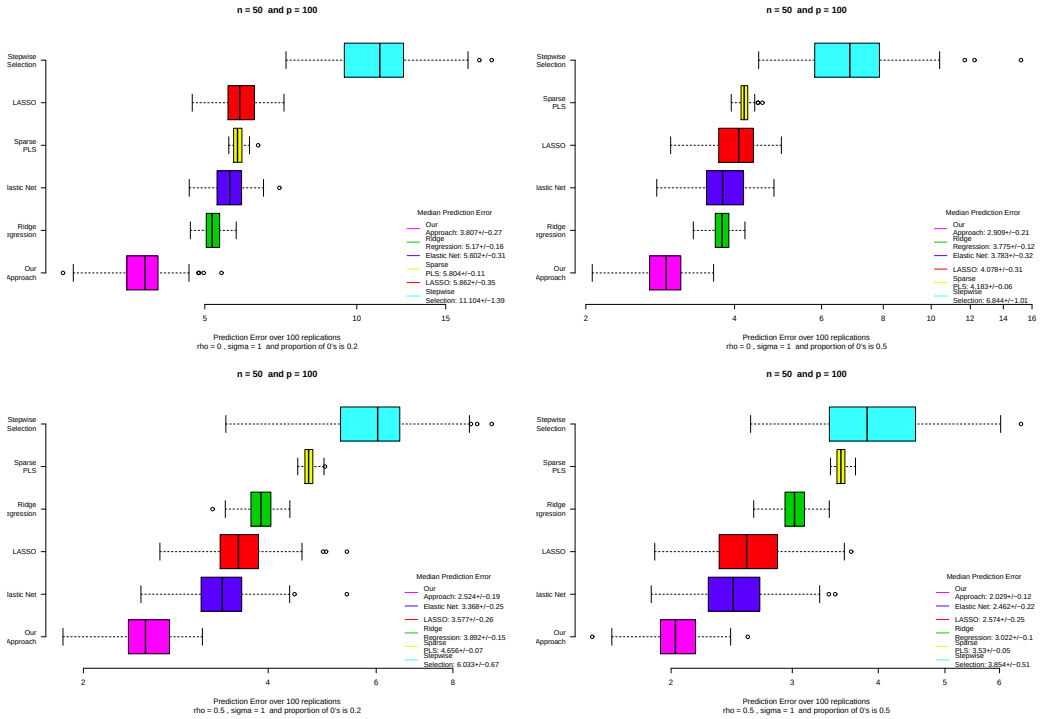


FIG. 16 – Performance de CLERE avec nombre de groupes fixé (*our approach*, boxplots du bas) vis-à-vis des méthodes classiques d'un point de vue erreur de prédiction, dans le cas $n = 50$ et $p = 100$ avec variables explicatives indépendantes ($\rho = 0$, haut) et corrélées ($\rho = 0.5$, bas), et avec 20% (gauche) et 50% (droite) de β_j nuls.

6

APPLICATIONS

Sommaire

Revues avec comité de lecture

[R14] Jacques, J., Bouveyron, C., Girard, S., Devos, O., Duponchel, L. and Ruckebusch, C. *Gaussian mixture models for the classification of high-dimensional vibrational spectroscopy data*, Journal of Chemometrics, 24, 719-727, 2010.

[R15] Langlois-Jacques, C. and Jacques, J. *Détection d'hétérogénéité au sein de mesures de qualité de l'environement*, La Revue Modulad, 40, 41–52, 2009.

[R16] Jacques, J. and Devictor, N. and Lavergne, C. *Sensitivity Analysis in presence of Model Uncertainty and Correlated Inputs*, Reliability Engineering and System Safety, 91, 1126–1134, 2006.

Conférences internationales avec comité de lecture

[CI8] Lebret R., Biernacki, C., Iovleff S., Jacques, J., Preda, C., McCarthy A. and Delrieu O. *Genetic apistasis analysis using 'taxonomy3'*. 2nd International BIO-SI Workshop, Rennes, France, octobre 2011.

[CI9] Ruckebusch C., Bouveyron C. and Jacques, J. *Classification of High-Dimensional NIR Spectroscopic Data*. RENACQ-4, Beni Mellal, mars 2010.

[CI10] Lebret R., Iovleff S., Biernacki, C., Jacques, J., Preda, C., McCarthy A. and Delrieu O. *Rapid multivariate analysis of 269 Hapmap subjects and 1 million SNPs using 'taxonomy3'*. Cold Spring Harbor/Wellcome Trust meeting on Pharmacogenomics, Hinxton, UK, septembre 2009.

Conférences nationales

[CN17] Jacques J. *Classification supervisée et non supervisée de données spectroscopiques à base de modèle probabilistes*. Conférence invitée au congrès Chimiométrie 2012, Lille, décembre 2012.

6.1.1 *Identifications de facteurs génétiques responsables d'intolérances médicamenteuses*

Dans le cadre d'un contrat de collaboration sur deux années entre l'entreprise anglo-japonaise PGxIS et le laboratoire Painlevé (représenté par Christophe Biernacki, Serge Iovleff, Cristian Preda et moi-même), nous avons contribué au développement d'une méthodologie permettant d'identifier les facteurs génétiques responsables d'intolérances médicamenteuses. En effet, la mise en place d'un nouveau médicament passe par de multiples phases, dont l'une consiste à tester le médicament sur un échantillon d'individus pour constater de son efficacité. Malheureusement, il arrive parfois qu'un médicament entraîne des effets indésirables chez certains cobayes qui, bien souvent, résultent du patrimoine génétique des individus. Si l'on pouvait détecter les facteurs génétiques responsables d'une intolérance à un médicament, il serait alors possible de proposer des traitements individualisés aux patients. C'est ce à quoi s'attache l'entreprise de pharmacogénétique PGxIS.

Pour effectuer ce type d'étude, nous disposons d'un ensemble d'invidus cas-contrôle pour lesquels sont mesurés des marqueurs génétiques (SNP, déjà définis dans le chapitre 5). Ces SNP peuvent prendre trois valeurs différentes, que l'on notera ici AA, Aa et aa. Les bases de données sur lesquelles travaille PGxIS sont généralement de l'ordre de 1000 cas et 50 témoins, pour un million de SNPs.

L'objectif du contrat de collaboration était d'améliorer, d'apporter des justifications théoriques et d'implémenter en C++ la méthodologie initiée par Delrieu et Bowman [36], appelée Taxonomy3. Cette méthode consiste en les principaux points suivants :

- plutôt que de travailler avec les variables qualitatives que sont les SNPs, chaque modalité est remplacée par le logarithme du rapport des fréquences de la modalité chez les cas et les contrôles : par exemple, pour le $j$ème SNP

$$AA \to lbf_j(AA) = \log \frac{f_{j,\text{cas}}(AA)}{f_{j,\text{contrôle}}(AA)}$$

  où $f_{j,\text{cas}}(AA)$ est la fréquence empirique de la modalité AA pour le SNP j chez les cas et $f_{j,\text{contrôle}}(AA)$ celle pour les contrôles.
- une ACP est réalisée sur la matrice des lbf. La figure 17 présente un exemple de résultat d'une telle ACP. On y voit que la première composante principale distingue très clairement les cas des témoins. L'idée est alors de détecter les SNPs significativement corrélées avec cette première composante principale.
- la significativité de la corrélation des SNPs avec la première composante principale est testée en réalisant des permutations aléatoires des statuts cas-contrôle. Ainsi, il est possible d'obtenir une p-value correspondant à la probabilité que la corrélation observée entre un SNP et la première composante, responsable de la distinction cas-contrôle, soit non significative. Pour estimer des p-values de très faibles valeurs, le nombre de permutations à réaliser doit être très important. Afin de réduire ce nombre, une modélisation de la distribution des p-values par mélange gaussien à l'aide du logiciel MIXMOD [14] est réalisée, ce qui permet d'estimer précisément

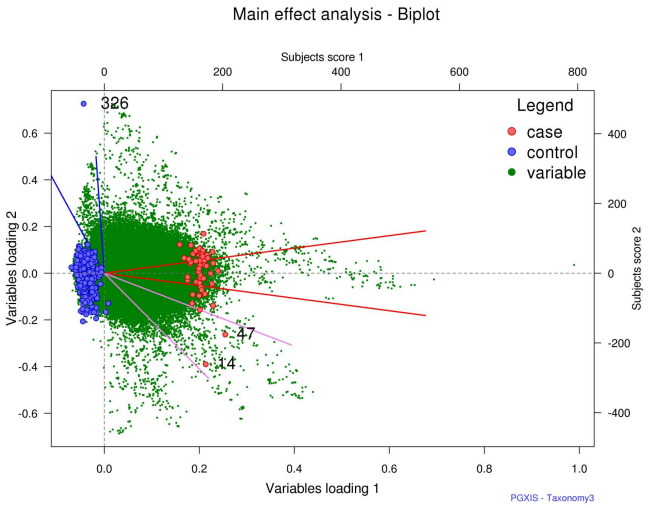


FIG. 17 – Représentation des SNPs (en vert) et des individus (cas en rouge et témoins en bleu) sur le premier plan factoriel.

les p-values les plus faibles. La figure 18 représente un exemple de ces p-values en fonction de l'emplacement du SNP sur les différents chromosomes. Un trait rouge indique le seuil au dessous duquel (l'échelle des ordonnées est inversée sur la figure 18) les p-values correspondent à des SNPs significatifs de la distinction cas-contrôle, seuil estimé à l'aide de la procédure pour tests multiples de Šidák [94].

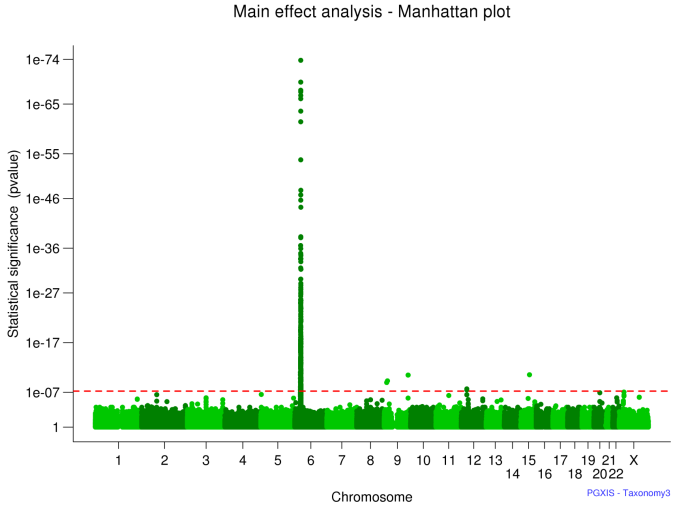


FIG. 18 – p-value des tests de significativité des corrélations des SNPs avec la première composante principale.

La méthodologie Taxonomy3 ainsi développée a été validée sur différents jeux de données biologiques [CI8,CI10] en retrouvant des gènes dont l'effet sur l'intolérance médi-

camenteuse étudiée était connu, et a également permis de détecter de nouveaux gènes potentiellement responsables de cette intolérance.

6.1.2 *Applications en génétique animale et médicale*

Les deux thèses auxquelles le chapitre 5 est consacré sont destinées à développer des modèles de prédiction d'une variable quantitative sur la base de marqueurs génétiques (SNP).
La thèse de Julie Hamon, financée par la société Gènes Diffusion, spécialisée dans la sélection génétique animale, a pour objectif d'expliquer des traits quantitatifs comme la production de lait ou la qualité de la viande à partir des SNPs. Plus que la réalisation d'un modèle prédictif, l'objectif est de déterminer un sous-ensemble de SNPs responsable des variations du trait considéré. Ainsi, à partir de l'analyse d'un nombre réduit de SNPs, et donc pour un coût réduit, il sera possible de déterminer les qualités de l'animal vis-à-vis du trait recherché.
La thèse de Loïc Yengo, ingénieur de recherche au sein de l'équipe de *génomique et physiologie moléculaire des maladies métaboliques* de l'Institut de Biologie de Lille, s'intéresse quant à elle principalement aux maladies de l'obésité et du diabète. L'objectif est néanmoins le même que dans le cas de la sélection animale, c'est-à-dire d'identifier un groupe de SNPs permettant d'expliquer un trait comme le taux d'insuline par exemple. Étant donné le modèle proposé dans cette thèse, l'idée pourra être d'identifier les variables du groupe ayant le coefficient de régression le plus fort, et si besoin de rechercher au sein de ces variables celles dont les appartenances à ce groupe sont les plus fortes.

6.2 APPLICATIONS EN CHIMIOMÉTRIE

Je présente dans cette section plusieurs travaux réalisés dans le cadre d'une collaboration avec le laboratoire Laboratoire de Spectrochimie Infrarouge et Raman (LASIR) de l'Université Lille 1 sous la forme d'un groupe de travail intitulé MADD (Mesure et Analyse des Données de grande Dimension) que Cyril Rückebusch (LASIR) et moi -même avons initié. Ces travaux ont fait l'objet d'un article [R13] et d'une conférence internationale [CI9], et ont également conduit à l'organisation de deux manifestations pluridisciplinaires, rassemblant une centaine d'élèves-ingénieurs, partenaires industriels et enseignants-chercheurs autour de la thématique de l'analyse des données de grande dimension (EPIC09) et de l'analyse des données biologiques (EPIC10). Ces deux thématiques sont en effet d'un intérêt tout particulier tant pour les statisticiens que pour les chimiométriciens.

6.2.1 *Discrimination de tissus*

Une des problèmes auxquels sont souvent confrontés les chimiométriciens est la classification (supervisée) des données issues de spectroscopie (infra-rouge, proche infra-rouge ou Raman). Ces données sont généralement des spectres échantillonnés à plusieurs milliers de longueurs d'ondes (2800 pour le jeu de données présenté ci-après) pour seulement

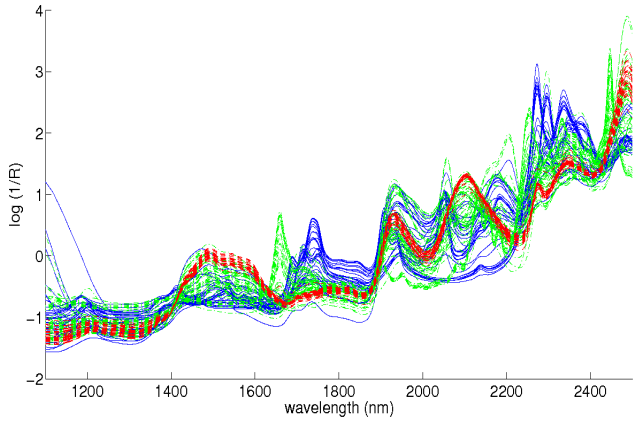


FIG. 19 – Spectres proche infra-rouge (3 classes).

quelques centaines de spectres mesurés. Du fait du grand nombre de variables en regard du nombre d'observations, les techniques habituellement utilisées pour classer ce type de données sont généralement les méthodes SVM [104], SIMCA [107] ou PLS-DA [8].

Les données considérées ici (figure 19) sont des spectres proche infra-rouge (NIR) de tissus, répartis en trois classes en fonction de leur composition. L'idée est de chercher à déterminer une règle de classification qui permettrait de retrouver automatiquement la composition du tissu en fonction de son spectre. Le jeu de données contenait 202 spectres mesurés en proche infra-rouge sur 2800 longueurs d'ondes.

Nous avons utilisé dans [R13] une méthode probabiliste de classification de données de grande dimension (HDDA [20]), et nous l'avons comparé aux méthodes classiques (SVM, SIMCA et PLS-DA) sur la base du taux de bonnes classifications (CCR) estimé par validation croisée ou à l'aide d'un échantillon test (table 12). Pour cette expérience, les paramètres de chaque méthode ont été réglés par validation croisée (5-fold). Il apparaît que la méthode probabiliste HDDA surpasse les autres méthodes à la fois en performance de classification mais également en temps de calcul (table 12). En outre, l'étude a posteriori des paramètres des modèles HDDA a permis aux collègues chimiométriciens de réaliser des interprétations fines des résultats.

Méthodes	CCR en CV	CCR en test	Temps (sec.)
HDDA	**92.3**	**96.7**	**3**
SVM	88.5	91.2	182
PLS-DA	87.7	84.7	59
SIMCA	-	82.4	-

TAB. 12 – Taux de classifications correctes en 3 classes des spectres NIR.

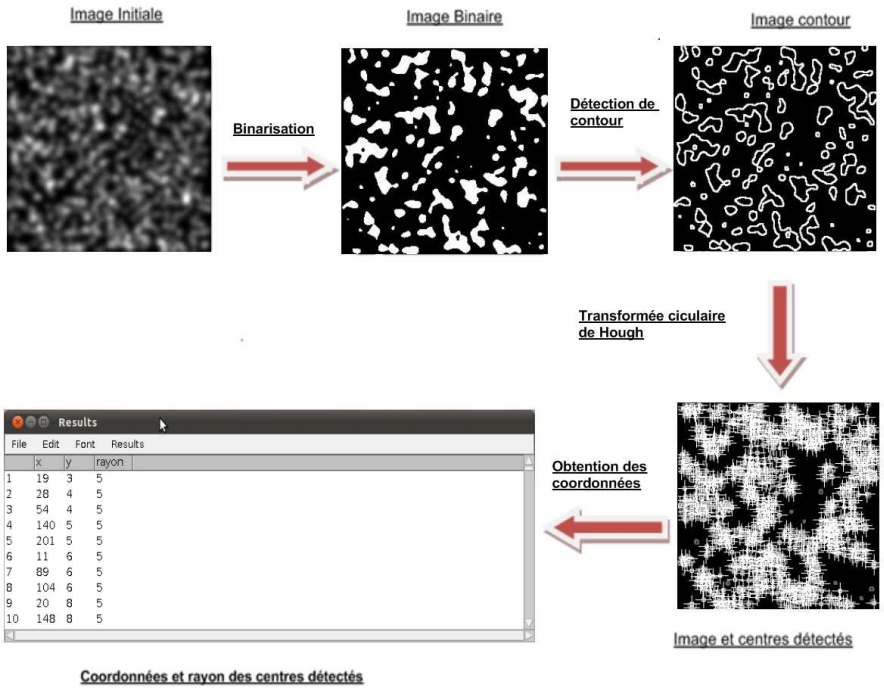


FIG. 20 – Première méthodologie de détection des Quantum Dots.

6.2.2 *Détection de nanocristaux fluorescents*

La microscopie à fluorescence est une technique couramment utilisée en imagerie moléculaire en biologie. Les limites de cette technologie (limite de résolution de diffraction) rend impossible la distinction de fluorophores éloignés de moins de 200 nm. Des travaux récents [57, 60] ont montré qu'il était possible d'exploiter les propriétés de clignotement de certains nanocristaux particuliers, les *Quantum Dots* (QDs), pour localiser leur coordonnées précises dans une image résultant d'une microscopie à fluorescence. Concrètement, on observe une série d'images dans lesquelles sont présentes des tâches lumineuses produites par les QDs. Le clignotement des QDs fait que les tâches lumineuses associées à chaque QDs sont présentes aléatoirement dans chaque image. Ainsi, on peut espérer à partir d'un grand nombre d'images de discerner chaque QDs présents dans un amas de QDs qui ne seraient pas discernables sur une même image.

L'objectif de la collaboration entre le LASIR et le Laboratoire Paul Painlevé (représenté par Radu Stoïca et moi-même) dans ce domaine, débutée en 2011 et toujours en cours, est de développer une technique mathématique permettant de localiser les coordonnées précises de chaque QDs. Un premier travail a été initié, en tentant de détecter les QDs présents dans une unique image. Pour ce faire, des techniques classiques d'analyse d'image ont été utilisées par un étudiant stagiaire que nous avons dirigé : binarisation de l'image, utilisation d'opérateurs morphologiques, détection de contours, transformée de Hough [49]. Ces travaux, résumés par la figure 20, ont donné des résultats encourageants.

6.2.3 *Segmentation de la surface de Mars*

Dans cette dernière application en chimiométrie, les données, fournies par le Laboratoire de Planétologie de Grenoble [10, 12] ont été obtenues à l'aide du satellite imageur OMEGA. Le sol de la planète Mars a été observé avec une résolution comprise entre 300 et 3000 mètres en fonction de l'altitude du satellite. Une photo de la zone étudiée est présentée par la figure 21 (partie gauche). En chaque pixel de l'image de taille 300×128, un spectre (visible et infra-rouge) a été mesuré (figure 21, partie droite) pour une plage de longueur d'onde de 0.36 à 5.2 microns.

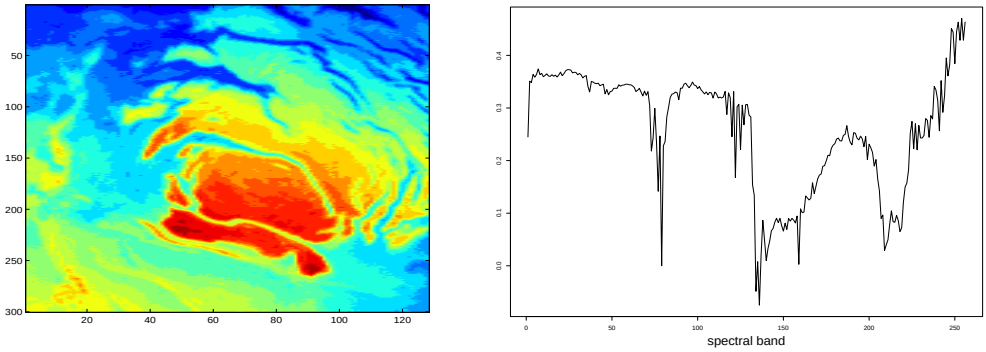


FIG. 21 – Image de la zone étudiée de la planète Mars (gauche) et exemple de spectre (droite).

Dans cette application, l'objectif était de réaliser une classification non supervisée des spectres dans l'objectif de caractériser les matériaux composant le sol de Mars. En considérant les spectres comme des courbes, et en utilisant la méthode de classification de courbe Funclust présentée dans le chapitre 4, nous avons réalisé une classification en 8 classes (nombres de matériaux différents attendus par les experts), présentée par la figure 22.

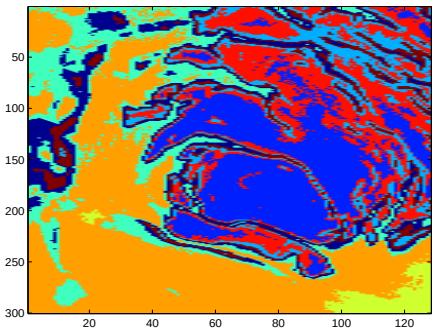


FIG. 22 – Classification en 8 groupes par la méthode Funclust.

La classification obtenue est très en accord avec la partition construite par les experts sur la base de photographies (figure 21) et de modèle physique, bien qu'aucune information spatiale n'ait été utilisée. En particulier, des classes spécifiques sous la forme de fines bandes séparant des classes plus importantes ont été identifiées : par exemple, les classes magenta et cyan séparent la classe bleue et la classe orange. Cela reflète la présence de matériaux particuliers (mélange de carbonate et de glace par exemple) à la frontière des principales matières (glace et poussière).

6.3 APPLICATION AU CONTRÔLE QUALITÉ

Dans le cadre d'une collaboration avec la société AGLAE (Association Générale des Laboratoires d'Analyse de l'Environnement), nous travaillons à la mise en place de procédure de contrôle qualité externe. Dans un but d'accréditation aux normes qualité en vigueur, les laboratoires d'analyses de l'environnement font appels à AGLAE pour organiser des essais interlaboratoires, qui consistent à envoyer des échantillons d'un même matériau, dans lequel une certaine entité doit être mesurée, à un ensemble de laboratoires clients et à analyser les mesures effectuées. Étant donné la nature des matériaux analysés (eaux, boues...), la quantité à mesurer n'est généralement pas connue de façon exacte, et l'appréciation de la qualité des mesures des laboratoires se fait en comparant les résultats des laboratoires les uns aux autres. Dans le cadre de mesures chimiques, les entités mesurées sont quantitatives continues ($\mathbb{R}$), et un ajustement par une loi normale (voir un mélange de loi normale [R14] dans certains cas) permet de définir les bornes au-delà desquelles les mesures des laboratoires sont qualifiées d'atypiques. En prenant en compte la façon dont sont organisés les essais interlaboratoires (deux échantillons sont en réalité envoyés à chaque laboratoire, qui mesurent chaque échantillon deux fois), des techniques d'analyse de variance classiques sont utilisées.
Dans le cas de mesures microbiologiques, les entités ne sont plus continues mais discrètes ($\mathbb{N}$) puisqu'on cherche généralement à dénombrer un type donné de bactéries. Dans le domaine écotoxicologique, les mesures étant cette fois des évolutions de concentrations en fonction de la dilution du milieu, peuvent s'apparenter à des fonctions. Ces deux domaines nécessitent donc le développement d'outils d'analyse de variance spécifiques aux variables aléatoires discrètes et fonctionnelles, ce qui constitue le sujet d'une thèse CIFRE que je co-encadrerai avec Cristian Preda (Univ. Lille 1) à partir de janvier 2013.

6.4 DIFFUSION DE CODES INFORMATIQUES

Un des objectifs de mes travaux de recherche est de rendre disponible les outils que je développe sous la forme de codes informatiques ou *packages*. Ainsi, sont disponibles sur ma page web les packages et codes suivants :

- analyse des données de rang : ce package pour **R** fournit les outils permettant d'estimer les modèles ISR et Φ de Mallows pour des données de rang univariées homogènes (cf. chapitre 3). Une version pour données de rang multivariées hétérogènes,

développée en C++ au cours d'un stage de Master cette année, sera proposée sur le site du CRAN[1] avant la fin de l'année.

- apprentissage adaptatif en régression : ce package pour **R** fournit des outils pour transférer un modèle de régression d'une population de référence à une nouvelle population avec seulement quelques observations (cf. chapitre 2).
- analyse de sensibilité (travaux réalisés au cours de ma thèse, non abordés dans ce document) : une interface graphique fonctionnant sous Matlab permet de réaliser des analyses de sensibilité à l'aide des principales méthodes classiques. Cette interface est à ce jour un des seuls outils libre permettant de réaliser une analyse de sensibilité sous Matlab. Le calcul d'indices multidimensionnels, développés dans [R15], est également disponible dans cette interface. Une version sous la forme d'un code informatique pour le logiciel **R** est également disponible.

Comme cela sera mentionné dans le chapitre suivant, il est prévu dans un futur proche de proposer un certain nombre d'autres packages **R** sur le site du CRAN, afin de favoriser la diffusion de mes travaux de recherche.

[1] http ://cran.r-project.org/

7 CONCLUSIONS ET PERSPECTIVES

Ce mémoire retrace mes activités de recherche dans le domaine de l'apprentissage statistique des données complexes, que sont les données de rang et les données ordinales (chapitre 3), les données fonctionnelles (chapitre 4), les données de grande dimension (chapitre 5) ainsi que les données issues de populations différentes (chapitre 2). Ces travaux ouvrent bon nombre de pistes de recherche, présentées ci-après, qui définissent une partie de mes projets de recherche pour les années à venir.

7.1 ANALYSE DES DONNÉES DE RANG

Le modèle ISR s'avérant être un modèle très pertinent et prometteur, il serait intéressant de poursuivre plus en avant son exploitation. En effet, nous avons vu qu'une des originalités majeures de ce modèle était la prise en compte des ordres de présentation des objets à classer. On convient aisément que ces ordres de présentation ont une importance, notamment lorsqu'on assimile la génération d'une donnée de rang à un algorithme de tri. Or, dans un certain nombre d'applications, comme dans les élections irlandaises analysées dans le chapitre 3, l'ordre de présentation des candidats est connu (ordre alphabétique). Nous pourrions donc intégrer la connaissance de cet ordre dans le modèle, ce qui en outre réduirait sa complexité. Sans aller jusqu'à intégrer la connaissance de cet ordre de présentation, nous pourrions intégrer dans le modèle uniquement l'existence d'un ordre de présentation identique pour toutes les observations. Les propriétés de ces nouveaux modèles devront néanmoins être étudiées à nouveaux de façon théorique, car plusieurs propriétés du modèle ISR reposent sur la non connaissance de ces ordres de présentation. En outre, nous avons considéré une probabilité de bonne comparaison identique tout au long de l'algorithme de classement. Or, il serait possible et assez réaliste de supposer que cette probabilité ne soit pas constante : en effet, nous pourrions supposer que la personne réalisant le classement se fatigue au fur et à mesure qu'elle classe les objets, et ainsi que la probabilité de bon classement diminue pour tendre vers $\frac{1}{2}$ (situations où le classement est fait aléatoirement) à la fin de l'algorithme de tri. Il est également possible que les personnes réalisant le classement prennent plus de précautions en classant les objets occupant les premières positions qu'en classant ceux de la fin. Il semble en effet que ce soit le cas par exemple pour les données des élections irlandaises : passés les deux premiers candidats, les électeurs voulant absolument classer tous les candidats ont tendance à classer les candidats restant suivant leur ordre d'apparition sur la liste électorale. Ces deux situations de probabilité de bon classement non constante pourraient donner naissance à deux nouveaux modèles pour données de rang, issue du modèle ISR.

7.2 analyse des données fonctionnelles

Les travaux de Delaigle et Hall [34] ont apporté une justification théorique à l'utilisation de la densité de probabilité des composantes principales, résultantes d'une analyse en composantes principales fonctionnelle, pour définir des modèles de mélange pour données fonctionnelles. A partir de cela, nous avons proposé des modèles de classification automatique pour les données fonctionnelles multivariées.
Comme nous l'avons vu, la convergence des algorithmes d'estimation proposés doit encore être étudiée théoriquement. Mais cette étude théorique a plus d'applications que les seuls modèles que nous avons proposés. En effet, le même problème de convergence apparaît dans les modèles de mélange faisant intervenir un paramètre de régularisation ajusté au fur et à mesure des itérations de l'algorithme. C'est le cas par exemple de la méthode HDDC [19] mais également de bon nombre d'autres méthodes de classification en grande dimension [77, 79, 81, 102]. L'idée que nous pensons développer est de montrer que, sous certaines conditions d'évolution des paramètres de régularisation, l'algorithme considéré est un algorithme de type EM généralisé [37], dont les propriétés assurent la convergence.
Nous ambitionnons également de nous intéresser au cas des variables aléatoires fonctionnelles qualitatives, que l'on rencontre par exemple lorsque l'on suit l'état d'un patient au cours du temps ou le statut marital d'un individu au cours de sa vie. Là encore l'idée est de s'appuyer sur les méthodes factorielles pour de tels processus à valeurs dans un espace qualitatif [38].

7.3 intégration de données hétérogènes

[CN18] Jacques J. *Classification automatique de données hétérogènes*. 44èmes Journées de Statistique de la SFDS, Bruxelles, mai 2012.

Une autre thématique qui m'intéresse particulièrement, et qui n'a pas été abordée dans ce mémoire, est l'intégration de données hétérogènes, c'est-à-dire de données de natures différentes. En effet, nous avons développé des algorithmes de classification automatique pour des données diverses (de rang, ordinales, fonctionnelles), mais comment prendre en compte ces données dans un même modèle ? Par exemple, dans le domaine médical, les données sont souvent constituées de paramètres physiologiques quantitatifs (âge, poids, taille...), qualitatifs (sexe, fumeur ou non fumeur...), mais également de données fonctionnelles (ECG, courbes de température...) voir de données ordinales lorsqu'on demande au patient d'évaluer un niveau de douleur sur une échelle ordinale.
Il a été montré qu'une approche naïve supposant l'indépendance de variables qualitatives et quantitatives apportait déjà des résultats très intéressants ([83],[CN18]). Outre la possibilité d'uniformiser les variables en passant par une analyse factorielle de données mixtes [59, 84], un certain nombre de travaux ont été développés récemment, basés sur :

- l'utilisation de modèles experts, établissant une dépendance entre les variables de différents types : Gormley et Murphy [53] font dépendre la loi de variables de rang de covariables quantitatives tandis que Ng et McLachlan [83] définissent la loi de variables continues conditionnellement à des variables qualitatives,

- le recours à des variables latentes : Browne et McNicholas [23] considèrent que les variables de différentes natures sont toutes issues de variables latentes continues,
- l'utilisation d'une méthode à noyau permettant de plonger des variables de différentes natures dans un même espace de Hilbert à noyau auto-reproduisant [18].

Ces techniques engendrent plusieurs questions. Les modèles experts utilisés induisent une asymétrie de traitement des données de différentes natures, et il serait intéressant de définir une dépendance mutuelle entre variables de différentes natures, en faisant attention aux problèmes d'identifiabilité que cela induit. Les travaux de Bouveyron *et al*. [18] plongeant les variables dans un même espace fonctionnel, impliquent la question du choix de la loi de probabilité à utiliser dans cet espace. Les travaux du chapitre 4 peuvent-ils fournir une réponse efficace ? L'hypothèse de normalité des composantes principales que nous avons faite est-elle vraisemblable dans une telle situation ?

7.4 PACKAGES R

Enfin, dans le domaine de recherche à but applicatif qui est le mien, l'apprentissage statistique, je pense que le développement de nouveaux modèles doit être accompagné d'un outil logiciel robuste, efficace et libre de droit, afin de pouvoir être utilisé par les praticiens de différentes communautés. Le logiciel **R**, à travers la possibilité de diffuser ses propres packages, est à mon avis le logiciel idéal. Aussi, je m'attache actuellement à diriger la création de trois packages, implémentés en C++ dans un soucis d'efficacité (en temps de calculs) :

- un package dédié à l'analyse des données de rang, proposant, outre la modélisation d'un jeu de données de rangs par le modèle ISR présenté dans le chapitre 3, un algorithme de classification automatique de données de rang permettant de prendre en compte les données multivariées ainsi que les données de rang partielles.
- un package dédié à l'analyse des données ordinales, proposant également un algorithme de classification automatique de données ordinales multivariées.
- un package dédié à la classification non supervisée de données hétérogènes (qualitatives et quantitatives), utilisant des modèles d'indépendance des variables qualitatives et quantitatives conditionnellement à la classe d'appartenance.

Il est à noter qu'à ce jour, aucun package **R** n'est dédié à la classification automatique de données de rang ou de données ordinales.
Un dernier package dédié à la classification automatique de données fonctionnelles est également à l'étude, mais avant cela, les points théoriques évoqués en 7.2 doivent être résolus afin de rendre plus robustes les algorithmes présentés dans le chapitre 4.

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

[2] A. Agresti. *Analysis of Ordinal Categorical Data*. Wiley, second edition, 2010.

[3] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6) :716–723, 1974.

[4] D. M. Allen. The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, 16 :125–127, 1974.

[5] R. Arnold, A. Nallapati and W.W. Cohen. A comparative study of methods for transductive transfer learning. In IEEE Computer Society, editor, *7th IEEE International Conference on Data Mining Workshops*, pages 77–82, Washington, DC, USA, 2007.

[6] M. Bacchiani and B. Roark. Adaptation of maximum entropy classifier : Little data can help a lot. In *Conference on Empirical Methods in Natural Language Processing*, 2003.

[7] J.D. Banfield and A.E. Raftery. Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49 :803–821, 1993.

[8] M. Barker and W. Rayens. Partial least squares for discrimination. *J. Chemometrics*, 17 :166–173, 2003.

[9] W. Benter. Computer-based horse race handicapping and wagering systems : A report. In W.T. Ziemba, V.S. Lo, and D.B. Haush, editors, *Efficiency of racetrack betting markets*. London : Academic Press, 1994.

[10] C. Bernard-Michel, S. Douté, M. Fauvel, L. Gardes, and S. Girard. Retrieval of Mars surface physical properties frim OMEGA hyperspectral images using regularized sliced inverse regression. *Journal of Geophysical Research*, 114 :E06005, 2009.

[11] J.R. Berrendero, A. Justel, and M. Svarc. Principal components for multivariate functional data. *Computational Statistics and Data Analysis*, 55 :2619–2634, 2011.

[12] J.-P. Bibring and 42 co-authors. *OMEGA : Observatoire pour la Minéralogie, l'Eau, les Glaces et l'Activité*, page 37 49. ESA SP-1240 : Mars Express : the Scientific Payload, 2004.

[13] C. Biernacki, F. Beninel, and V. Bretagnolle. A generalized discriminant rule when training population and test population differ on their descriptive parameters. *Biometrics*, 58(2) :387–397, 2002.

[14] C. Biernacki, G. Celeux, G. Govaert, and F. Langrognet. Model-based cluster and discriminant analysis with the mixmod software. *Computational Statistics and Data Analysis*, 51 :587–600, 2006.

[15] U. Bockenholt. Multivariate thurstonian models. *Psychometrika*, 55(2) :391–403, 1990.

[16] U. Böckenholt. Applications of Thurstonian models to ranking data. In *Probability models and statistical analyses for ranking data (Amherst, MA, 1990)*, volume 80 of *Lecture Notes in Statist.*, pages 157–172. Springer, New York, 1993.

[17] H.D. Bondell and B.J. Reich. Simultaneous regression shrinkage, variable selection and supervised clustering of predictors with oscar. *Biometrics*, 64 :115–123, 2008.

[18] C. Bouveyron, M. Fauvel, and S. Girard. Kernel discriminant analysis and clustering with parsimonious gaussian process models. Technical report, Laboratoire SAMM, Université Paris 1 Panthéon-Sorbonne, 2012.

[19] C. Bouveyron, S. Girard, and C. Schmid. High Dimensional Data Clustering. *Computational Statistics and Data Analysis*, 52 :502–519, 2007.

[20] C. Bouveyron, S. Girard, and C. Schmid. High dimensional discriminant analysis. *Comm. Statist. Theory Methods*, 36(14) :2607–2623, 2007.

[21] R.A. Bradley and M.E. Terry. Rank analysis of incomplete block designs. I. The method of paired comparisons. *Biometrika*, 39 :324–345, 1952.

[22] V. Bretagnolle. personnal communication. 2006.

[23] R.P. Browne and P.D. McNicholas. Model-based clustering, classification, and discriminant analysis of data with mixed type. *Journal of Statistical Planning and Inference*, 142 :2976–2984, 2012.

[24] R. Cattell. The scree test for the number of factors. *Multivariate Behaviour Research*, 1(2) :245–276, 1966.

[25] G. Celeux and J. Diebolt. The SEM algorithm : a probabilistic teacher algorithm from the EM algorithm for the mixture problem. *Computational Statistics Quaterly*, 2(1) :73–92, 1985.

[26] G. Celeux and G. Govaert. Clustering criteria for discrete data and latent class models. *Journal of Classification*, 8 :157–176, 1991.

[27] G. Celeux and G. Govaert. Parsimonious gaussian models in cluster analysis. *Pattern Recognition*, 28 :781–793, 1995.

[28] G. Celeux, M. Hurn, and C. Robert. Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, 95 :957–970, 2000.

[29] J-M. Chiou and P-L. Li. Functional clustering and identifying substructures of longitudinal data. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 69(4) :679–699, 2007.

[30] D. E. Critchlow. *Metric methods for analyzing partially ranked data*, volume 34 of *Lecture Notes in Statistics*. Springer-Verlag, Berlin, 1985.

[31] S. Dabo-Niang. Density estimation by orthogonal series in an infinite dimensional space : application to processes of diffusion type I. *Journal of Nonparametric Statistics*, 16(1-2) :171–186, 2004. The International Conference on Recent Trends and Directions in Nonparametric Statistics.

[32] S. Dabo-Niang. Kernel density estimator in an infinite-dimensional space with a rate of convergence in the case of diffusion process. *Applied Mathematics Letters. An International Journal of Rapid Publication*, 17(4) :381–386, 2004.

[33] H. Daumé III and D. Marcu. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research.*, 26 :101–126, 2006.

[34] A. Delaigle and P. Hall. Defining probability density for a distribution of random functions. *The Annals of Statistics*, 38 :1171–1193, 2010.

[35] A. D'Elia and D. Piccolo. A mixture model for preferences data analysis. *Computational Statistics and Data Analysis*, 49(3) :917–934, 2005.

[36] O. Delrieu and Bowman C. Visualizing gene determinants of disease in drug discovery. *Pharmacogenomics*, 7(3) :311–329, 2006.

[37] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data (with discussion). *Journal of the Royal Statistical Society. Series B*, 39 :1–38, 1977.

[38] J-C. Deville and G. Saporta. *Data Analyss and Informatics*, chapter Analyse harmonique qualitative, pages 375–389. North-Holland, 1980.

[39] J. Diebolt and C. Robert. Estimation of finite mixture distributions through bayesian sampling. *Journal of the Royal Statistical Society. Series B*, 56(2) :363–375, 1994.

[40] B. S. Everitt. *An introduction to latent variable models*. Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1984.

[41] W. Fan, I. Davidson, B. Zadrozny, and P.S. Yu. An improved categorization of classifier's sensitivity on sample selection bias. In *Fifth IEEE Conference on Data Mining*, 2005.

[42] F. Ferraty and P. Vieu. *Nonparametric functional data analysis*. Springer Series in Statistics. Springer, New York, 2006.

[43] M.A. Fligner and J.S. Verducci. Distance based ranking models. *J. Roy. Statist. Soc. Ser. B*, 48(3) :359–369, 1986.

[44] M.A. Fligner and J.S. Verducci. Multistage ranking models. *J. Amer. Statist. Assoc.*, 83(403) :892–901, 1988.

[45] S. Frühwirth-Schnatter and S. Kaufmann. Model-based clustering of multiple time series. *Journal of Business and Economic Statistics*, 26 :78–89, 2008.

[46] A. Geman and D. Geman. Stochastic relaxation, gibbs distributions and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Matching Intelligence*, 6 :721–741, 1984.

[47] G. Giordan, M. et Diana. A clustering method for categorical ordinal data. *Communications in Statistics – Theory and Methods*, 40 :1315–1334, 2011.

[48] M. Goldfeld and R.E. Quandt. A markov model for switching regressions. *Journal of Econometrics*, 1 :3–16, 1973.

[49] R.C. Gonzalez and R.E. Woods. *Digital Image Processing*. Addison Wesley, 1992.

[50] I.C. Gormley and T.B. Murphy. Analysis of Irish third-level college applications data. *J. Roy. Statist. Soc. Ser. A*, 169(2) :361–379, 2006.

[51] I.C. Gormley and T.B. Murphy. A latent space model for rank data. In *Proceedings of the 23th International Conference on Machine Learning*, Pittsburgh, PA, 2006.

[52] I.C. Gormley and T.B. Murphy. Exploring voting blocs within the irish electorate : A mixture modeling approach. *J. Amer. Statist. Assoc.*, 103(483) :1014–1027, 2008.

[53] I.C. Gormley and T.B. Murphy. A mixture of experts model for rank data with applications in election studies. *Annals of Applied Statistics*, 2(4) :1452–1477, 2008.

[54] C. Gouget. *Utilisation des modèles de mélange pour la classification automatique de données ordinales*. PhD thesis, Université de Technologie de Compiègne, 2006.

[55] J.A. Hartigan and M.A. Wong. Algorithm as 1326 : A k-means clustering algorithm. *Applied Statistics*, 28 :100–108, 1978.

[56] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. Springer Series in Statistics. Springer-Verlag, New York, 2009.

[57] S.W. Hell and J. Wichmann. Breaking the diffraction resolution limit by stimulated emission : stimulated-emission-depletion fluorescence microscopy. *Optics Letters*, 19(11), 1994.

[58] C. Hennig. *Classification in the Information Age*. Springer-Verlag, Heidelberg, 1999.

[59] M. Hill and J. Smith. Principal component analysis of taxonomic data with multi-state discrete characters. *Taxon*, 25 :249–255, 1976.

[60] J-I. Hotta, E. Fron, P. Dedecker, K.P.F. Janssen, C. Li, K. Mullen, B. Harke, J. Buckers, S.W. Hell, and J. Hofkenst. Spectroscopic rationale for efficient stimulated-emission depletion microscopy fluorophores. *Journal of the American Chemical Society*, 135 :5021–5023, 2010.

[61] J. Huang, A. Smola, A. Gretton, K.M. Borgwardt, and B. Scholkopf. Correcting sample selection bias by unlabeled data. In *19th Annual Conference on Neural Information Processing Systems*, 2007.

[62] M. Hurn, A. Justel, and C.P. Robert. Estimating mixtures of regressions. *Journal of Computational and Graphical Statistics*, 12(1) :55–79, 2003.

[63] F. Ieva, A.M. Paganoni, D. Pigoli, and V. Vitelli. ECG signal reconstruction, landmark registration and functional classification. In *7th Conference on Statistical Computation and Complex System*, Padova, 2011.

[64] G.M. James and C.A. Sugar. Clustering for sparsely sampled functional data. *J. Amer. Statist. Assoc.*, 98(462) :397–408, 2003.

[65] M.G. Kendall. A new measure of rank correlation. *Biometrika*, 30 :81–93, 1938.

[66] M.G. Kendall and B.B. Smith. On the method of paired comparisons. *Biometrika*, 31 :324–345, 1940.

[67] D.E. Knuth. *The Art of Computer Programming*, volume 3 : Sorting and Searching. Addison-Wesley Professional, second edition, 1998.

[68] N.D. Lawrence and J.C. Platt. Learning to learn with the informative vector machine. In *Proceedings of the 21th International Conference on Machine Learning*, Banff, Alberta, Canada, 2004.

[69] G. Lebanon and J. Lafferty. Cranking : Combining rankings using conditional models on permutations. In *Proceedings of the 19th International Conference on Machine Learning*, Sydney, Australia, 2002.

[70] G. Lebanon and Y. Mao. Non-parametric modeling of partially ranked data. *J. Mach. Learn. Res.*, 9 :2401–2429, 2008.

[71] C. Lévéder, P.A. Abraham, E. Cornillon, E. Matzner-Lober, and N. Molinari. Discrimination de courbes de prétrissage. In *Chimiométrie 2004*, pages 37–43, Paris, 2004.

[72] A. Lourme and C. Biernacki. Simultaneous Gaussian Model-Based Clustering for Samples of Multiple Origins. *Computational Statistics*, in press, 2012.

[73] R.D. Luce. *Individual choice behavior : A theoretical analysis*. John Wiley & Sons Inc., New York, 1959.

[74] J.I. Marden. *Analyzing and modeling rank data*, volume 64 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London, 1995.

[75] G. McLachlan and T. Krishnan. *The EM algorithm and extensions*. Wiley Interscience, New York, 1997.

[76] G. McLachlan, D. Peel, and R. Bean. Modelling high-dimensional data by mixtures of factor analyzers. *Comput. Statist. Data Anal.*, 41 :379–388, 2003.

[77] G. McLachlan, D. Peel, and R. Bean. Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics and Data Analysis*, 41 :379–388, 2003.

[78] G.J. McLachlan and D. Peel. *Finite Mixture Models*. Wiley, New York, 2000.

[79] P. McNicholas and B. Murphy. Parsimonious gaussian mixture models. *Statistics and Computing*, 18(3) :285–296, 2008.

[80] N. Molinari, J-F. Durand, and R. Sabatier. Bounded optimal knots for regression splines. *Computational Statistics and Data Analysis*, 45 :159–178, 2004.

[81] A. Montanari and C. Viroli. Heteroscedastic factor mixture analysis. *Statistical Modeling : An International journal*, 10(4) :441–460, 2010.

[82] T.B. Murphy and D. Martin. Mixtures of distance-based models for ranking data. *Comput. Statist. Data Anal.*, 41(3-4) :645–655, 2003.

[83] S.K. Ng and G.J. McLachlan. *Machine Learning Research Progress*, chapter Expert networks with mixed continuous and categorical feature variables : a location modeling approach, pages 355–368. Nova, Hauppauge, New York, 2010.

[84] J. Pagès. Analyse factorielle de données mixtes. *Revue Statistique Appliquée*, LII(4) :93–111, 2004.

[85] S.J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10) :1345–1359, 2010.

[86] R.L. Plackett. The analysis of permutations. *J. Roy. Statist. Soc. Ser. C Appl. Statist.*, 24(2) :193–202, 1975.

[87] C. Preda, G. Saporta, and C. Lévéder. PLS classification of functional data. *Comput. Statist.*, 22(2) :223–235, 2007.

[88] J. O. Ramsay and B. W. Silverman. *Functional data analysis*. Springer Series in Statistics. Springer, New York, second edition, 2005.

[89] S. Ray and B. Mallick. Functional clustering by Bayesian wavelet methods. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 68(2) :305–332, 2006.

[90] C. Robert. *The Bayesian Choice*. Springer, 2007.

[91] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2) :461–464, 1978.

[92] Y. She. Sparse regression with exact clustering. *Electronic Journal of Statistics*, 4 :1055–1096, 2010.

[93] H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2) :227–244, 2000.

[94] Z. Šidák. Rectangular confidence regions for the means of multivariate normal distributions. *J. Amer. Statist. Assoc.*, 62 :626–633, 1967.

[95] A. Storkey and M. Sugiyama. *Mixture regression for covariate shift*, pages 1337–1344. Advances in Neural Information Processing Systems 19. MIT Press, Cambridge, 2007.

[96] M. Sugiyama. Active learning in approximately linear regression based on conditional expectation of generalization error. *Journal of Machine Learning Research*, 7 :141–166, 2006.

[97] M. Sugiyama and K-R. Müller. Input-dependent estimation of generalization error under covariate shift. *Statistics & Decisions*, 23, 2005.

[98] M. Sugiyama and Krauledat M. Müller, K-R. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8 :985–1005, 2007.

[99] E-G. Talbi. *Metaheuristics : From Design to Implementation*. Wiley, 2009.

[100] T. Tarpey and K.J. Kinateder. Clustering functional data. *J. Classification*, 20(1) :93–114, 2003.

[101] L.L. Thurstone. A law of comparative judgment. *Psychological Review*, 79 :281–299, 1927.

[102] M. E. Tipping and C. Bishop. Mixtures of principal component analyzers. *Neural Computation*, 11(2) :443–482, 1999.

[103] R.D. Tuddenham and M.M. Snyder. Physical growth of california boys and girls from birth to eighteen years. *Univ. Calif. Publ. Chld Devlpmnt*, 1 :188–364, 1954.

[104] V. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

[105] A. Vignal, D. Milan, M. SanCristobal, and A. Eggen. A review on snp and other types of molecular markers and their use in animal genetics. *Genetics Selection Evolution*, 34(3) :275–305, 2002.

[106] J.H. Ward. Hierarchical grouping to optimize an objective function. *J. Amer. Statist. Assoc.*, 58 :236–244, 1963.

[107] S. Wold. Pattern recognition by means of disjoint principal component models. *Patt. Recogn.*, 8 :127–139, 1976.

[108] B. Zadrozny. Learning and evaluating classifiers under sample selection bias. In *21st International Conference on Machine Learning*, 2004.

[109] Z. Zhang, E. Ersoz, C-Q. Lai, R.J. Todhunter, H.K. Tiwari, M.A. Gore, P.J. Bradbury, J. Yu, D.K. Arnett, J.M. Ordovas, and E.S. Buckler.

Contribution à l'apprentissage statistique
à base de modèles génératifs pour données complexes

Résumé : Ce mémoire synthétise les activités de recherche que j'ai menées de 2005 à 2012, sur la thématique de l'apprentissage statistique des données complexes, abordée par le biais de modèles probabilistes paramétriques génératifs. Plusieurs types de données complexes sont considérées. Les données issues de populations différentes ont été abordées en proposant des modèles de lien paramétriques entre populations, permettant d'adapter les modèles statistiques d'une population vers l'autre, en évitant une lourde collecte de nouvelles données. Les données de rang, définissant un classement d'objets selon un ordre de préférence, les données ordinales, qui sont des données qualitatives ayant des modalités ordonnées, et les données fonctionnelles, où l'observation statistique consiste en une ou plusieurs courbes, ont également été étudies. Pour ces trois types de données, des modèles génératifs probabilistes ont été définis et utilisés en classification automatique de données multivariées. Enfin les données de grande dimension, que l'on rencontre lorsque le nombre de variables du problème dépasse celui des observations, ont été étudiées dans un cadre de régression. Deux approches, fruits de deux thèses de doctorat que je co-encadre, sont proposés : l'une utilisant des algorithmes d'optimisation combinatoire pour explorer de façon efficace l'espace des variables, et l'autre définissant un modèle de régression regroupant ensemble les variables ayant un effet similaire.

Mots-clefs : apprentissage statistique, apprentissage adaptatif, modèles génératifs, données de rang, données ordinales, données fonctionnelles, grande dimension, classification automatique.

Abstract: This manuscript presents my research activities, which mainly focus on designing parametric, parsimonious and meaningful generative models for complex data. Several kinds of complex data have been studied. Data sampled from different populations (transfer learning) has been addressed by designing parametric models for the link between the different populations. Thus, statistical models can be adapted from one population to another one by sparing a large collect of new data. Ranking data, which results from ranking of objects by a judge according to a preference order, ordinal data, which are categorical data with ordered categories, and functional data, in which the statistical unit consists of one or several curves, have also been studied. For this three kinds of complex data, generative models have been developed and used for the clustering of multidimensional data. The last kind of complex data, high dimensional data, has been studied in a regression context. In this domain, two approaches are proposed by two Ph.D. students I co-supervise: the first one uses combinatorial optimization algorithms in order to efficiently explore the feature space and the second one defines a regression model in which the variables having a similar effect on the output are grouped together.

Key-words: statistical learning, transfer learning, generative model, ranking data, ordinal data, functional data, high dimensional problem, clustering.