N° d'ordre : 3871

THÈSE

présentée à

L'UNIVERSITÉ BORDEAUX I

ÉCOLE DOCTORALE DE MATHÉMATIQUES ET INFORMATIQUE

par **Vanessa KUENTZ**

POUR OBTENIR LE GRADE DE

DOCTEUR

SPÉCIALITÉ : **Mathématiques Appliquées - Statistique**

****

Contributions à la réduction de dimension

****

Soutenue le 20 novembre 2009 à l'Institut de Mathématiques de Bordeaux

Devant la commission d'examen composée de :

M. Bernard BERCU	PROF.	Université Bordeaux I	Président du jury
Mme Marie CHAVENT	MCF-HDR.	Université Victor Segalen Bordeaux II	Codirectrice de thèse
M. François HUSSON	MCF-HDR.	Agrocampus Ouest, Rennes	Examinateur
M. Henk H.A.L. KIERS	PROF.	University of Groningen	Rapporteur
M. Jean-Michel POGGI	PROF.	Université Paris-Sud	Rapporteur
M. Gilbert SAPORTA	PROF.	CNAM Paris	Rapporteur
M. Jérôme SARACCO	PROF.	Université Montesquieu Bordeaux IV	Directeur de thèse

- 2009 -

A toi papa,

Remerciements

Je tiens tout d'abord à remercier très chaleureusement mes deux directeurs de thèse, Marie Chavent et Jérôme Saracco, pour leur disponibilité, leurs remarques pertinentes, leurs précieux conseils et les nombreuses connaissances statistiques qu'ils m'ont permis d'acquérir. La façon dont ils ont codirigé cette thèse m'a permis de m'épanouir durant ces trois années. J'ai toujours senti un soutien moral, un intérêt pour mon travail, une attention particulière et une ambiance de travail toujours placée sous le signe de la bonne humeur. Tout cela a été une grande source de motivation, qui m'a donné confiance en moi et m'a permis d'évoluer sereinement dans le monde de la Recherche.

Je remercie également très sincèrement Henk Kiers, Professeur à l'Université de Groningen (Pays-Bas), Jean-Michel Poggi, Professeur à l'Université Paris-Sud et Gilbert Saporta, Professeur au CNAM à Paris, pour avoir accepté d'être les rapporteurs de mon travail de thèse. Je les remercie pour le temps qu'ils ont accordé à la relecture de ce manuscrit et pour l'intérêt qu'ils lui ont accordé.

Merci à Bernard Bercu, professeur à l'Université Bordeaux 1, et François Husson, Maître de Conférences à Agrocampus Ouest (Rennes), pour m'avoir fait l'honneur de participer à mon jury de thèse.

Merci infiniment à Chantal Lacomblez sans qui cette merveilleuse aventure qu'a été ma thèse n'aurait pas eu lieu.

Je remercie vivement tous les professeurs de l'UFR Sciences et Modélisation de l'Université Bordeaux 2 qui m'ont vu "naître" il y a déjà huit ans. Merci pour la qualité de l'enseignement dont j'ai bénéficié et pour leur amitié. Un merci tout particulier à Brigitte, Bedr'Eddine, Manue, Christine, Vincent, Fred', Olivier et Jean-Baptiste.

Je remercie Gridou pour son amitié, ses fous rires et sa bonne humeur. Merci également à Pierrick, Jessica, Danaëlle et Vincent pour leur amitié. Les soirées passées ensemble ont été des moments de détente et de plaisir indispensables à la bonne réussite de ma thèse.

Je voudrais également remercier Benoît pour sa sympathie et son entrain, travailler avec lui est un réel plaisir.

Merci aux doctorants de Bordeaux 1 qui sont d'ailleurs tous pour la plupart déjà docteurs. Je pense à Bertrand, Fabien, Julien, Delphine et Aubin, Jean. Une mention particulière pour Ludi qui m'entoure de son amitié. Merci pour son humour, les discussions que nous avons et tous les bons moments passés ensemble.

Enfin mes plus grands remerciements vont à ma famille sans qui je n'écrirais pas ces lignes aujourd'hui. Merci à ma maman et ma soeurette pour l'amour qu'elles me donnent et le soutien qu'elles m'apportent. Elles ont été un élément clé dans la réussite de ma thèse. Merci à Jean de faire partie de ma vie. Sa présence, ses conseils, ses encouragements et son réconfort dans les moments de doute sont pour moi très précieux. Merci à mes grands-parents pour leur amour. Merci également à Marie-Jeanne et Alain pour leur gentillesse.

Table des matières

Chapitre 1

Présentation générale

1.1 Description de la thèse

Le travail développé dans ce mémoire de thèse s'articule autour de la réduction de dimension. Cette thématique centrale en Statistique consiste à rechercher des sous-espaces de faibles dimensions tout en minimisant la perte d'information statistique. En effet, avec l'évolution des moyens informatiques, les bases de données sont de plus en plus grandes. Ainsi certaines variables apportent parfois le même type d'information et cette redondance peut masquer des phénomènes intéressants. Lors d'analyses ultérieures, il est alors plus aisé et souvent plus judicieux de travailler avec des variables synthétiques résumant l'information principale. Selon la problématique et l'objectif poursuivi, plusieurs approches sont possibles. Dans ma thèse, j'ai successivement adopté une approche exploratoire (statistique multidimensionnelle), et une approche "modélisation" (avec comme cadre de référence un modèle semiparamétrique de régression). Remarquons que les méthodes et algorithmes proposés dans cette thèse ont été implementés sous le langage de programmation `R`.

Dans la suite de ce mémoire, les travaux que j'ai réalisés durant ma thèse sont regroupés en trois chapitres : le Chapitre 2 concerne les méthodes de type "clustering" (ou classification non supervisée), le Chapitre 3 porte sur la régression inverse par tranchage (SIR, pour Sliced Inverse Regression) et le Chapitre 4 regroupe différentes applications en statistique multidimensionnelle réalisées dans le cadre de collaborations universitaires ou industrielles et du doctorat-conseil. Chaque chapitre est organisé de la façon suivante : tout d'abord une section synthétise les travaux réalisés dans la thématique du chapitre, ensuite chaque section contient un article à paraître ou soumis à l'heure actuelle. Ce format permet ainsi de lire les différentes sections d'un chapitre indépendamment les unes des autres. Décrivons maintenant plus précisément ces trois chapitres principaux.

Dans le Chapitre 2 s'intitulant "Réduction de dimension via des méthodes de statistique multidimensionnelle", je me suis intéressée à la réduction de dimension dans le cas de variables qualitatives.

Dans un premier temps, j'aborde le problème de la rotation en Analyse des Correspondances Multiples (ACM). Cette méthode factorielle, qui permet une description statistique multidimensionnelle de données qualitatives, peut parfois fournir des composantes principales difficiles à interpréter. Nous avons donc voulu définir une étape supplémentaire de rotation afin de faciliter la lecture des résultats. Pour cela, nous avons utilisé un algorithme pratique de rotations planaires successives de paires de facteurs. Notre principale contribution a été la définition de l'expression analytique de l'angle optimal pour le critère de rotation choisi. Les variables étant qualitatives, ce critère de rotation est basé sur les rapports de corrélation entre les variables et les composantes principales. Ainsi l'utilisation de la rotation en ACM permet d'obtenir des composantes principales plus clairement reliées aux variables et donc implicitement des groupes de variables. Une collaboration avec Voies Navigables de France (VNF), qui est décrite dans la section 4.1, nous a permis d'illustrer les intérêts pratiques de la rotation en ACM.

Dans la section suivante, j'étudie la classification de variables qui est une alternative pour la recherche de composantes interprétables puisqu'elle permet d'organiser les variables en groupes homogènes afin de faire ressortir une structure. Le praticien peut alors sélectionner une variable, ou construire une variable synthétique, dans chaque groupe. Une approche simple et fréquemment utilisée pour obtenir une partition des variables est de définir une matrice de dissimilarités entre ces variables et d'appliquer ensuite des méthodes de classification sur tableaux de dissimilarités. Cette approche utilisant des algorithmes définis pour la classification d'observations, peut être appliquée aussi bien pour des variables quantitatives que qualitatives. Parallèlement, certaines méthodes ont été développées spécifiquement pour la classification de variables, cependant, à notre connaissance, peu sont capables de gérer des données qualitatives. Dans les méthodes de classification de variables qualitatives que nous avons proposées, le critère d'homogénéité d'une classe est égal à la somme des rapports de corrélation entre les variables de la classe et une variable latente quantitative. Nous avons alors montré que la variable latente maximisant ce critère peut être obtenue en réalisant une ACM des variables de la classe. Dans ce cadre, différents algorithmes (de type nuées dynamiques, classification hiérarchique ascendante et descendante) ont été proposés. Afin d'illustrer et de comparer empiriquement les différents algorithmes proposés, nous avons utilisé le jeu de données réelles issu de la collaboration avec VNF.

Dans le Chapitre 3 intitulé "Réduction de dimension via la méthode SIR (Sliced Inverse Regression)", je me suis intéressée à un modèle de régression semiparamétrique et plus particulièrement à la méthode de régression inverse par tranchage (SIR). La régression étudie la relation entre une covariable et une variable réponse. En régression paramétrique, la fonction de lien est une fonction algébrique simple des variables explicatives et des méthodes de type moindres carrés ou maximum de vraisemblance, entre autres, peuvent être utilisées pour trouver le meilleur ajustement global. En régression nonparamétrique, la classe de fonctions ajustées est élargie pour obtenir davantage de flexibilité, un ajustement local peut être obtenu par des procédures de lissage sophistiquées (méthodes à noyau ou splines de lissage par exemple). Mais lorsque la dimension

de la covariable devient importante, le nombre d'observations nécessaires pour le lissage local croît de manière exponentielle avec cette dimension. Pour surmonter ce "fléau de la dimension", des modèles semiparamétriques de régression ont été développés. Nous nous intéressons spécifiquement à la méthode SIR introduite par Li (1991). Soient une variable dépendante y et une variable explicative $\mathbf{x}$ multidimensionnelle. Une possibilité est de s'intéresser à la distribution conditionnelle de y sachant $\mathbf{x}$. Les méthodes de réduction de dimension supposent que l'on peut remplacer $\mathbf{x}$ par un vecteur de dimension inférieure $(\mathbf{x}'\beta_1, \ldots, \mathbf{x}'\beta_K)$, avec $K < p$, sans perdre d'information sur le lien entre $\mathbf{x}$ et y. On cherche alors à estimer une base du sous-espace de réduction de dimension effective (engendré par les vecteurs $\beta_k, k = 1, \ldots, K$). L'idée de base des méthodes SIR est d'échanger le rôle de $\mathbf{x}$ et y (afin de réduire la dimension du problème) et d'étudier les moments conditionnels de $\mathbf{x}$ sachant y. Un tranchage est alors réalisé sur la variable dépendante pour faciliter l'estimation de ces moments. Les méthodes SIR sont basées sur une propriété géométrique de la courbe de régression inverse qui repose sur une condition cruciale de linéarité de la distribution de la covariable. Cependant cette condition est difficile à vérifier et est souvent violée en pratique lorsqu'on travaille sur de vraies données. Théoriquement, l'utilisation de SIR pour estimer le sous-espace de réduction de dimension effective n'est alors plus justifiée.

Nous avons proposé une approche qui permet de gérer cette situation. L'idée est de partitionner l'espace des prédicteurs pour que la condition de linéarité soit vérifiée dans chaque classe. Des estimateurs sont alors obtenus à l'aide de SIR dans chaque classe et finalement ces estimateurs sont combinés pour donner un estimateur du sous-espace de réduction de dimension effective utilisant la totalité de l'espace. Nous avons appelé "Cluster-based SIR" cette méthode. La convergence ainsi que la normalité asymptotique de l'estimateur ont été obtenues, et une étude sur des données simulées a démontré la supériorité de cette approche face à SIR. Enfin une application sur un jeu de données réelles, largement étudié dans la littérature sur SIR, a été faite et montre l'intérêt d'utiliser la méthode Cluster-based SIR.

Une autre adaptation de la méthode SIR, utilisant le bootstrap, est proposée pour améliorer l'estimation de la base du sous-espace de réduction de dimension effective lorsque le nombre d'observations de l'échantillon est faible ou lorsque le modèle est bruité. L'idée est de générer plusieurs échantillons bootstrap, d'estimer les directions de réduction de dimension effective dans chaque réplication, puis finalement de les combiner afin d'estimer le sous-espace recherché. Plusieurs façons de combiner les différents estimateurs ont été proposés et ont donné naissance à des méthodes de type "Bagging SIR". Nous avons obtenu des résultats de convergence asymptotique pour l'un de ces estimateurs. Le bon fonctionnement pratique des différentes approches a été illustré sur des simulations numériques.

Le Chapitre 4 intitulé "Applications et collaborations interdisciplinaires" présente les travaux que j'ai réalisés dans le cadre de collaborations universitaires, industrielles ou du doctorat-conseil fait chez Danone Research à Paris.

Le premier sujet sur lequel j'ai travaillé concerne une étude sur la pollution atmosphérique, qui a été réalisée en 2006 en réponse à l'Appel à Proposition de Recherche

"Particules" du programme PRIMEQUAL/PREDIT initié par le Ministère de l'Ecologie, du Développement et de l'Aménagement Durable et l'Agence de l'Environnement et de la Maîtrise de l'Energie. La question étudiée était l'identification et la quantification des contributions relatives des sources de poussières fines à un environnement. Dans la méthodologie que nous avons mise au point, l'utilisation d'une technique de rotation dans l'Analyse en Composantes Principales (ACP) s'est révélée pertinente. Ainsi mon premier article a consisté à dresser une présentation synthétique du modèle d'Analyse en Facteurs, peu décrit dans les ouvrages francophones mais au contraire très présent dans la littérature anglo-saxonne, afin de maîtriser les justifications théoriques de la rotation en ACP. Cet article a été publié dans la Revue *Modulad* et est disponible dans la section 4.2. Plus généralement, la méthodologie mise au point pour répondre à la problématique a donné lieu à deux articles, un à paraître dans *Environmetrics* (disponible dans la section 4.3) et un autre publié dans *Case Studies in Business, Industry and Government Statistics*.

Dans un second temps, je décris la collaboration avec des universitaires en Economie de l'Université Montpellier 1. L'objectif de ce travail était d'étudier l'impact de l'adoption en 2005 d'une nouvelle loi en comptabilité financière sur les incorporels en France. Afin de définir une typologie sur les entreprises en fonction de leur réaction face à la mise en place de cette loi, nous avons remplacé l'étape classique d'ACP préalable à la classification des observations, par une classification hiérarchique descendante des variables quantitatives. De plus, une procédure bootstrap, visant à étudier la stabilité des partitions construites, a permis de choisir un nombre de classes de variables. Ce travail a donné lieu à un article soumis dans une revue internationale de comptabilité financière que je tiens à disposition pour de plus amples détails.

Je présente ensuite le travail réalisé avec l'entreprise de Marketing Enform pour le traitement statistique d'une enquête de satisfaction commanditée par VNF. L'objectif était de mettre en lumière, à l'aide de méthodes statistiques descriptives et inférentielles, les grandes tendances d'opinion au sein des navigants plaisanciers sur le Canal des Deux Mers durant l'été et l'automne 2008. Il s'agissait ainsi de fournir à VNF un outil d'aide à la décision pour l'amélioration de l'offre de services. Le rapport synthétique qui a été remis à VNF est disponible sur demande.

Puis je décris le doctorat-conseil réalisé chez Danone Research à Paris durant l'année universitaire 2008-2009. L'objectif de ce travail était d'étudier les justifications théoriques de l'utilisation de méthodes multi-tableaux pour les études cliniques. Par ailleurs, l'utilisation de l'Analyse Factorielle Multiple a permis de mettre en exergue de nouveaux résultats ayant une lecture interprétable par les biométriciens.

Enfin pour clore ce chapitre de présentation générale, je donne, dans la section suivante, la liste des travaux que j'ai réalisés durant ma thèse.

1.2 Liste des travaux

1.2.1 Articles parus dans des revues à comité de lecture

Revues internationales à comité de lecture

Kuentz, V., Saracco, J. (2009). Cluster-based Sliced Inverse Regression. A paraître dans *Journal of the Korean Statistical Society.*

Kuentz, V., Liquet, B., Saracco, J. (2008). Bagging versions of Sliced Inverse Regression. A paraître dans *Communications in Statistics - Theory and Methods.*

Chavent, M., Guéguan, H., Kuentz, V., Patouille B., Saracco J. (2008). PCA and PMF based methodology for air pollution sources identification and apportionment. A paraître dans *Environmetrics.*

Chavent, M., Guégan, H., Kuentz, V., Patouille, B., Saracco, J. (2007). Apportionment of air pollution by source at a French urban site. *Case Studies in Business, Industry and Government Statistics*, Vol 1-Issue 2, 119-129.

Revues nationales à comité de lecture

Chavent, M., Kuentz, V., Saracco, J. (2007). Analyse en Facteurs : présentation et comparaison des logiciels SAS, SPAD et SPSS. *La revue Modulad*, 37, p 1-30.

1.2.2 Articles parus dans des actes de conférences avec comité de lecture

Conférences internationales *(Le nom de l'orateur est souligné.)*

Chavent, M., Kuentz, V., Saracco, J. (2009). A partitioning method for the clustering of categorical variables. *Classification and Data Analysis Group (CLADAG 2009).* Catania, Italy.

Chavent, M., Kuentz, V., Saracco, J. (2009). A partitioning method for the clustering of categorical variables. *Proceedings of the International Federation of Classification Societies (IFCS 2009), Springer-Verlag.* Dresden, Germany.

Kuentz, V., Saracco, J. (2008). A cluster-based approach for Sliced Inverse Regression. *XVIIIth Symposium of Computational Statistics (COMPSTAT2008).* Porto, Portugal.

Chavent, M., Kuentz, V., Saracco, J. (2008). Divisive hierarchical clustering of quantitative and qualitative variables. *Xth European Symposium on Statistical Methods for*

the Food Industry (AGROSTAT 2008), Louvain La Neuve, Belgium.

Chavent, M., Guéguan, H., Kuentz, V., Patouille, B., Saracco, J. (2007). Pollution sources detection via principal component analysis and rotation. *XIIth International Symposium of Applied Stochastic Models and Data Analysis (ASMDA 2007)*, Chania, Greece.

Conférences nationales

Chavent, M., Kuentz, V., Saracco, J. (2009). Une méthode de partitionnement pour la classification de variables qualitatives. *XVIèmes Rencontres de la Société Francophone de Classification, Grenoble.*

Chavent, M., Kuentz, V., Liquet, B. (2009). Données manquantes en ACM : l'algorithme NIPALS. *XVIèmes Rencontres de la Société Francophone de Classification, Grenoble.*

Chavent, M., Kuentz, V., Saracco, J. (2009). Une solution analytique pour la rotation planaire en Analyse Factorielle des Correspondances Multiples. *41èmes Journées de Statistique, SFdS, Bordeaux.*

Chavent, M., Kuentz, V. , Saracco, J. (2007). Une approche divisive de classification hiérarchique de variables quantitatives. *XIVèmes Rencontres de la Société Francophone de Classification*, ENST, Paris.

Chavent, M., Guégan, H., Kuentz, V., Patouille, B., Saracco, J. (2007). Quantification de sources de pollution sur un site urbain français. *39èmes Journées de Statistique*, SFdS, Angers.

1.2.3 Articles soumis

Chavent, M., Kuentz, V., Saracco, J. (2009). Rotation in Multiple Correspondence Analysis : a planar rotation iterative procedure.

Chavent, M., Kuentz, V., Saracco, J. (2009). Clustering of categorical variables around latent variables.

Bessieux Ollier C., Chavent M., Kuentz V., Walliser E. (2008). The effects of adopting mandatory IFRS on intangible assets : the case of France.

Chapitre 2

Réduction de dimension via des méthodes de statistique multidimensionnelle

Dans ce chapitre, la problématique de la réduction de dimension est abordée sous l'angle de la "statistique multidimensionnelle". Plus précisément je me suis intéressée à des méthodes exploratoires comme l'analyse factorielle et la classification non supervisée. Tout d'abord je présente une synthèse des travaux réalisés dans ce cadre, puis les articles (actuellement soumis) qui en ont découlé.

2.1 Synthèse des travaux

Les travaux présentés ici s'articulent autour de l'analyse d'un tableau de données multidimensionnelles et plus précisément autour du traitement de données qualitatives. Nous considérons donc une matrice de données où n observations sont décrites sur p variables qualitatives. Dans ce cadre, nous nous sommes intéressés à deux problèmes ayant trait à la réduction de dimension : la rotation en Analyse des Correspondances Multiples (ACM) et la classification de variables.

Pour des données quantitatives, l'Analyse en Composantes Principales (ACP) est une méthode classique de réduction de dimension. Elle construit un ensemble de composantes principales qui résument au mieux, en terme de variance du nuage de points, l'information contenue dans les données initiales. L'ACP peut aussi se définir comme un problème d'approximation de la matrice des données centrées réduites par le produit de deux matrices de rang inférieur : la matrice des scores (composantes principales standardisées) et la matrice des "loadings" (que l'on trouve sous le terme saturations dans quelques ouvrages en français mais qui est peu utilisé). Cette approximation n'est pas unique mais déterminée à une rotation près. Cette liberté de changement de base permet d'obtenir de nouvelles composantes principales après rotation plus facilement interprétables. Les critères de rotation utilisés sont généralement basés sur les corrélations entre les variables (quantitatives) et les composantes principales, l'idée étant

d'obtenir après rotation des corrélations au carré proches de 0 ou de 1. Mon travail a donc consisté à traiter cette question de la rotation dans le cas de variables qualitatives, c'est-à-dire en ACM. Pour cela, nous avons considéré l'ACM comme une ACP pondérée de la matrice des profils lignes du tableau disjonctif complet construit à partir des données qualitatives. Nous avons ensuite utilisé un critère de rotation basé non plus sur des corrélations comme en ACP, mais sur des rapports de corrélation entre les variables (qualitatives) et les composantes principales. Nous avons pu déterminer une solution exacte à l'optimisation de ce critère de rotation dans le cas planaire (deux dimensions). Nous avons ensuite utilisé cette solution dans une procédure de rotation planaire itérative, permettant d'effectuer une rotation dans le cas où le nombre de composantes principales retenues à l'issue de l'ACM est supérieur à deux.

Ce travail sur la rotation en ACM a été réalisé en lien avec celui concernant la classification de variables qualitatives. L'utilisation de la rotation en ACM permet en effet d'obtenir des composantes principales plus clairement reliées ou non aux variables et donc implicitement des groupes de variables. En pratique, une partition de l'ensemble des variables peut être obtenue en associant chaque variable à la composante principale avec laquelle son rapport de corrélation est le plus grand.

Notre travail sur la classification de variables qualitatives est basé sur l'optimisation d'un critère de partitionnement. Ce critère utilise, pour mesurer l'homogénéité d'une classe, non pas des distances euclidiennes comme dans le cas de la classification d'observations, mais des mesures d'association entre les variables. Plus précisément, le critère mesurant l'homogénéité d'une classe est un critère d'adéquation entre les variables qualitatives de la classe et une variable "centrale" quantitative, appelée variable latente. Cette adéquation est égale à la somme des rapports de corrélation entre les variables de la classe et la variable latente de la classe. Le problème consiste alors à définir la variable latente qui maximise ce critère. Nous avons vérifié que la première composante principale de l'ACM des variables de la classe est une solution de ce problème d'optimisation, et peut donc jouer le rôle de variable latente. Ce résultat une fois obtenu, nous avons développé plusieurs algorithmes : un algorithme de partitionnement de type nuées dynamiques (k-means, centres mobiles) et deux algorithmes de classification hiérarchique (ascendante et descendante).

2.1.1 Une procédure itérative pour la rotation en ACM (Analyse des Correspondances Multiples)

La méthode de rotation en ACM présentée dans cette section a donné lieu à la rédaction d'un article intitulé "Rotation in Multiple Correspondence Analysis : a planar rotation iterative procedure" écrit en collaboration avec Marie Chavent et Jérôme Saracco. Cet article a été soumis à la revue *Advances in Data Analysis and Classification* et est disponible dans la section 2.2.

Remarque : Mes premiers travaux ont concerné la rotation en ACP et ont été réalisés dans le cadre d'une collaboration portant sur des problématiques autour de la pollution atmosphérique. Ces travaux et les articles correspondant sont décrits dans le Chapitre 4.

2.1.1.1 Introduction

La rotation est un outil couramment utilisé en ACP. Comme énoncé précédemment, réaliser une ACP revient à approximer la matrice des données centrées réduites par le produit de la matrice des scores et de la matrice des loadings. Cette dernière joue un rôle majeur dans l'interprétation des résultats puisqu'elle contient les corrélations entre les variables et les scores. Appliquer la même matrice de rotation orthogonale à la matrice des loadings et à la matrice des scores garantit que la matrice des loadings après rotation contient toujours les corrélations entre les variables et les scores après rotation. Le principe de la recherche d'une structure simple consiste alors à choisir la "meilleure" matrice de rotation, c'est-à-dire celle permettant d'obtenir dans la matrice des loadings après rotation des corrélations au carré les plus proches possibles de 0 ou de 1. Pour cela, différents critères de rotation ont été proposés. Le plus courant est le critère Varimax (Kaiser, 1958) qui vise à maximiser la somme des variances des colonnes de la matrice des loadings. Dans le cas où deux composantes principales sont retenues, Kaiser (1958) donne l'écriture de la solution analytique de l'angle optimal pour effectuer la rotation planaire. Pour un nombre de composantes supérieur à deux, l'auteur propose d'utiliser cette solution dans un algorithme pratique consistant à appliquer des rotations successives de paires de facteurs.

Le problème de la rotation a été beaucoup moins étudié en ACM. Kiers (1991) s'est intéressé à ce problème dans le cadre de la méthode PCAMIX qu'il a développée pour l'analyse de données mixtes (qualitatives et quantitatives). Cette méthode inclut ainsi comme cas particuliers l'ACP et l'ACM. Il propose un critère de rotation qui dans le cas purement qualitatif est basé sur les rapports de corrélation entre les variables qualitatives et les composantes principales. Il utilise l'algorithme de De Leeuw et Pruzansky (1978) pour optimiser ce critère. Dans notre travail sur la rotation en ACM, nous avons utilisé ce même critère de rotation et nous avons défini, dans le cas particulier de deux dimensions (rotation planaire), l'expression analytique de l'angle optimal de rotation. Dans le cas de plus de deux dimensions, nous utilisons la procédure de rotations successives planaires, proposée par Kaiser (1958) pour la rotation en ACP.

2.1.1.2 Un bref panorama de l'ACM

L'ACM est le nom français (Benzécri, 1973 ; Lebart, Morineau et Warwick, 1984) pour une méthode de quantification de données qualitatives. Cette méthode a été proposée par plusieurs auteurs sous différents noms. On peut citer entres autres "Homogeneity Analysis" (Gifi, 1990), "Quantification Method" (Hayashi, 1954), "Dual Scaling" (Nishisato, 1980, 1994). Ces méthodes, reposant sur des fondements théoriques différents, conduisent en général à des solutions équivalentes (Tenenhaus et Young, 1985). Récemment, Greenacre et Blasius (2006) ont dressé un panorama de ces différentes "écoles" statistiques.

L'ACM peut être vue comme une ACP pondérée des profils lignes ou bien des profils colonnes du tableau disjonctif complet. Cette analyse duale nous permet de définir la relation barycentrique entre les coordonnées des composantes principales des observations et des modalités. Par la suite, le lien entre les notions de contribution et de

rapport de corrélation est rappelé. Enfin l'approximation de rang inférieur permettant la liberté de rotation des composantes est précisée.

ACP pondérée des profils lignes. Ici l'ACM est définie comme l'application d'une ACP pondérée (Benzécri, 1973 ; Greenacre, 1984 : Chapitre 3). Nous disposons de n observations faites sur p variables qualitatives. Notons q_j l'ensemble des modalités de la $j$ème variable et $\mathbf{G}$ le tableau disjonctif complet associé, de dimension $n \times q$ où $q = \sum_{j=1}^{p} q_j$. En traitant la matrice $\mathbf{G}$ comme une matrice des correspondances, nous divisons $\mathbf{G}$ par np (la somme de ses éléments) pour obtenir la matrice des fréquences : $\mathbf{F} = \frac{1}{np}\mathbf{G}$. Les sommes marginales des lignes et des colonnes définissent respectivement les vecteurs $\mathbf{r} \in \mathbb{R}^n$ et $\mathbf{c} \in \mathbb{R}^q$ des poids des lignes et des colonnes. Nous notons $\mathbf{D_r} = \text{diag}(\mathbf{r})$ et $\mathbf{D_c} = \text{diag}(\mathbf{c})$ les matrices diagonales de ces masses. Remarquons que le $i$ème élément de $\mathbf{r}$ est donné par $f_{i.} = \frac{1}{n}$ et le $s$ème élément de $\mathbf{c}$ par $f_{.s} = \frac{n_s}{np}$ où n_s est le nombre d'observations possédant la modalité s.

L'ACM est alors définie comme une ACP pondérée de la matrice des profils lignes centrés $\mathbf{D_r^{-1}}(\mathbf{F} - \mathbf{rc}^t)$ avec la métrique (du chi-deux) $\mathbf{D_c^{-1}}$ utilisée pour comparer deux profils. D'un point de vue géométrique, il s'agit de trouver $k \leq \text{rang}(\mathbf{F})$ composantes orthogonales telles que la variance des $\mathbf{D_c^{-1}}$- projections des n profils lignes centrés soit maximale. On introduit la matrice suivante :

$$\widetilde{\mathbf{F}} = \mathbf{D_r^{-1/2}}(\mathbf{F} - \mathbf{rc}^t)\mathbf{D_c^{-1/2}} \tag{2.1}$$

et on montre que la matrice $\mathbf{X}$ de dimension $n \times k$ des composantes principales des observations (lignes) est définie par $\mathbf{X} = \mathbf{D_r^{-1/2}}\widetilde{\mathbf{F}}\mathbf{V}_k$, où $\mathbf{V}_k$ est la matrice $q \times k$ des vecteurs propres correspondant aux k plus grandes valeurs propres $\lambda_1, ..., \lambda_k$ de la matrice $\widetilde{\mathbf{F}}^t\widetilde{\mathbf{F}}$. On appellera aussi $\mathbf{X}$ la matrice des coordonnées factorielles des observations.

ACP pondérée des profils colonnes. L'analyse duale des profils colonnes est obtenue en remplaçant les lignes par les colonnes, c'est-à-dire en transposant la matrice $\mathbf{F}$ et en répétant ce qui est énoncé ci-dessus, avec les pondérations et les métriques appropriées. On montre alors que la matrice des composantes principales des modalités (colonnes) est donnée par $\mathbf{Y} = \mathbf{D_c^{-1/2}}\widetilde{\mathbf{F}}^t\mathbf{U}_k$, où $\mathbf{U}_k$ est la matrice $n \times k$ des vecteurs propres associés aux k plus grandes valeurs propres de la matrice $\widetilde{\mathbf{F}}\widetilde{\mathbf{F}}^t$. On appellera également $\mathbf{Y}$ la matrice des coordonnées factorielles des modalités.

Décomposition en Valeurs Singulières. D'un point de vue pratique, pour obtenir les matrices des coordonnées principales des observations et des modalités, on utilise la Décomposition en Valeurs Singulières (DVS) suivante : $\widetilde{\mathbf{F}}^t = \mathbf{U}\Lambda\mathbf{V}^t$ où $\mathbf{U}^t\mathbf{U} = \mathbf{V}^t\mathbf{V} = \mathbb{I}_r$, avec $r = \text{rang}(\mathbf{F})$, et Λ est la matrice diagonale des valeurs singulières rangées par ordre décroissant. La matrice des coordonnées principales des observations (resp. modalités) est alors donnée par :

$$\mathbf{X} = \mathbf{D_r^{-1/2}}\mathbf{U}_k\Lambda_k \text{ (resp. } \mathbf{Y} = \mathbf{D_c^{-1/2}}\mathbf{V}_k\Lambda_k).$$

On peut alors en déduire la matrice des coordonnées factorielles standardisées des observations (resp. modalités) :

$$\mathbf{X}^* = \mathbf{X}\Lambda_k^{-1} = \mathbf{D_r}^{-1/2}\mathbf{U}_k \text{ (resp. } \mathbf{Y}^* = \mathbf{Y}\Lambda_k^{-1} = \mathbf{D_c}^{-1/2}\mathbf{V}_k). \tag{2.2}$$

Propriété barycentrique. En se souvenant à partir de la définition de $\mathbf{F}$ que $f_{is} = \frac{g_{is}}{np}$ (avec $g_{is} = 1$ si l'observation i possède la modalité s, et 0 sinon), $f_{i.} = \frac{1}{n}$ et $f_{.s} = \frac{n_s}{np}$, le terme général de $(\mathbf{F} - \mathbf{r}\mathbf{c}^t)$ est donné par $\frac{g_{is}}{np} - \frac{n_s}{n^2 p}$. Il s'en suit que le terme général $y_{s\alpha}$ de la matrice $\mathbf{Y}$ des coordonnées factorielles des modalités (coordonnée de la modalité s sur l'axe α) est égal à :

$$\begin{aligned} y_{s\alpha} &= \sum_{i=1}^{n} \frac{np}{n_s}\left(\frac{g_{is}}{np} - \frac{n_s}{n^2 p}\right)x^*_{i\alpha} \\ &= \frac{1}{n_s}\sum_{i=1}^{n} g_{is}x^*_{i\alpha} - \frac{1}{n}\sum_{i=1}^{n} x^*_{i\alpha} \\ &= \bar{x}^*_{s\alpha}. \end{aligned}$$

Ainsi le terme général $y_{s\alpha}$ de la matrice $\mathbf{Y}$ vaut $\bar{x}^*_{s\alpha}$ qui est la moyenne de la αème composante principale standardisée des observations. Ce résultat s'écrit matriciellement de la manière suivante $\mathbf{Y} = \mathbf{D_c}^{-1}(\mathbf{F} - \mathbf{r}\mathbf{c}^t)^t\mathbf{X}^*$, avec la matrice $\mathbf{X}^*$ définie en (2.2). Cette relation barycentrique permet une représentation simultanée des observations et des modalités dans le graphique appelé graphique asymétrique des colonnes (traduit de l'anglais "asymmetric map of the columns").

Lien entre contribution et rapport de corrélation. Tout d'abord, rappelons que la contribution absolue de la variable j à l'inertie de la αème composante principale des modalités (αème colonne de $\mathbf{Y}$) est donnée par $c_{j\alpha} = \sum_{s\in\mathcal{M}_j} f_{.s}y^2_{s\alpha}$, avec $\mathcal{M}_j$ l'ensemble des modalités de la variable j.

D'autre part, rappelons que le rapport de corrélation $\eta^2_{j\alpha}$ entre la variable j et la αème composante principale standardisée des observations $\mathbf{x}^*_\alpha$ (αème colonne de $\mathbf{X}^*$) est égal à la variance inter-groupe (pour les groupes définis par les modalités de la variable j) divisée par la variance de $\mathbf{x}^*_\alpha$. Comme $\bar{\mathbf{x}}^*_\alpha = 0$ et $\mathbb{V}(\mathbf{x}^*_\alpha) = 1$, on a :

$$\eta^2_{j\alpha} = \frac{\sum_{s\in\mathcal{M}_j}\frac{n_s}{n}(\bar{x}^*_{s\alpha} - 0)^2}{1},$$

et comme d'après la relation quasi-barycentrique $y_{s\alpha} = \bar{x}^*_{s\alpha}$, on a bien :

$$\eta^2_{j\alpha} = p \times c_{j\alpha}. \tag{2.3}$$

Ces valeurs de rapport de corrélation vont jouer un rôle similaire aux loadings dans la rotation en ACP.

Remarquons que le rapport de corrélation est également appelé mesure de discrimination par Gifi (1990). Cette mesure peut être vue comme la corrélation au carré entre une variable qualitative optimalement quantifiée et une composante principale (Gifi, 1990, p.96).

Approximation de rang inférieur. Eckart et Young (1936) ont montré qu'une approximation de rang k au sens des moindres carrés de la matrice $\widetilde{\mathbf{F}}$ est obtenue en sélectionnant les k plus grandes valeurs propres et les vecteurs propres associés de la DVS de la matrice. De plus, comme

$$||\widetilde{\mathbf{F}} - \mathbf{U}_k \Lambda_k \mathbf{V}_k^t||^2 = ||\mathbf{D_r}^{-1}(\mathbf{F} - \mathbf{r}\mathbf{c}^t)\mathbf{D_c}^{-1} - \mathbf{X}^*\mathbf{Y}^t||^2,$$

la matrice $\mathbf{X}^*\mathbf{Y}^t$ est une approximation de rang k au sens des moindres carrés de $\mathbf{D_r}^{-1}(\mathbf{F} - \mathbf{r}\mathbf{c}^t)\mathbf{D_c}^{-1}$. Cette approximation de rang inférieur est à l'origine de la rotation en ACM.

2.1.1.3 Recherche d'une structure simple en ACM par rotation

Objectifs. Notons $\widetilde{\mathbf{X}}^* = \mathbf{X}^*\mathbf{T}$, et $\widetilde{\mathbf{Y}} = \mathbf{Y}\mathbf{T}$, avec $\mathbf{T}$ une matrice orthonormale de rotation ($\mathbf{T}\mathbf{T}^t = \mathbf{T}^t\mathbf{T} = \mathbb{I}_k$). Comme $\mathbf{X}^*\mathbf{Y}^t = \widetilde{\mathbf{X}}^*\widetilde{\mathbf{Y}}^t$, l'approximation de rang inférieur de la matrice $\mathbf{D_r}^{-1}(\mathbf{F} - \mathbf{r}\mathbf{c}^t)\mathbf{D_c}^{-1}$ par le produit de deux matrices n'est pas unique. Il s'en suit que la matrice $\mathbf{X}^*$ des coordonnées factorielles standardisées des observations et la matrice $\mathbf{Y}$ des coordonnées factorielles des modalités sont déterminées à une rotation près. L'idée est d'utiliser cette propriété de non-unicité de la solution de l'ACM afin d'obtenir des composantes plus facilement interprétables. En notant $\tilde{\eta}^2_{j\alpha}$ le rapport de corrélation entre la variable j et la αème colonne de $\widetilde{\mathbf{X}}^*$, on peut montrer que :

$$\tilde{\eta}^2_{j\alpha} = p \sum_{s \in \mathcal{M}_j} f_{.s} \tilde{y}^2_{s\alpha},$$

où $\sum_{s \in \mathcal{M}_j} f_{.s} \tilde{y}^2_{s\alpha}$ est la contribution de la variable j à l'inertie de la αème colonne de $\widetilde{\mathbf{Y}}$, c'est-à-dire la αème composante principale des modalités après rotation.

On cherche alors la matrice $\mathbf{T}$ telle que les rapports de corrélation entre les variables et les composantes principales après rotation soient les plus proches possible de 0 ou 1. Ainsi des groupes de variables qualitatives vont apparaître, ayant des rapports de corrélation élevés sur la même composante, modérés sur quelques composantes et négligeables sur les autres.

Critère de rotation basé sur Varimax. Le critère de rotation que nous avons considéré est basé sur le critère Varimax utilisé en ACP. Pour plus de détails sur la rotation Varimax en ACP, le lecteur peut se référer à l'article intitulé "Analyse en Facteurs : présentation et comparaison des logiciels SAS, SPAD et SPSS" publié dans *Modulad* et disponible dans la section 4.2.

Il s'agit ici d'appliquer la fonction Varimax à la matrice $p \times k$ des rapports de corrélation (jouant le rôle des corrélations en ACP), sachant que la matrice de rotation $\mathbf{T}$ est appliquée à $\mathbf{Y}$ de dimension $q \times k$. Cela conduit à un critère un peu plus compliqué

que l'application direct de Varimax à la matrice $\widetilde{\mathbf{Y}} = \mathbf{YT}$:

$$\begin{aligned} h(\mathbf{T}) &= \sum_{\alpha=1}^{k} \left\{ \frac{\sum_{j=1}^{p} (\tilde{\eta}_{j\alpha}^2)^2}{p} - \left(\frac{\sum_{j=1}^{p} \tilde{\eta}_{j\alpha}^2}{p} \right)^2 \right\} \\ &= \sum_{\alpha=1}^{k} \left\{ p \sum_{j=1}^{p} \left(\sum_{s \in \mathcal{M}_j} f_{.s} \tilde{y}_{s\alpha}^2 \right)^2 - \left(\sum_{j=1}^{p} \sum_{s \in \mathcal{M}_j} f_{.s} \tilde{y}_{s\alpha}^2 \right)^2 \right\}. \end{aligned} \tag{2.4}$$

Trouver la matrice de rotation optimale consiste alors à résoudre le problème d'optimisation sous contraintes suivant :

$$\begin{array}{ll} \max\limits_{\mathbf{T}} & h(\mathbf{T}), \\ \text{s.c.} & \mathbf{T}\mathbf{T}^t = \mathbf{T}^t\mathbf{T} = \mathbb{I}_k. \end{array} \tag{2.5}$$

Procédure de rotation planaire itérative. Le critère de rotation proposé en ACP par Kaiser (1958) vise à maximiser la somme des variances des colonnes de la matrice des loadings après rotation (de dimension $p \times k$). Une solution analytique dans le cas $k = 2$ a été définie par Kaiser (1958). Pour une dimension supérieure à deux, Kaiser (1958) propose de réaliser des rotations planaires (en utilisant la solution analytique) successives de toutes les paires de facteurs. L'idée est d'effectuer la rotation des composantes 1 et 2, puis 1 et 3, ..., 1 et k, ..., $(k-1)$ et k de façon itérative jusqu'à convergence, c'est-à-dire jusqu'à obtenir $\frac{k(k-1)}{2}$ rotations successives fournissant un angle de rotation planaire nul. Il a été montré que lorsque chaque rotation planaire est globalement optimale dans le plan considéré, cette procédure améliore à chaque itération la fonction Varimax. La partie centrale de la définition de cette procédure réside donc dans l'écriture de la solution analytique pour la rotation dans un plan.

Nous avons utilisé la même approche pour résoudre le problème d'optimisation de la rotation en ACM. Pour utiliser cette même procédure, nous avons déterminé la solution analytique de l'angle optimal de rotation planaire, c'est-à-dire résolu le problème d'optimisation (2.5) pour $k = 2$.

Solution planaire explicite. En dimension $k = 2$, la matrice de rotation orthogonale $\mathbf{T}$ s'écrit simplement en fonction d'un angle de rotation θ :

$$\mathbf{T} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}. \tag{2.6}$$

Le problème d'optimisation (2.5) s'écrit alors comme un problème non contraint :

$$\left\{ \max_{\theta \in \mathbb{R}} h(\theta). \right. \tag{2.7}$$

Remarque. L'expression de $h(\theta)$ et les détails du calcul de la solution analytique sont disponibles dans la section 2.2 contenant le papier soumis sur ce travail. Afin de ne pas

surcharger de calculs mathématiques ce mémoire de thèse, seules les idées principales pour l'obtention de la solution sont données ici.

Pour trouver cette solution, nous exprimons h en fonction de θ et nous écrivons la dérivée de h par rapport à θ sous la forme suivante :

$$\frac{\partial h}{\partial \theta} = 2(a + b\cos 4\theta + c\sin 4\theta),$$

avec

$$a = (p-1)\sum_{j=1}^{p}\sum_{s\in\mathcal{M}_j}\sum_{t\in\mathcal{M}_j} f_{.s}f_{.t}\alpha_{st}\beta_{st} - \sum_{j=1}^{p}\sum_{l=1,l\neq p}^{p}\sum_{s\in\mathcal{M}_j}\sum_{t\in\mathcal{M}_l} f_{.s}f_{.t}\alpha_{st}\beta_{st},$$

$$b = (p-1)\sum_{j=1}^{p}\sum_{s\in\mathcal{M}_j}\sum_{t\in\mathcal{M}_j} f_{.s}f_{.t}\delta_{st}\gamma_{st} - \sum_{j=1}^{p}\sum_{l=1,l\neq p}^{p}\sum_{s\in\mathcal{M}_j}\sum_{t\in\mathcal{M}_l} f_{.s}f_{.t}\delta_{st}\gamma_{st},$$

$$c = (p-1)\sum_{j=1}^{p}\sum_{s\in\mathcal{M}_j}\sum_{t\in\mathcal{M}_j} f_{.s}f_{.t}\frac{1}{2}(\gamma_{st}^2 - \delta_{st}^2) - \sum_{j=1}^{p}\sum_{l=1,l\neq p}^{p}\sum_{s\in\mathcal{M}_j}\sum_{t\in\mathcal{M}_l} f_{.s}f_{.t}\frac{1}{2}(\gamma_{st}^2 - \delta_{st}^2),$$

et

$$\begin{aligned}
\alpha_{st} &= y_{s1}y_{t1} + y_{s2}y_{t2},\\
\beta_{st} &= y_{s2}y_{t1} - y_{s1}y_{t2},\\
\gamma_{st} &= y_{s2}y_{t1} + y_{s1}y_{t2},\\
\delta_{st} &= y_{s1}y_{t1} - y_{s2}y_{t2}.
\end{aligned}$$

Il suffit alors de résoudre l'équation $a + b\cos 4\theta + c\sin 4\theta = 0$. L'astuce consiste à diviser cette équation par $(b^2+c^2)^{1/2}$:

$$\frac{a}{(b^2+c^2)^{1/2}} + \frac{b}{(b^2+c^2)^{1/2}}\cos 4\theta + \frac{c}{(b^2+c^2)^{1/2}}\sin 4\theta = 0$$

et à introduire l'angle $\varphi \in]-\pi, +\pi]$ tel que $\cos(\varphi) = \frac{b}{(b^2+c^2)^{1/2}}$ et $\sin(\varphi) = \frac{c}{(b^2+c^2)^{1/2}}$. L'équation se réécrit alors :

$$\frac{a}{(b^2+c^2)^{1/2}} + \cos(\varphi)\cos 4\theta + \sin(\varphi)\sin 4\theta = 0,$$

soit

$$\frac{a}{(b^2+c^2)^{1/2}} + \cos(4\theta - \varphi) = 0.$$

Il existe donc deux solutions (correspondant au maximum et au minimum de la fonction) à condition que $|a| \leq (b^2+c^2)^{1/2}$, ce qui est nécessairement vérifié car la fonction à maximiser $h(\theta)$ ne dépend que de $\cos(\theta)$ et de $\sin(\theta)$, et elle est donc périodique (de période $\pi/2$) et dérivable. Par conséquent, la dérivée doit s'annuler en chaque maximum et minimum. Ces deux solutions sont alors :

$$\theta = \frac{1}{4}(\varphi \pm \arccos(-\frac{a}{\sqrt{b^2+c^2}})).$$

Finalement, l'angle optimal de rotation planaire correspond à la solution pour laquelle la valeur de la fonction h est la plus grande.

Une rapide étude sur simulations nous a permis de vérifier l'exactitude de la solution analytique et de visualiser l'impact de la rotation. Grâce à la rotation, les variables sont plus clairement reliées aux composantes principales. De plus, un graphique montre que les coordonnées des modalités après rotation sont mieux alignées sur les composantes principales.

2.1.1.4 Application sur des données réelles

La procédure de rotation planaire itérative proposée en ACM a été appliquée sur un jeu de données réelles obtenu dans le cadre d'une collaboration avec Voies Navigables de France (décrite dans le Chapitre 4). Afin d'illustrer les intérêts pratiques d'une telle approche en analyse des données, nous nous sommes limités à quatre variables. Les intérêts pratiques du choix d'un faible nombre de variables sont limités mais cela nous a permis d'illustrer le phénomène de rotation et ses avantages, en particulier au travers de graphiques. Deux groupes de variables sont apparus, facilitant ainsi l'interprétation des deux composantes principales. De plus la représentation des modalités sur le premier plan factoriel a mis en exergue un meilleur pouvoir discriminant des composantes après rotation. Ainsi il a été plus aisé d'interpréter et de donner un nom aux composantes et par la suite de mieux comprendre le comportement des navigants.

2.1.1.5 Conclusions et perspectives

Nous avons proposé une procédure de rotation planaire itérative pour la rotation en ACM. Cette algorithme pratique repose sur l'écriture de la solution analytique en dimension deux. Le critère de rotation est basé sur Varimax et sur la notion de rapport de corrélation. L'intérêt potentiel de la rotation en ACM a été illustré sur des données réelles. Concernant les perspectives de travail, il serait intéressant d'appliquer l'approche sur un jeu de données plus complexe. Dans le cas de variables quantitatives, ten Berge (1984) montre que la rotation Varimax peut être interprétée comme un cas particulier de diagonalisation de matrices symétriques et que la solution donnée par De Leeuw et Pruzansky (1978) est équivalente à celle proposée par Kaiser (1958). Ainsi il serait intéressant de faire le lien entre l'approche que j'ai décrite dans ma thèse et la procédure de rotation proposée par Kiers (1991). Enfin une autre piste de recherche serait la détermination de l'expression analytique de la matrice de rotation lorsque le nombre de composantes principales retenues à l'issue de l'ACM est supérieur à deux.

2.1.2 Classification de variables qualitatives autour de variables latentes

Les algorithmes proposés dans ce chapitre ont donné lieu à l'article intitulé "Clustering of categorical variables around latent variables" écrit en collaboration avec Marie Chavent et Jérôme Saracco. Cet article a été soumis à la revue *Computational Statistics and Data Analysis* et est disponible dans la section 2.3.

2.1.2.1 Introduction

Dans de nombreuses applications, on s'intéresse à la classification des variables et non pas à celle des individus. C'est le cas par exemple en analyse sensorielle (mise en place de groupes de descripteurs), en biochimie (classification de gènes), en marketing (segmentation d'un panel de consommateurs), en économie (détection de stratégies financières), etc. L'idée est alors de chercher des groupes de variables liées c'est-à-dire porteuses de la même information. Un autre objectif poursuivi par la classification de variables est la suppression des redondances entre les variables et ainsi la réduction de la dimension du tableau de données. Dans ce cas, il est nécessaire de sélectionner dans chaque classe une variable ou de résumer chaque classe de variables par une variable synthétique encore appelée variable latente.

Une approche assez simple et couramment utilisée pour la classification de variables consiste à calculer une matrice de dissimilarités entre les variables puis à appliquer une méthode de classification dédiée aux observations et capable de traiter une matrice de dissimilarités. Pour les variables quantitatives, de nombreuses mesures de dissimilarités peuvent être utilisées. Elles font intervenir par exemple le coefficient de corrélation, la mesure d'association de Soffritti (1999), la distance basée sur l'opérateur d'Escoufier, etc. Concernant les variables qualitatives, le nombre de mesures d'association disponibles est tout aussi important : χ^2, Rand, Belson, Jordan, etc. (voir par exemple Abdallah et Saporta, 1998 ou Derquenne, 1997).

Parallèlement des méthodes ont été développées spécifiquement pour la classification de variables. Pour des données quantitatives, on peut citer entre autres l'approche de Hastie et al. (2000) en biologie génomique ou encore le récent travail de Vichi et Saporta (2009) qui permet une classification simultanée des observations et des variables. La plus célèbre est sans doute la procédure VARCLUS du logiciel SAS. Cette procédure complexe avec peu de justifications des options offertes fournit une hiérarchie ou une partition des variables quantitatives. Une autre approche consiste à utiliser un algorithme de classification qui fournit simultanément des classes de variables et leurs prototypes, qui sont dans ce cas des variables latentes. Deux algorithmes de partitionnement de ce type existent déjà pour la classification de variables quantitatives : la méthode CLV (Clustering of variables around Latent Variables) proposée par Vigneau et Qannari (2003) et la méthode Diametrical Clustering développée par Dhillon et al. (2003). Ces deux méthodes utilisent dans leur version de base la même variable latente. Cette variable maximise l'homogénéité de la classe définie comme la somme des corrélations au carré des variables de la classe à cette variable latente. Elle est obtenue grâce à une ACP des variables de la classe.

A notre connaissance, il existe moins de méthodes qui ont été proposées spécifiquement pour la classification de variables qualitatives. On peut citer entre autres l'Analyse de la Vraisemblance du Lien de Lerman (1990, 1993) qui est une méthode de classification hiérarchique de variables quantitatives ou qualitatives. Dans notre travail, nous avons étendu la méthode CLV au cas de variables qualitatives. La variable latente d'une classe maximise l'homogénéité de la classe, définie cette fois comme la somme des rapports de corrélation entre les variables qualitatives de la classe et cette variable latente

quantitative. Nous avons montré que cette variable latente peut être obtenue par une ACM des variables de la classe. Nous avons également défini deux algorithmes de classification hiérarchique utilisant le même critère d'homogénéité : un algorithme ascendant et un descendant, ce dernier étant inspiré de la méthode VARCLUS de SAS. Ces différentes approches ont été appliquées sur les données réelles provenant de la collaboration avec Voies Navigables de France (VNF).

2.1.2.2 Un critère de partitionnement basé sur les rapports de corrélation

Soit un ensemble de n objets décrits sur un ensemble de p variables qualitatives. On note $\mathcal{V} = \{\mathbf{x}_1, \ldots, \mathbf{x}_p\}$ l'ensemble des p colonnes de la matrice des données. Pour plus de simplicité, on confondra la notion de variable et son vecteur de n réalisations. On parlera donc de la variable $\mathbf{x}_j$, avec $\mathbf{x}_j \in \mathcal{M}_j^n$, où $\mathcal{M}_j$ est l'ensemble des modalités de j.

Critère d'homogénéité d'une classe. On note $\mathcal{C} \subset \mathcal{V}$ une classe de variables qualitatives et $\mathbf{y}$ un vecteur de $\mathbb{R}^n$ appelé variable latente. Le critère d'homogénéité de $\mathcal{C}$ mesure l'adéquation entre les variables de la classe et la variable latente $\mathbf{y}$:

$$S(\mathcal{C}) = \sum_{\mathbf{x}_j \in \mathcal{C}} \eta^2(\mathbf{x}_j, \mathbf{y}), \tag{2.8}$$

où $\eta^2(\mathbf{x}_j, \mathbf{y})$ est le rapport de corrélation entre $\mathbf{x}_j$ et $\mathbf{y}$. Son expression est donnée par $\eta^2(\mathbf{x}_j, \mathbf{y}) = \dfrac{\sum_{s \in \mathcal{M}_j} n_s(\bar{y}_s - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$, avec $\bar{y}_s$ la moyenne de $\mathbf{y}$ calculée sur les observations possédant la modalité s. Cette mesure appartient à $[0, 1]$ et mesure le lien entre la variable qualitative $\mathbf{x}_j$ et la variable latente numérique $\mathbf{y}$.

Définition de la variable latente d'une classe. Dans la classe $\mathcal{C}$, la variable latente $\mathbf{y}$ maximise le critère d'homogénéité $S(\mathcal{C})$:

$$\mathbf{y} = \arg\max_{\mathbf{u} \in \mathbb{R}^n} \sum_{\mathbf{x}_j \in \mathcal{C}} \eta^2(\mathbf{x}_j, \mathbf{u}). \tag{2.9}$$

Proposition 1 *La variable latente* $\mathbf{y}$ *de* $\mathcal{C}$ *peut se définir des deux façons suivantes :*

(a) $\mathbf{y}$ *est le premier vecteur propre normé à 1 de la matrice* $\widetilde{\mathbf{F}}\widetilde{\mathbf{F}}^t$*, avec* $\widetilde{\mathbf{F}}$ *défini en* (2.1).

(b) $\mathbf{y}$ *est colinéaire à la première composante principale issue de l'ACM des variables de* $\mathcal{C}$.

La démonstration de cette proposition est fournie dans la section 2.3.2.

Critère d'homogénéité d'une partition de variables qualitatives. On note $\mathcal{P}_K = \{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$ une partition de $\mathcal{V}$ en K classes et $\mathcal{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_K\}$ un ensemble de K variables latentes. Le critère de partitionnement utilisé est additif et vise à maximiser l'homogénéité des classes de la partition :

$$H(\mathcal{P}_K) = \sum_{k=1}^{K} S(\mathcal{C}_k), \tag{2.10}$$

où $S(\mathcal{C}_k)$ est défini en (2.8) pour une classe générique.

2.1.2.3 Différents algorithmes de classification

Je présente dans cette section les trois algorithmes de classification que nous avons définis avec le critère d'homogénéité H.

Algorithme de type nuées dynamiques. Il s'agit d'un algorithme itératif qui alterne une étape de représentation et une étape d'affectation.

(a) *Initialisation* : Cette étape peut se faire de plusieurs façons.
- La première solution consiste à calculer les K premières composantes principales de l'ACM de toutes les variables de $\mathcal{V}$. Ces composantes jouent alors le rôle de variables latentes des K classes puis on passe à l'étape (c) pour créer une partition initiale.
- Une seconde solution consiste à effectuer une rotation des K premières composantes principales de l'ACM des variables de $\mathcal{V}$ (selon la procédure décrite dans la section 2.1.1), avant de passer à l'étape (c). Cette idée s'inspire de la procédure VARCLUS qui utilise la rotation en ACP pour démarrer avec une meilleure partition des variables. Dans notre cas, en appliquant une rotation, les valeurs des rapports de corrélation entre les variables et les composantes principales seront faibles ou élevées et l'affectation des variables aux classes sera plus facile et ainsi probablement meilleure.
- Enfin une troisième solution consiste à sélectionner au hasard K variables de $\mathcal{V}$ et à transformer chacune de ces variables qualitatives en une variable quantitative, jouant ainsi le rôle de variables latentes initiales. Cette quantification est réalisée en faisant une ACM de cette variable. Puis l'étape (c) permet de créer une partition initiale. En pratique, lorsqu'on choisit ce type d'initialisation aléatoire, l'algorithme complet (étapes (a)+(b)+(c)+(d)) est répété plusieurs fois avec différents tirages aléatoires initiaux et finalement la meilleure partition au sens de notre critère de partionnement H est retenue.

(b) *Représentation* : Pour tout $k = 1, ..., K$, on définit la variable latente $\mathbf{y}_k$ de $\mathcal{C}_k$ comme le premier vecteur propre normalisé de la matrice $\widetilde{\mathbf{F}}_k \widetilde{\mathbf{F}}_k^t$, où $\widetilde{\mathbf{F}}_k$ est défini en (2.1) pour une classe générique.

(c) *Affectation* : Chaque variable est affectée à la classe dont la variable latente est la plus proche au sens du rapport de corrélation : pour tout $j = 1, ..., p$, on

cherche ℓ tel que $\ell = \arg \max_{k=1,\ldots,K} \eta^2(\mathbf{x}_j, \mathbf{y}_k)$. Notons $\mathcal{C}_k$ la classe de $\mathbf{x}_j$. Alors si $\ell \neq k$, $\mathcal{C}_\ell \leftarrow \mathcal{C}_\ell \cup \{\mathbf{x}_j\}$ et $\mathcal{C}_k \leftarrow \mathcal{C}_k \backslash \{\mathbf{x}_j\}$.

(d) *Arrêt* : L'algorithme s'arrête lorsqu'aucune variable ne change de classe dans l'étape (c).

Proposition 2 *La variable latente d'une classe optimisant le critère S d'homogénéité d'une classe, on montre facilement que l'algorithme proposé converge vers un maximum local de H.*

Preuve. Nous montrons que le critère d'homogénéité H croît jusqu'à convergence. Pour cela, nous devons montrer que $H(\mathcal{P}^n, \mathcal{Y}^n) \leq H(\mathcal{P}^n, \mathcal{Y}^{n+1}) \leq H(\mathcal{P}^{n+1}, \mathcal{Y}^{n+1})$, où l'exposant n représente la $n$ème iteration de l'algorithme.

La première inégalité est vérifiée puisque la variable latente d'une classe $\mathcal{C}_k$ est définie telle que $S(\mathcal{C}_k^n, \mathbf{y}_k^n) \leq S(\mathcal{C}_k^n, \mathbf{y}_k^{n+1})$. Ensuite en sommant sur k, nous obtenons $H(\mathcal{P}^n, \mathcal{Y}^n) \leq H(\mathcal{P}^n, \mathcal{Y}^{n+1})$.

Finalement selon la définition de l'étape d'affectation, nous obtenons que $\sum_{k=1}^K \sum_{\mathbf{x}_j \in \mathcal{C}_k^n} \eta^2(\mathbf{x}_j, \mathbf{y}_k^{n+1}) \leq \sum_{k=1}^K \sum_{\mathbf{x}_j \in \mathcal{C}_k^{n+1}} \eta^2(\mathbf{x}_j, \mathbf{y}_k^{n+1})$, ce qui démontre la deuxième inégalité. □

Classification ascendante hiérarchique. Nous avons proposé une approche de classification hiérarchique ascendante basée sur l'optimisation du même critère H. D'après la Proposition 1 (a), ce critère peut s'écrire de la façon suivante :

$$H(\mathcal{P}_K) = \sum_{k=1}^{K} p_k \lambda_k,$$

où p_k est le nombre de variables dans $\mathcal{C}_k$ et λ_k est la plus grande valeur propre de la matrice $\widetilde{\mathbf{F}}_k \widetilde{\mathbf{F}}_k^t$, avec $\widetilde{\mathbf{F}}_k$ définie en (2.1) (cette définition est à adapter au cas de variables d'une classe $\mathcal{C}_k$).

Dans l'algorithme, on part de la partition des singletons puis on procède par agrégations successives de deux classes jusqu'à l'obtention d'une seule classe $\mathcal{V}$. A chaque étape, on agrège les deux classes $\mathcal{C}_l$ et $\mathcal{C}_m$ de la partition $\mathcal{P}_K = \{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$ qui permettent d'obtenir la partition $\mathcal{P}_{K-1}$ (en $K-1$ classes) la meilleure au sens de H. Le critère de partitionnement H étant additif, on a :

$$H(\mathcal{P}_{K-1}) = H(\mathcal{P}_K) - (S(\mathcal{C}_l) + S(\mathcal{C}_m) - S(\mathcal{C}_l \cup \mathcal{C}_m)), \tag{2.11}$$

et on utilise donc comme mesure d'agrégation :

$$D(\mathcal{C}_l, \mathcal{C}_m) = S(\mathcal{C}_l) + S(\mathcal{C}_m) - S(\mathcal{C}_l \cup \mathcal{C}_m) = \lambda_l + \lambda_m - \lambda_{l \cup m}.$$

Une classe $\mathcal{C} = \mathcal{C}_l \cup \mathcal{C}_m$ de la hiérarchie $\mathcal{H}$ de $\mathcal{V}$ obtenue avec cette mesure d'agrégation est alors indicée par $h(\mathcal{C}) = \lambda_l + \lambda_m - \lambda_{l \cup m}$.

On montre dans l'annexe de l'article donné dans la section 2.3 que cet indice est bien toujours positif. En revanche, on sait que pour ne pas avoir d'inversion dans le dendrogramme, il faut vérifier la monotonie de cet indice, c'est-à-dire vérifier que $\forall \mathcal{A}, \mathcal{B} \in \mathcal{H}$, si $\mathcal{A} \subset \mathcal{B}$, alors $h(\mathcal{A}) \leq h(\mathcal{B})$. Ce résultat n'a pas encore été démontré.

Classification descendante hiérarchique. Nous avons également développé une approche hiérarchique descendante. On part d'une partition en une seule classe $\mathcal{V}$ et on divise successivement une classe en deux sous-classes jusqu'à obtenir la partition des singletons. A chaque étape, une classe $\mathcal{C}_l$ de la partition $\mathcal{P}_K = \{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$ est divisée en deux classes de manière à obtenir la partition $\mathcal{P}_{K+1}$ (avec $K+1$ classes) la meilleure au sens de H. Plus précisément les deux procédures suivantes doivent être définies : la façon de diviser une classe en deux sous-classes et le choix de la classe à diviser.

- *Division d'une classe en deux sous-classes.* Afin de couper optimalement la classe $\mathcal{C}_l$ en deux sous-classes $(\mathcal{A}_l, \bar{\mathcal{A}}_l)$, l'énumération complète de l'ensemble des $2^{p_l-1}-1$ (avec p_l le nombre de variables de $\mathcal{C}_l$) bipartitions possibles n'est pas envisageable dès que p_l augmente. L'algorithme des nuées dynamiques décrit précédemment peut donc être utilisé pour obtenir une partition en deux classes localement optimale pour le critère H.

- *Choix de la classe à diviser.* Le critère H étant additif, on a $H(\mathcal{P}_{K+1}) = H(\mathcal{P}_K) + h(\mathcal{C}_l)$ où $\mathcal{C}_l = (\mathcal{A}_l \cup \bar{\mathcal{A}}_l)$ est la classe qui a été divisée pour obtenir $\mathcal{P}_{K+1}$. Donc choisir de diviser la classe qui donne la meilleure partition $\mathcal{P}_{K+1}$ au sens de H revient à choisir de diviser la classe $\mathcal{C}_l$ de $\mathcal{P}_K$ pour laquelle $h(\mathcal{C}_l)$ est maximum. Cela veut dire qu'en pratique il faut avoir trouvé une bipartition de toutes les classes de $\mathcal{P}_K$ avant de choisir celle qui sera effectivement divisée.

Dans l'approche descendante, l'ensemble des classes obtenues après $K-1$ divisions peut être vu comme une hiérarchie $\mathcal{H}_K$ dont les singletons sont les K classes de la partition $\mathcal{P}_K$ obtenues à la dernière étape de l'algorithme. Cette hiérarchie peut être considérée comme une hiérarchie partielle à mi-chemin entre les niveaux haut et bas. Elle est appelée hiérarchie haute dans Mirkin (2005) et est indexée par h. Le fait d'indicer la hiérarchie par le critère qui a permis le choix de la classe à diviser, ici h, garantit qu'une classe coupée avant les autres dans l'algorithme sera représentée plus haut dans le dendrogramme. De ce fait, lorsque les divisions sont stopées avant d'obtenir la partition en singletons, cela permet de s'assurer que la hiérarchie partielle ainsi obtenue $\mathcal{P}_K$ correspond à celle que l'on aurait obtenue sur les $K-1$ plus hauts niveaux du dendrogramme de la hiérarchie complète $\mathcal{H}_n$.

2.1.2.4 Etude d'un jeu de données réelles

Les différents algorithmes de classification ont été appliqués sur le jeu de données des navigants plaisanciers du Canal des Deux Mers (provenant de la collaboration avec VNF). L'objectif de la classification de variables sur cette étude de cas a été d'examiner la redondance éventuelle entre les variables afin de sélectionner un sous-ensemble d'attributs utilisables lors d'enquêtes futures. Ainsi cela permettra non seulement de baisser le coût d'édition des questionnaires mais aussi de réduire le temps moyen de réponse à l'ensemble du questionnaire. Par ailleurs, la suppression d'information redondante et la création des variables latentes peut être judicieuse et bénéfique lors d'analyses statistiques ultérieures.

Illustration de l'algorithme de classification ascendante hiérarchique. Tout d'abord, l'algorithme de classification ascendante hiérarchique a été appliqué sur un sous-ensemble des variables afin d'illustrer les bénéfices pratiques de cette approche. Le choix du nombre de classes a été fait en examinant le dendrogramme obtenu ainsi que l'évolution du critère d'agrégation, mais aussi en fonction de l'interprétation des partitions obtenues. La procédure bootstrap décrite dans la section 4.1 pour la classification hiérarchique descendante, que l'on peut également utiliser pour l'algorithme de classification hiérarchique ascendante ou de type nuées dynamiques, a confirmé ce choix du nombre de classes. Un calcul du rapport de corrélation moyen, c'est-à-dire la moyenne des rapports de corrélation entre les variables de chaque classe et la variable latente correspondante a également permis de s'assurer de l'homogénéité des classes obtenues. Enfin un sous-ensemble de questions a pu être sélectionné en choisissant dans chaque classe la variable dont le rapport de corrélation avec la variable latente est le plus fort.

Etude empirique et comparaison des différents algorithmes. Afin d'étudier et de comparer les algorithmes, nous avons calculé la proportion d'homogénéité expliquée par une partition $\mathcal{P}_K$ donnée. Cette mesure est égale au rapport entre le gain en homogénéité obtenu en passant de 1 à K classes et le gain maximal que l'on peut avoir avec la partition en singletons. Ainsi cette valeur vaut 0% pour la partition en une seule classe et 100% pour la partition en singletons. Comme cette valeur augmente en fonction du nombre de classes, elle ne peut être utilisée que pour comparer des partitions avec le même nombre de classes.

- *Impact de l'initialisation de l'algorithme des nuées dynamiques sur la qualité des partitions.* Les trois procédures d'initialisation décrites plus haut (composantes principales couplées ou non à une rotation, et initialisations multiples aléatoires) ont été comparées en fonction de la proportion d'homogénéité expliquée. Cette étude a montré que la rotation est bénéfique en terme d'homogénéité et que l'initialisation au hasard semble être une stratégie efficace.

- *Comparaison des différentes approches.* L'étude empirique menée sur le jeu de données réelles a mis en exergue de meilleurs résultats avec la classification hiérarchique ascendante en terme d'homogénéité que son homologue descendant. Une explication possible à la faiblesse de l'approche descendante est que la division d'une classe en deux sous-classes n'est pas globalement optimale puisqu'elle est obtenue avec l'algorithme de type nuées dynamiques. Au contraire, dans la version ascendante, l'agrégation de deux classes est la meilleure possible au sens du critère de partitionnement.

2.1.2.5 Conclusions et perspectives

Nous avons étendu un critère existant pour la classification de variables quantitatives au cas de données qualitatives. Le critère de partitionnement est basé sur la notion de rapport de corrélation et l'ACM est utilisée pour construire la variable latente d'une classe. Un algorithme de partitionnement et deux algorithmes hiérarchiques ont été proposés pour le même critère d'homogénéité et comparés empiriquement sur des données réelles.

Concernant des pistes de travail futures, il reste tout d'abord à démontrer la monotonie de la fonction h utilisée pour indicer les hiérarchies. Notons qu'en pratique, sur les jeux de données réelles ou simulées que nous avons considérées, nous n'avons jamais observé de phénomènes d'inversion. Un autre point important en pratique serait de calculer la complexité des algorithmes proposés.

Un travail qui n'a pas encore été réalisé est la comparaison des méthodes de classification proposées avec les méthodes existantes de classification de variables qualitatives, par exemple l'Analyse de la Vraisemblance du Lien de Lerman (1990, 1993) ou encore les méthodes classiques sur tableaux de dissimilarités. Une étude par simulations est envisagée.

Une approche courante pour la classification d'observations issues de variables qualitatives consiste à réaliser d'abord une ACM puis à effectuer une classification sur les coordonnées factorielles issues de cette ACM (en utilisant un nombre de composantes restreint). Cependant certains auteurs (voir par exemple DeSarbo et al., 1990 ; De Soete et Caroll, 1994 ; Vichi et Kiers, 2001) soulignent les effets néfastes de cette procédure qu'ils nomment "tandem analysis". Selon eux, l'ACM identifie parfois des composantes qui contribuent peu à la détection d'une structure dans les observations et qui au contraire masquent l'information taxinomique. En effet, l'objectif de l'ACM est d'identifier une première composante principale qui explique le plus de variation possible dans le nuage de points puis une seconde qui lui est orthogonale et qui explique un grand pourcentage d'inertie et ainsi de suite. Ainsi on peut concevoir que des informations relatives à la structure des observations puissent être masquées par la création de ces composantes qui visent seulement à reconstruire au mieux la variance initiale. Au contraire, la classification de variables supprime l'information redondante et la création des variables latentes se fait au vue de la réorganisation des variables en classes homogènes. Nous souhaiterions donc comparer sur un exemple réel la "qualité " des partitions des observations obtenues d'une part avec les coordonnées factorielles de l'ACM, et d'autre part avec les variables latentes de l'une des méthodes de classification de variables que nous avons proposées. Pour cela, nous pensons utiliser de nouveau les données issues de la collaboration avec VNF. En effet, nous disposons pour ces données d'une classification "experte" des navigants plaisanciers et il serait intéressant de voir quelle approche retrouve au mieux cette structure en classes.

Enfin, lors du traitement statistique des enquêtes des navigants plaisanciers, nous avons été confrontés au problème des données manquantes et nous avons dû supprimer les individus avec des réponses manquantes afin d'appliquer les algorithmes de classification de variables qualitatives. Un travail sur l'utilisation de l'algorithme NIPALS (entre autres) pour la gestion des données manquantes en ACM est en cours et des premiers programmes et résultats ont été obtenus.

2.2 Rotation in Multiple Correspondence Analysis : a planar rotation iterative procedure

Abstract. Multiple Correspondence Analysis (MCA) is a well-known multivariate method for statistical description of categorical data. Similarly to what is done in Principal Component Analysis (PCA) and Factor Analysis, the MCA solution can be rotated to increase the components simplicity. The idea behind a rotation is to find subsets of variables which coincide more clearly with the rotated components. This implies that maximizing components simplicity can help in factor interpretation and in variables clustering. In this paper, we propose a two-dimensional analytic solution for rotation in MCA. Similarly to what is done by Kaiser (1958) for PCA, this planar solution is computed in a practical algorithm applying successive pairwise planar rotations for optimizing the rotation criterion. This criterion is a varimax-based one relying on the correlation ratio between the categorical variables and the MCA components. A simulation study is used to illustrate the proposed solution. An application on a real data set shows the possible benefits of using rotation in MCA.

Keywords : categorical data, multiple correspondence analysis, correlation ratio, rotation.

2.2.1 Introduction

Multiple Correspondence Analysis (MCA) is the french name (Benzécri, 1973 ; Lebart, Morineau and Warwick, 1984) for a multivariate quantification method of categorical data. This method has been proposed by many different authors under various names. Among others we can mention the Dutch Homogeneity Analysis (Gifi, 1990), the Japanese Quantification Method (Hayashi, 1954), the Canadian Dual Scaling (Nishisato, 1980, 1994). All these methods with different theoretical foundations lead usually to equivalent solutions (Tenenhaus and Young, 1985). A recent survey of various approaches from different statistical "schools" can be found in Greenacre and Blasius (2006).

In the present paper, our treatment and interpretation of MCA resemble that of PCA (Benzécri, 1973 ; Greenacre, 1984 : chapter 3). Indeed, MCA is concerned with observations of p categorical variables for each n samples and may be viewed as a form of PCA applicable to categorical variables rather than quantitative variables. However, special emphasis will be placed on the fact that, as in Correspondence Analysis (CA) (Greenacre, 1984 : chapter 2 and appendix), MCA solutions are neatly encapsulated in the Singular Value Decomposition (SVD) of a suitably transformed matrix. More precisely, the relationship between MCA and the lower rank approximation approach of biplot (Greenacre, 1993 or Gower and Hand, 1996) provides the mathematical scaffolding for applying rotation methods in MCA.

In Principal Component Analysis (PCA) and Factor Analysis, objective criteria have been proposed for the attainment of simple structure. The varimax criterion introduced by Kaiser (1958) is by far the most commonly used criterion for rotation in PCA. This

criterion aims at maximizing the sum over the columns of the squared elements of the loading matrix. The loading matrix plays indeed a major part in the interpretation of the results since it contains the correlations between the variables and the principal components. The idea is to get components for which the interpretation is easier, that is to rotate the loading matrix and the standardized principal components such that groups of variables appear, having high loadings on the same component, moderate on a few components and negligible on the remaining ones. Because the lower-rank approach in PCA gives the freedom for orthogonal rotation, the only consequence is that the percentage of variance explained is redistributed along newly rotated axes, while still conserving the variance explained by the solution as a whole. In practice, defining the best orthogonal rotation matrix sums up to a constrained optimization problem. When a solution requires only two dimensions the rotation occurs in a plane and the rotation matrix can be written according to a rotation angle θ leading to an unconstrained real optimization problem. When the interpretation of three or more dimensions is required, the analytic expression of θ optimizing the Varimax criterion is used by Kaiser (1958) in a practical algorithm applying successive pairwise planar rotations. Several other algorithms for the maximization of the Varimax criterion have since been proposed in literature : see for instance Neudecker, (1981) ; Sherin (1966) or ten Berge (1984).

As has already been pointed, MCA and thus CA too, is a particular case of weighted PCA. Despite this close relationship with a method in which rotation is quite common, rotation in CA has not received much attention : Van de Velden and Kiers (2003, 2005) and Greenacre (2006) explicitly considered rotation in CA. Their results, however, do not carry out over to rotation in MCA. Adachi (2004) considered oblique rotation in MCA. Oblique and orthogonal rotation involves the same problem of maximizing a simplicity criterion. Only the imposed constraints differ. Since fewer constraints are imposed in oblique rotation, it is generally possible to obtain simpler solution than in orthogonal rotation (Browne, 2001). Despite this advantage, orthogonal rotations are commonly used in practice. Indeed, the orthogonality leads to direct interpretation of the rotated axes : the orthogonally rotated loadings can be directly interpreted as correlations between the variables and the rotated standardized principal components and graphical representations remain possible. Kiers (1991) considered orthogonal rotation in PCAMIX. This method, developed for the analysis of a mixture of categorical and numerical variables, includes PCA and MCA as special cases. The several rotation techniques proposed for simple structure in PCAMIX solution can then be applied to MCA solutions. In PCA, the rotation criteria are defined on the correlations between variables and principal components. For qualitative variables, however, the correlation can not be used and another coefficient has to be chosen to express the link between a categorical variable and a (quantitative) component. Kiers (1991) used for rotation in PCAMIX, and then in MCA, the discrimination measure (Gifi, 1990) defined as the contribution of a component to the inertia of a variable that is accounted for. This measure can be interpreted as the squared correlation between a variable optimally quantified and a principal component (Gifi, 1990, p.96), or alternatively, as the well-known correlation ratio. The idea of simple structure in MCA is to rotate the component coordinates such that groups of categorical variables appear, having high correlation ratio on the same

component, moderate on a few components and negligible on the remaining ones. The research of simple structure in MCA can then be operated by applying orthogonal rotation criteria to the correlation ratio matrix. Kiers (1991) gave in this framework a matrix formulation of the orthomax criterion (including varimax) which permits interpreting this rotation problem as a simultaneous diagonalization of a set of symmetric matrices (ten Berge, 1984), and proposed to use the algorithm of de Leeuw and Pruzansky (1978) for that simultaneous diagonalization.

The main contribution of this paper is the definition of the analytic expression of the angle θ for orthogonal planar rotation in MCA, optimizing the correlation ratio based Varimax criterion. The relevance of finding an analytic solution for two dimensional MCA is first that this solution can be used in divisive clustering of categorical data, which was our first motivation for this work. Moreover, this planar solution can be used, as the planar solution proposed by Kaiser (1958), in a practical algorithm applying successive pairwise planar rotations for optimizing the rotation criterion in more than two dimensions. This procedure is an alternative to that proposed by Kiers (1991). We also try to give in this paper a pedagogic and relatively detailed presentation of the problem of rotation in MCA, which has not been extensively studied yet. Therefore, we remind the relations between the french geometric presentation of PCA and MCA, and the matrix lower-rank approximation approach of biplots.

In Section 2.2.2 we recall the principles of MCA. In Section 2.2.3 we consider the rotation problem to obtain simple structure in MCA and we give the expression of the analytic solution for two-dimensional rotation. A simulated example is used to illustrate planar rotation. A real data application is treated in Section 2.2.4 to show the potential benefits of using rotation in MCA. Finally concluding remarks are given in Section 2.2.5.

2.2.2 Recall on multiple correspondence analysis

In this section the theory of MCA is summarized in order to define the terms and notation for the later sections. The basic data we start with are n observations on p categorical variables. Suppose variable j can assume q_j different values. We can code the data using indicator matrices (also known as dummies). Indicator matrix $\mathbf{G}_j$ is $n \times q_j$. It consists of zeroes and ones, and it has exactly one element equal to one in each row, indicating in which category of variable j object i belongs. By concatenating the $\mathbf{G}_j$ we obtain the $n \times q$ matrix $\mathbf{G}$, with q the sum of the q_j.

MCA is defined in this paper as the application of simple CA to the indicator matrix $\mathbf{G}$. Hence CA, and then MCA too, are defined as the application of weighted PCA to the indicator matrix $\mathbf{G}$ (Benzécri, 1973 ; Greenacre, 1984 : chapter 3). More precisely, $\mathbf{G}$ is divided by its grand total np to obtain the so-called "correspondence matrix" $\mathbf{F} = \frac{1}{np}\mathbf{G}$, so that $\mathbf{1}_n^t \mathbf{F} \mathbf{1}_q = 1$, where, generically, $\mathbf{1}_i$ is an $i \times 1$ vector of ones. Furthermore, the row and column marginals define respectively the vectors $\mathbf{r} = \mathbf{F}\mathbf{1}_q$ and $\mathbf{c} = \mathbf{F}^t\mathbf{1}_n$, that is the vectors of row and column masses. Let $\mathbf{D_r} = \text{diag}(\mathbf{r})$ and $\mathbf{D_c} = \text{diag}(\mathbf{c})$ be the diagonal matrices of these masses. In this particular case, the ith element of $\mathbf{r}$ is $f_{i.} = \frac{1}{n}$ and the sth element of $\mathbf{c}$ is $f_{.s} = \frac{n_s}{np}$ where n_s is the frequency of category s.

Weighted PCA of the row profiles. The objects are described here by the row profiles which are points in $\mathbb{R}^q$ calculated by dividing the rows of $\mathbf{F}$ by their row marginals. They are weighted by the row masses in $\mathbf{r}$ and their centroid (weighted average) turns out to be exactly the vector of marginal column totals $\mathbf{c}^t$. MCA is then defined as the application of PCA to the centered matrix $\mathbf{D_r^{-1}}(\mathbf{F}-\mathbf{rc}^t)$ with distances between profiles measured by the chi-squared metric defined by $\mathbf{D_c^{-1}}$. From a geometrical point of view, this weighted PCA searches for $k \leq \text{rank}(\mathbf{F})$ orthogonal principal axes such that for each principal axe, the variance of the $\mathbf{D_c^{-1}}$-projections of the n profiles is maximal. The coordinates of the n projected row profiles on these principal axes are called *row principal coordinates*. Note that row (resp. coordinates) is sometimes replaced by object (resp. scores). The $n \times k$ matrix $\mathbf{X}$ of row principal coordinates is defined by :

$$\mathbf{X} = \mathbf{D_r^{-1/2}}\widetilde{\mathbf{F}}\mathbf{V}_k, \tag{2.12}$$

where $\widetilde{\mathbf{F}} = \mathbf{D_r^{-1/2}}(\mathbf{F}-\mathbf{rc}^t)\mathbf{D_c^{-1/2}}$ and $\mathbf{V}_k$ is the $q \times k$ matrix of eigenvectors corresponding to the k largest eigenvalues $\lambda_1, ..., \lambda_k$ of the matrix $\widetilde{\mathbf{F}}^t\widetilde{\mathbf{F}}$ (see Appendix 1 for a short recall on this wellknown result). Similarly to what is done in PCA, these projected row profiles can be plotted, for visualization and interpretation, in the different planes defined by these principal axes called *row principal planes*.

Weighted PCA of the column profiles. The categories are described here by the column profiles which are points in $\mathbb{R}^n$ calculated by dividing the columns of $\mathbf{F}$ by their column marginals. The dual analysis of columns profiles can be defined simply by interchanging rows with columns and all associated entities, i.e. transposing the matrix $\mathbf{F}$ and repeating all the above. The metrics used to define the principal axes in the weighted PCA of the centered profiles matrix $\mathbf{D_c^{-1/2}}(\mathbf{F}-\mathbf{rc}^t)^t$ are $\mathbf{D_c}$ on $\mathbb{R}^q$ and $\mathbf{D_r^{-1}}$ on $\mathbb{R}^n$. The coordinates of the q projected column profiles on these principal axes are called *column principal coordinates*. Note that column (resp. coordinates) is sometimes replaced by category (resp. scores). The $q \times k$ matrix $\mathbf{Y}$ of columns principal coordinates is defined by :

$$\mathbf{Y} = \mathbf{D_c^{-1/2}}\widetilde{\mathbf{F}}^t\mathbf{U}_k, \tag{2.13}$$

where $\mathbf{U}_k$ is the $n \times k$ matrix of eigenvectors corresponding to the k largest eigenvalues $\lambda_1, ..., \lambda_k$ of the matrix $\widetilde{\mathbf{F}}\widetilde{\mathbf{F}}^t$. These projected column profiles can be plotted, for visualization and interpretation, in the planes defined by these principal axes called *column principal planes*.

Use of SVD. The computational algorithm to obtain the principal coordinates of the row and column profiles with respect to principal axes is obtained with SVD :

$$\widetilde{\mathbf{F}} = \mathbf{U}\Lambda\mathbf{V}^t \tag{2.14}$$

where $\mathbf{U}^t\mathbf{U} = \mathbf{V}^t\mathbf{V} = \mathbb{I}_r$, Λ is the diagonal matrix with singular values on the diagonal, in weakly descending order, and r is the rank of $\widetilde{\mathbf{F}}$. It follows indeed from (2.14) that

expression (2.12) (resp. (2.13)) of the row (resp. column) principal coordinates matrix can be rewritten :

$$\mathbf{X} = \mathbf{D_r}^{-1/2}\mathbf{U}_k\Lambda_k \text{ (resp. } \mathbf{Y} = \mathbf{D_c}^{-1/2}\mathbf{V}_k\Lambda_k). \tag{2.15}$$

The barycentric property. From (2.14) and (2.15), we obtain :

$$\mathbf{Y} = \mathbf{D_c}^{-1}(\mathbf{F} - \mathbf{r}\mathbf{c}^t)^t\mathbf{X}^* \tag{2.16}$$

where $\mathbf{X}^* = \mathbf{D_r}^{-1/2}\mathbf{U}_k = \mathbf{X}\Lambda_k^{-1}$ is the $n \times k$ matrix of the standardized row coordinates called *row standard coordinates*. Equation (2.16) can be interpreted in terms of *reciprocal averaging* and is called *barycentric property* : the principal coordinate of a category is the average of the standard coordinates of the objects in that category. The corresponding formula is $y_{s\alpha} = \frac{1}{n_s}\sum_{i=1}^{n} g_{is}x^*_{i\alpha} = \bar{x}^*_{s\alpha}$, where $y_{s\alpha}$ is the (s,α)-element of $\mathbf{Y}$ and g_{is} is the (i,s)-element of $\mathbf{G}$ (see Appendix 2 for details on the barycentric property). This barycentric property permits a simultaneous representation of the objects and the categories in the so called *asymmetric map of the columns*.

Contribution and correlation ratio. The absolute contribution of the variable j to the inertia of the column principal component α (αth column of $\mathbf{Y}$) is $c_{j\alpha} = \sum_{s\in\mathcal{M}_j} f_{.s}y^2_{s\alpha}$, where $\mathcal{M}_j$ is the set of categories of variable j. Remembering moreover that $y_{s\alpha} = \bar{x}^*_{s\alpha}$ and the sample mean (resp. variance) of the αth column of $\mathbf{X}^*$ is equal to zero (resp. one), we have the following relation between the absolute contribution $c_{j\alpha}$ and the correlation ratio between the variable j and the row standard component α (αth column of $\mathbf{X}^*$) :

$$\eta^2_{j\alpha} = \frac{\sum_{s\in\mathcal{M}_j}\frac{n_s}{n}(\bar{x}^*_{s\alpha} - 0)^2}{1} = p \times c_{j\alpha}. \tag{2.17}$$

Remembering that in PCA the loadings are correlations between the variables and the components, the correlation ratios, called discrimination measure in Gifi (1990), are interpreted in MCA as *squared loadings*.

The lower rank approximation approach. As shown by Eckart and Young (1936), a rank k least squares approximation of $\widetilde{\mathbf{F}}$ is obtained by selecting in the k largest singular values and corresponding singular vectors. Now, as

$$||\widetilde{\mathbf{F}} - \mathbf{U}_k\Lambda_k\mathbf{V}_k^t||^2 = ||\mathbf{D_r}^{-1}(\mathbf{F} - \mathbf{r}\mathbf{c}^t)\mathbf{D_c}^{-1} - \mathbf{X}^*\mathbf{Y}^t||^2,$$

the matrix $\mathbf{X}^*\mathbf{Y}$ is a rank k least squares approximation of $\mathbf{D_r}^{-1}(\mathbf{F} - \mathbf{r}\mathbf{c}^t)\mathbf{D_c}^{-1}$. This lower rank approximation gives the freedom for rotation in MCA.

2.2.3 Simple structure in MCA.

Let $\widetilde{\mathbf{X}}^* = \mathbf{X}^*\mathbf{T}$, and $\widetilde{\mathbf{Y}} = \mathbf{Y}\mathbf{T}$, where $\mathbf{T}\mathbf{T}^t = \mathbf{T}^t\mathbf{T} = \mathbb{I}_k$. Then, as $\mathbf{X}^*\mathbf{Y}^t = \widetilde{\mathbf{X}}^*\widetilde{\mathbf{Y}}^t$, we immediately see that the lower rank approximation is not unique and that the MCA solution $\mathbf{X}^*$ and $\mathbf{Y}$ is not unique over orthogonal rotations. This non-uniqueness can be exploited to improve the interpretability of the original solution by means of rotation. Clearly, rotation of the column principal coordinates matrix $\mathbf{Y}$ to simple structure must be followed by the same rotation of the row standard coordinates matrix $\mathbf{X}^*$. To simplify the interpretation of the correlation ratios, the matrices $\mathbf{Y}$ and $\mathbf{X}^*$ are rotated in such a way that when considering one variable few correlation ratios are large (close to 1) and as many as possible are close to zero.

The Varimax-based function. After rotation of $\mathbf{X}^*$ and $\mathbf{Y}$, the relation (2.17) remains true :

$$\tilde{\eta}^2_{j\alpha} = p \sum_{s \in \mathcal{M}_j} f_{.s} \tilde{y}^2_{s\alpha}, \tag{2.18}$$

where $\tilde{\eta}^2_{j\alpha}$ is the correlation ratio between the variable j and αth column of $\widetilde{\mathbf{X}}^*$. The Kaiser's Varimax function is applied to the $p \times k$ correlation ratio matrix, interpreted as squared correlations, but the rotation matrix $\mathbf{T}$ is applied to $\mathbf{Y}$ which leads to a more complicated function than in PCA :

$$\begin{aligned} h(\mathbf{T}) &= \sum_{\alpha=1}^{k} \left\{ \frac{\sum_{j=1}^{p} (\tilde{\eta}^2_{j\alpha})^2}{p} - \left(\frac{\sum_{j=1}^{p} \tilde{\eta}^2_{j\alpha}}{p} \right)^2 \right\} \\ &= \sum_{\alpha=1}^{k} \left\{ p \sum_{j=1}^{p} \left(\sum_{s \in \mathcal{M}_j} f_{.s} \tilde{y}^2_{s\alpha} \right)^2 - \left(\sum_{j=1}^{p} \sum_{s \in \mathcal{M}_j} f_{.s} \tilde{y}^2_{s\alpha} \right)^2 \right\}. \end{aligned} \tag{2.19}$$

The rotation of the $p \times k$ matrix $\mathbf{Y}$ can be formulated as objective,

$$\begin{array}{ll} \max\limits_{\mathbf{T}} & h(\mathbf{T}), \\ \text{s.t.} & \mathbf{T}\mathbf{T}^t = \mathbf{T}^t\mathbf{T} = \mathbb{I}_k. \end{array} \tag{2.20}$$

The rotation iterative procedure. In PCA, the Kaiser's procedure is aimed at maximizing the sum of variances of the squared columns of $\widetilde{\mathbf{A}}$, where $\widetilde{\mathbf{A}} = \mathbf{A}\mathbf{T}$ for a given $p \times k$ matrix $\mathbf{A}$ of factor loadings. Because a direct solution for the optimal $\mathbf{T}$ is not available, except for the case $k = 2$, Kaiser suggested an iterative procedure based on planar rotations. The idea is to alternately rotate all pairs of columns of $\mathbf{A}$. Each rotation is globally optimal for the plane under consideration, and improves the Varimax function, because the contribution of all $k - 2$ columns except the pair being rotated is not affected. The essential part of Kaiser's procedure is then the explicit formula of the Varimax angle of rotation.

In MCA, we propose to use the same iterative procedure for the optimization problem (2.20) : the single-plane rotations are made on dimension 1 with 2, 1 with 3, ..., 1

with $k, \ldots, (k-1)$ with k iteratively until the process converges, i.e. until $\frac{k(k-1)}{2}$ successive rotations providing an angle of rotation equal to zero are obtained. The definition of an explicit formula for the angle of rotation θ maximizing the rotation function h is then the essential part of our proposed generalization of the Kaiser's procedure to MCA.

The planar explicit solution. For $k = 2$, the rotation matrix $\mathbf{T}$ is defined by

$$\mathbf{T} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}. \tag{2.21}$$

The optimization problem (2.20) can then be rewritten :

$$\max_{\theta \in \mathbb{R}} h(\theta), \tag{2.22}$$

where the analytic expression of $h(\theta)$ is given in (2.30) in Appendix 3. The derivative of h gives (see Appendix 3 for details) :

$$\frac{\partial h}{\partial \theta} = 2(a + b\cos(4\theta) + c\sin(4\theta)), \tag{2.23}$$

where

$$\begin{aligned}
a &= (p-1)\sum_{j=1}^{p}\sum_{s\in\mathcal{M}_j}\sum_{t\in\mathcal{M}_j} f_{.s}f_{.t}\alpha_{st}\beta_{st} - \sum_{j=1}^{p}\sum_{l\neq j}\sum_{s\in\mathcal{M}_j}\sum_{t\in\mathcal{M}_l} f_{.s}f_{.t}\alpha_{st}\beta_{st}, \\
b &= (p-1)\sum_{j=1}^{p}\sum_{s\in\mathcal{M}_j}\sum_{t\in\mathcal{M}_j} f_{.s}f_{.t}\delta_{st}\gamma_{st} - \sum_{j=1}^{p}\sum_{l\neq j}\sum_{s\in\mathcal{M}_j}\sum_{t\in\mathcal{M}_l} f_{.s}f_{.t}\delta_{st}\gamma_{st}, \\
c &= \frac{(p-1)}{2}\sum_{j=1}^{p}\sum_{s\in\mathcal{M}_j}\sum_{t\in\mathcal{M}_j} f_{.s}f_{.t}(\gamma_{st}^2 - \delta_{st}^2) - \frac{1}{2}\sum_{j=1}^{p}\sum_{l\neq j}\sum_{s\in\mathcal{M}_j}\sum_{t\in\mathcal{M}_l} f_{.s}f_{.t}(\gamma_{st}^2 - \delta_{st}^2),
\end{aligned} \tag{2.24}$$

with

$$\begin{aligned}
\alpha_{st} &= y_{s1}y_{t1} + y_{s2}y_{t2}, \\
\beta_{st} &= y_{s2}y_{t1} - y_{s1}y_{t2}, \\
\gamma_{st} &= y_{s2}y_{t1} + y_{s1}y_{t2}, \\
\delta_{st} &= y_{s1}y_{t1} - y_{s2}y_{t2}.
\end{aligned} \tag{2.25}$$

Afterwards the trick to solve $a + b\cos(4\theta) + c\sin(4\theta) = 0$ consists in dividing each term by $(b^2+c^2)^{1/2}$ and introducing the angle $\varphi \in]-\pi, +\pi]$ such that $\cos(\varphi) = \frac{b}{(b^2+c^2)^{1/2}}$ and $\sin(\varphi) = \frac{c}{(b^2+c^2)^{1/2}}$. It gives

$$\frac{a}{(b^2+c^2)^{1/2}} + \cos(\varphi)\cos(4\theta) + \sin(\varphi)\sin(4\theta) = \frac{a}{(b^2+c^2)^{1/2}} + \cos(4\theta - \varphi) = 0.$$

As h only depends on $\cos(\theta)$ and $\sin(\theta)$, it is periodic (of period $\pi/2$) and differentiable and the derivative necessarily cancels for each minimum and maximum. Therefore $|a| \leq (b^2+c^2)^{1/2}$ and this equation has two solutions :

$$\hat{\theta} = \frac{1}{4}(\pm \arccos(-\frac{a}{(b^2+c^2)^1/2}) + \varphi), \tag{2.26}$$

corresponding to the minimum and the maximum of h, on condition of course that $|a| \leq (b^2+c^2)^{1/2}$. But this condition is necessarily verified because as h only depends on $\cos(\theta)$ and $\sin(\theta)$, it is periodic (of period $\pi/2$) and differentiable and the derivative necessarily cancel for each minimum and maximum.

An illustrative example. In this simulated example, we consider four binary variables $x_1, \ldots, x_4$ such that x_1 and x_2 (respectively x_3 and x_4) are strongly linked and not related to the other variables x_3 and x_4 (resp. x_1 and x_2). Then we have two groups of variables denoted $\mathcal{C}_1$ and $\mathcal{C}_2$. Let e_1 (resp. e_2, e_3, e_4) be a category of x_1 (resp. x_2, x_3, x_4) and $\mathbb{P}$ denote one probability measure. To generate a contingency table, the following log-linear model (see for instance Agresti (2002)) is used :

$$\begin{aligned}\log(\mathbb{P}(x_1 = e_1, \ldots, x_4 = e_4)) &= \log(\mu_{e_1e_2e_3e_4}) \\ &= (\lambda_{e_1}^{x_1} + \lambda_{e_2}^{x_2} + \beta_{e_1e_2}^{x_1x_2}) + (\lambda_{e_3}^{x_3} + \lambda_{e_4}^{x_4} + \beta_{e_3e_4}^{x_3x_4}) + \beta_{e_1e_4}^{x_1x_4},\end{aligned} \tag{2.27}$$

where $e_1, e_2, e_3, e_4 \in \{0, 1\}$. The parameters $\lambda_{e_1}^{x_1}, \lambda_{e_2}^{x_2}, \lambda_{e_3}^{x_3}$ and $\lambda_{e_4}^{x_4}$ designate the effect of each variable and the parameters $\beta_{e_1e_2}^{x_1x_2}$ and $\beta_{e_3e_4}^{x_3x_4}$ are interactions corresponding with cohesion terms in each group. The parameter $\beta_{e_1e_4}^{x_1x_4}$ is used to add some interactions between categories of variables belonging to different groups $\mathcal{C}_1$ and $\mathcal{C}_2$.

We simulate a contingency table corresponding to a global sample size $n = 1000$ using log-linear model (2.27) with the following values of the parameters $\lambda_0^{x_1} = \lambda_0^{x_3} = 1$, $\lambda_0^{x_2} = \lambda_0^{x_4} = 2$, $\beta_{00}^{x_1x_2} = -1.5$, $\beta_{00}^{x_3x_4} = -0.9$ and $\beta_{00}^{x_1x_4} = -0.5$. All the remaining parameters are set to zero. Thus the within groups cohesion parameters are high whereas the between groups interaction parameters are low in order to get well defined groups. We apply MCA on the categorical data corresponding with the generated contingency table. We retain $k = 2$ components and apply a planar rotation using the Varimax-based function h. Using (2.26) the corresponding analytic solution is $\hat{\theta} \approx \frac{\pi}{3}$. Figure 2.1 plots the criterion $h(\theta)$ for $\theta \in [-\pi, \pi]$ and we can verify on this figure that h is $\frac{\pi}{2}$-periodic and maximum in $\hat{\theta} \approx \frac{\pi}{3}$.

In order to visualize the impact of rotation on this simulated data, we plot in Figure 2.2 the four variables according to their correlation ratio to the first row standard component (in abscissa) and to the second row standard component (in ordinate) before and after planar rotation, respectively on the left and right side. As expected the variables are more clearly related to the components after rotation.

Let us also visualize in Figure 2.3 the impact of rotation on the representation of the categories on the first column principal plane of MCA : the principal coordinates of the categories before (resp. after) rotation are given in the first two columns of $\mathbf{Y}$ (resp. $\widetilde{\mathbf{Y}}$).

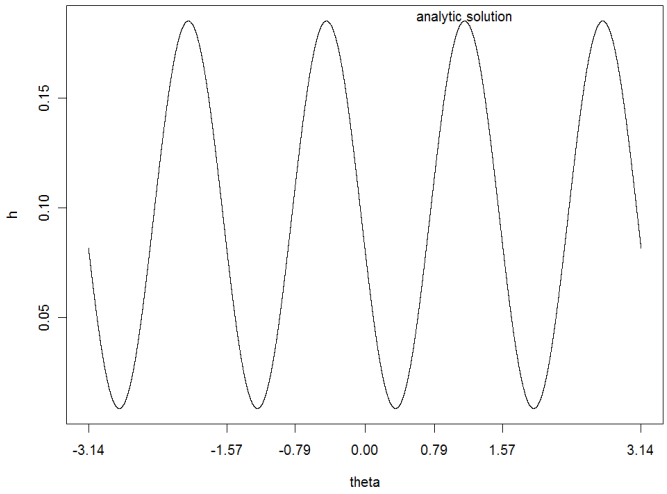

FIG. 2.1 – Graph of $\theta \mapsto h(\theta)$.

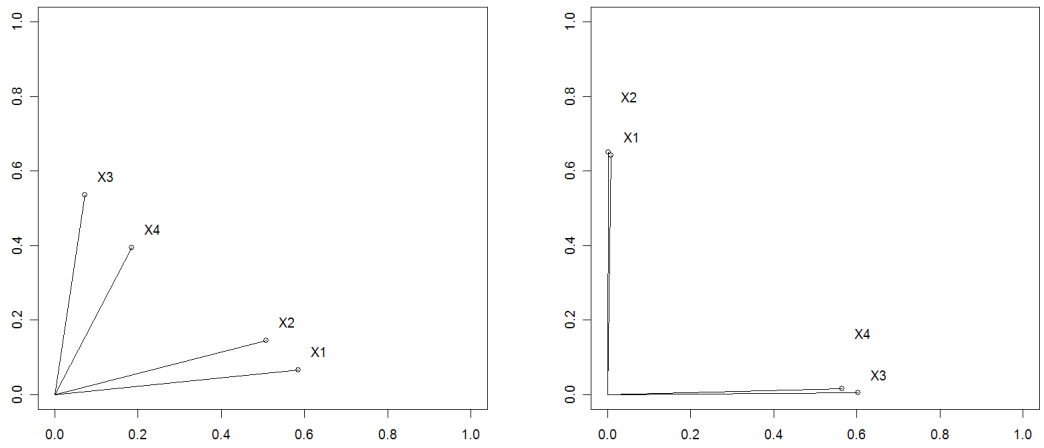

FIG. 2.2 – Plot of the correlation ratio matrix before rotation (on the left) and after planar rotation (on the right).

We see that after rotation the two categories of each variable are more clearly related to one of the two components. To conclude this simulated example provides expected results. Let us now study the impact of rotation on a real data set.

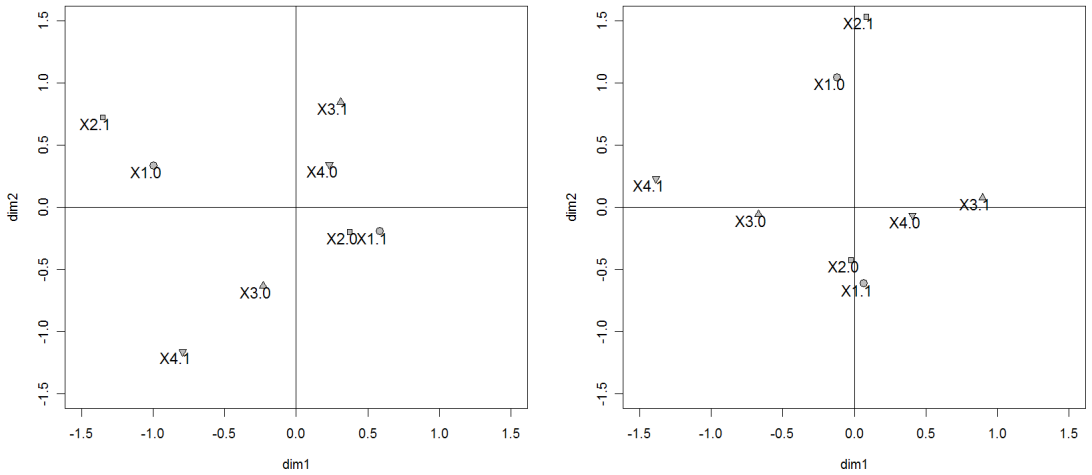

FIG. 2.3 – Plot of the categories in the first principal plane before rotation (on the left) and after rotation (on the right).

2.2.4 A real data application

In this section we apply this rotation methodology on a real data set in order to illustrate the benefits of using rotation in MCA. We consider a user satisfaction survey of pleasure craft operators on the "Canal des Deux Mers" located in South of France. This study has been realized from June to December 2008. It contains numerous questions with quantitative or qualitative answers. The sample size is $n = 1082$ pleasure craft operators. We focus here on a small number of qualitative variables in order to get clear graphical representations when plotting the categories on the principal plane. Although considering only four variables is of little practical interest, this application is useful to illustrate the rotation phenomenon. The four chosen variables are named “information”, “stopover”, “cleanliness” and “sailors”. They have each one three categories. The variable “information” deals with the quality of the information concerning sites worth visiting and its categories are 1-satisfactory, 2-unsatisfactory and 3-no opinion. The variable “stopover” is associated with the following question *What makes you decide to stop over at a particular place ?* and the possible answers are 1-necessity (supplies, time constraints, ...), 2-interest of stopover point (architecture, restaurant, landscape, ...) and 3-desire to be on dry land. The variable “cleanliness” is about the canal’s degree of cleanliness (1-clean, 2-average or 3-dirty). Finally the variable “sailors” is associated with the question *How would you describe other sailors you encountered ?* and its categories are 1-pleasant, 2-unpleasant and 3-do not know.

To visualize the effects of rotation on this data set, we first plot in Figure 2.4 the four variables according to their correlation ratio to the first two row standard principal components before and after rotation. We see that the association of the variables to the components is clearly easier after rotation. Thus two groups of variables appear,

the first one contains the variables "sailors" and "information" and the second one is composed of "cleanliness" and "stopover".

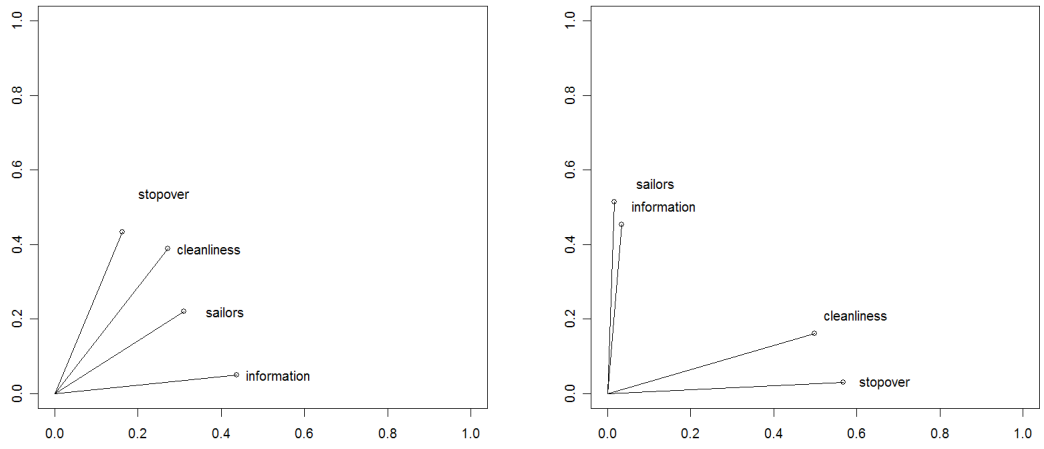

FIG. 2.4 – Plot of the correlation ratio matrix before rotation (on the left) and after planar rotation (on the right).

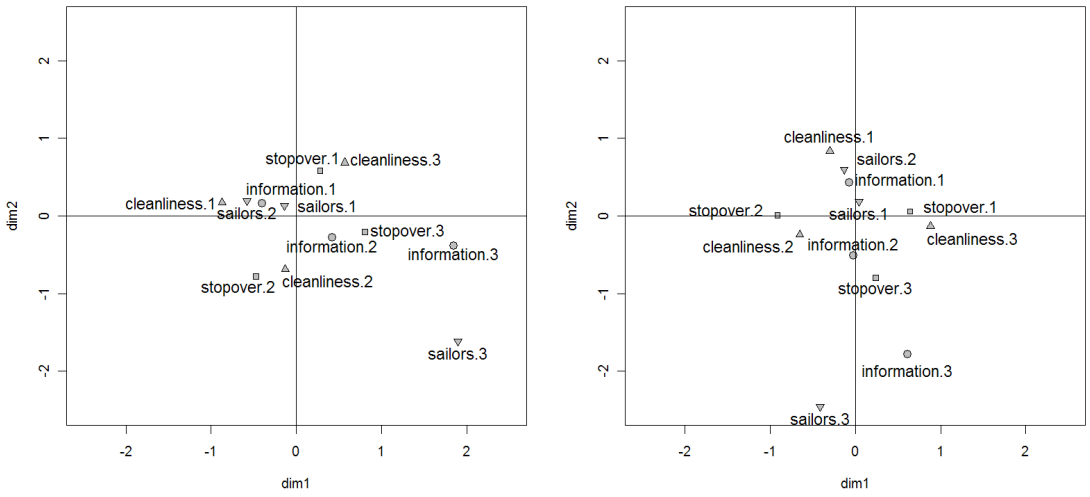

FIG. 2.5 – Plot of the categories in the first principal plane before rotation (on the left) and after rotation (on the right).

We observe in Figure 2.5 the impact of rotation on the representation of the categories on the first principal plane. The rotated components have a better discriminatory

capability than the initial ones. The first component represents on the left tourists who decide to stop over at a particular place because of interest and who think the canal's degree of cleanliness is average. On the contrary craft operators on the right stop over because of necessity and find the canal dirty. This component refers to the "expectations of pleasure craft operators" concerning the use of the canal. People who stop over because of necessity may not be pleased to be on dry land and are then quite demanding and critical of the canal out of hand. The second component could be labelled "opinion of the tourists" since it is discriminating between people with and people without an opinion either on the relationship with other sailors or on the information concerning sites worth visiting. Note that a second issue of discussion would be whether the respondents who scored the category 3-do not know when asked their opinion of other sailors are indeed individuals who do not encounter other sailors. Maybe they are people who like some and do not like others. Or people who do not feel like giving their opinion on other sailors. This latter view may be substantiated by the fact that these people also do not give their opinion on the information concerning sites worth visiting. This example on real data shows that rotation in MCA may help for the interpretation of the results since categories are better aligned along the components. Thus the labelling and interpretation of the components is easier.

2.2.5 Concluding remarks

In this article we propose a two-dimensional analytic solution for rotation in MCA using a Varimax-based criterion relying on the correlation ratio between the categorical variables and the MCA components. We have checked on a simulated example the accuracy of the given solution. We have also shown that rotation may be beneficial to real data since it may bring new elements for the interpretation of the results. However we are aware of the simplicity of the data we considered for an easier presentation and of the probable supplementary difficulty when dealing with more complex data sets.

When higher dimensionality is required, we use the practical algorithm of Kaiser (1958) which consists in computing the two-dimensional solution and then applying successive pairwise rotations. But although the Kaiser rotation procedure is a very popular techniques in data analysis, it is not without problems. Remedy against nonoptimal Varimax rotation have been proposed (Fraenkel, 1984 ; ten Berge, 1995) and may possibly be applied in the iterative planar rotation procedures proposed in this paper. ten Berge (1984) also showed that Varimax rotation can be interpreted as a special case of diagonalizing symmetric matrices and that the solution by De Leeuw and Pruzansky (1978) is essentially equivalent to the solution by Kaiser. We would like to obtain the same kind of result in MCA in order to link the rotation procedure proposed by Kiers (1991) for PCAMIX and thus MCA too, and the procedure proposed in this paper.

Moreover we think that the proposed planar solution in MCA can be used in divisive hierarchical methods for the clustering of qualitative variables. The well-known VARCLUS procedure of SAS software, planar rotation is used to help dividing at best a cluster of quantitative variables in two sub-clusters. The adaptation of this approach to qualitative variables is currently under investigation. Finally a future prospect on this

work would be to give the analytic expression of the rotation matrix for a dimension larger than two.

Acknowledgements

The authors are very greatful to the public corporation Voies Navigables de France in charge of managing the network of navigable waterways in Europe and Laurent Morillère, the manager of the private firm Enform, for providing the real data set.

Appendix

Appendix 1 : Short recall on row principal coordinates. Let $\mathbf{R} = \mathbf{D_r^{-1}F} - \mathbf{1}_n\mathbf{c}^t = \mathbf{D_r^{-1}}(\mathbf{F} - \mathbf{rc}^t)$ denote the $n \times q$ matrix of the centered row profiles. In a first step, MCA (or weighted PCA) searches for an axis with head vector $\mathbf{w}_1$ (of $\mathbf{D_c^{-1}}$-norm equal to 1) such that the vector $\mathbf{x}_1 = \mathbf{RD_c^{-1}w}_1$ of the $\mathbf{D_c^{-1}}$-projections of the rows of $\mathbf{R}$, has maximal variance (i.e. a maximal $\mathbf{D_r}$-norm). The first principal component $\mathbf{x}_1$ is then solution of the optimization problem :

$$\begin{cases} \max\limits_{\mathbf{x}\in\mathbb{R}^n} \parallel \mathbf{x} \parallel^2_{\mathbf{D_r}}, \\ \text{subject to } \mathbf{w}^t\mathbf{D_c^{-1}w} = 1, \end{cases} \tag{2.28}$$

which is equivalent, with the change of variable $\mathbf{v} = \mathbf{D_c^{-1/2}w}$, to the following simpler writting :

$$\begin{cases} \max\limits_{\mathbf{v}\in\mathbb{R}^p} \mathbf{v}^t\widetilde{\mathbf{F}}^t\widetilde{\mathbf{F}}\mathbf{v}, \\ \text{s.t.}\mathbf{v}^t\mathbf{v} = 1. \end{cases} \tag{2.29}$$

where $\widetilde{\mathbf{F}} = \mathbf{D_r^{-1/2}}(\mathbf{F} - \mathbf{rc}^t)\mathbf{D_c^{-1/2}}$. The first eigenvector $\mathbf{v}_1$ associated with the largest eigenvalue λ_1 of the matrix $\widetilde{\mathbf{F}}^t\widetilde{\mathbf{F}}$ is a solution of (2.29) and the sample variance of $\mathbf{x}_1$ is equal to λ_1. The other principal components are defined similarly by $\mathbf{x}_\alpha = \mathbf{RD_c^{-1/2}v}_\alpha$, for $\alpha = 2, \dots k$, where $\mathbf{v}_\alpha$ is the eigenvector associated with the αth largest eigenvalue λ_α of $\widetilde{\mathbf{F}}^t\widetilde{\mathbf{F}}$ and λ_α is the sample variance of $\mathbf{v}_\alpha$. The vectors $\mathbf{x}_\alpha$ are the k columns of the matrix of object scores $\mathbf{X} = \mathbf{RD_c^{-1/2}V}_k = \mathbf{D_r^{-1/2}}\widetilde{\mathbf{F}}\mathbf{V}_k$.

Appendix 2 : The barycentric property. Equation (2.15) of the column principal coordinate matrix gives $\mathbf{Y}^t = \Lambda_k\mathbf{V}_k^t\mathbf{D_c^{-1/2}}$. It follows from (2.14), that $\widetilde{\mathbf{F}}\mathbf{D_c^{-1/2}} = \mathbf{U}_k\mathbf{Y}^t$ and from (2.15) that $\mathbf{Y} = \mathbf{D_c^{-1}}(\mathbf{F} - \mathbf{rc}^t)^t\mathbf{D_r^{-1/2}U}_k = \mathbf{D_c^{-1}}(\mathbf{F} - \mathbf{rc}^t)^t\mathbf{X}^*$.

Remembering from the definition of $\mathbf{F}$ that $f_{is} = \frac{g_{is}}{np}$, $f_{i.} = \frac{1}{n}$ and $f_{.s} = \frac{n_s}{np}$, the

general term of $(\mathbf{F} - \mathbf{rc}^t)$ is then $\frac{g_{is}}{np} - \frac{n_s}{n^2 p}$. It gives :

$$\begin{aligned} y_{s\alpha} &= \sum_{i=1}^{n} \frac{np}{n_s}(\frac{g_{is}}{np} - \frac{n_s}{n^2 p})x^*_{i\alpha} \\ &= \frac{1}{n_s}\sum_{i=1}^{n} g_{is}x^*_{i\alpha} - \frac{1}{n}\sum_{i=1}^{n} x^*_{i\alpha} \\ &= \bar{x}^*_{s\alpha} \end{aligned}$$

Appendix 3 : Analytic expression of $h(\theta)$. For $k = 2$, criterion (2.19) simply writes

$$h(\theta) = \underbrace{p\sum_{j=1}^{p}\left(\sum_{s\in\mathcal{M}_j} f_{.s}\tilde{y}^2_{s1}\right)^2}_{M_1} + \underbrace{p\sum_{j=1}^{p}\left(\sum_{s\in\mathcal{M}_j} f_{.s}\tilde{y}^2_{s2}\right)^2}_{M_2} - \underbrace{\left(\sum_{j=1}^{p}\sum_{s\in\mathcal{M}_j} f_{.s}\tilde{y}^2_{s1}\right)^2}_{M_3} - \underbrace{\left(\sum_{j=1}^{p}\sum_{s\in\mathcal{M}_j} f_{.s}\tilde{y}^2_{s2}\right)^2}_{M_4} \tag{2.30}$$

where $\tilde{y}_{s1} = y_{s1}\cos\theta + y_{s2}\sin\theta$ and $\tilde{y}_{s2} = -y_{s1}\sin\theta + y_{s2}\cos\theta$ are the rotated loadings.

To maximize (2.30), we have to differentiate h with respect to θ and to set the derivative equal to zero. Note that this is only a necessary but not sufficient condition and we have to make sure it is a maximum. Let us first remark that $\frac{\partial\tilde{y}_{s1}}{\partial\theta} = \tilde{y}_{s2}$ and $\frac{\partial\tilde{y}_{s2}}{\partial\theta} = -\tilde{y}_{s1}$. Thus we have

$$\frac{\partial(M_1 + M_2)}{\partial\theta} = 4pA \quad \text{and} \quad \frac{\partial(M_3 + M_4)}{\partial\theta} = 4(A + B)$$

where

$$A = \sum_{j=1}^{p}\left\{(\sum_{s\in\mathcal{M}_j} f_{.s}\tilde{y}_{s1}\tilde{y}_{s2})(\sum_{t\in\mathcal{M}_j}(\tilde{y}^2_{t1} - \tilde{y}^2_{t2}))\right\}$$

and

$$B = \sum_{j=1}^{p}\sum_{l\neq j}^{p}(\sum_{s\in\mathcal{M}_j} f_{.s}\tilde{y}_{s1}\tilde{y}_{s2})(\sum_{t\in\mathcal{M}_l} f_{.t}(\tilde{y}^2_{t1} - \tilde{y}^2_{t2})).$$

It follows $\frac{\partial h}{\partial\theta} = 4(p-1)A - 4B$. Let us now remark that $\tilde{y}_{s1}\tilde{y}_{s2} = (y^2_{s2} - y^2_{s1})\frac{1}{2}\sin 2\theta + y_{s1}y_{s2}\cos 2\theta$, and $\tilde{y}^2_{t1} - \tilde{y}^2_{t2} = (y^2_{t1} - y^2_{t2})\cos 2\theta + 2y_{t1}y_{t2}\sin 2\theta$. Then we have

$$A = \sum_{j=1}^{p}\left\{\{\sum_{s\in\mathcal{M}_j} f_{.s}[(y^2_{s2} - y^2_{s1})\frac{1}{2}\sin 2\theta + y_{s1}y_{s2}\cos 2\theta]\} \times \{\sum_{t\in\mathcal{M}_j} f_{.t}[(y^2_{t1} - y^2_{t2})\cos 2\theta + 2y_{t1}y_{t2}\sin 2\theta]\}\right\}.$$

After a good deal of trigonometric identities and algebraic manipulations, we get :

$$\begin{aligned} A = \tfrac{1}{2}\textstyle\sum_{j=1}^{p}\sum_{s\in\mathcal{M}_j}\sum_{t\in\mathcal{M}_j} f_{.s}f_{.t} \quad & \{(y^2_{s2} - y^2_{s1})y_{t1}y_{t2} + (y^2_{t1} - y^2_{t2})y_{s1}y_{s2} \\ & + [(y^2_{s1} - y^2_{s2})y_{t1}y_{t2} + (y^2_{t1} - y^2_{t2})y_{s1}y_{s2}] \times \cos 4\theta \\ & + \tfrac{1}{2}[(y^2_{s2} - y^2_{s1})(y^2_{t1} - y^2_{t2}) + 2y_{s1}y_{s2}y_{t1}y_{t2}] \times \sin 4\theta\}. \end{aligned}$$

Then we have

$$A = \frac{1}{2}\sum_{j=1}^{p}\sum_{s\in\mathcal{M}_j}\sum_{t\in\mathcal{M}_j} f_{.s}f_{.t}\{\alpha_{st}\beta_{st} + \delta_{st}\gamma_{st}\cos 4\theta + \frac{1}{2}(\gamma_{st}^2 - \delta_{st}^2)\sin 4\theta\}$$

and

$$B = \frac{1}{2}\sum_{j=1}^{p}\sum_{l\neq j}^{p}\sum_{s\in\mathcal{M}_j}\sum_{t\in\mathcal{M}_j} f_{.s}f_{.t}\{\alpha_{st}\beta_{st} + \delta_{st}\gamma_{st}\cos 4\theta + \frac{1}{2}(\gamma_{st}^2 - \delta_{st}^2)\sin 4\theta\}$$

where the terms α_{st}, β_{st}, γ_{st} and γ_{st} are defined in (2.25). Finally, we get :

$$\frac{\partial h}{\partial \theta} = 2(a + b\cos 4\theta + c\sin 4\theta),$$

where the expression of a, b and c are given in (2.24).

References

Adachi, K. (2004). Oblique promax rotation applied to solutions in multiple correspondence analysis. *Behaviormetrika*, **31**, 1-12.

Agresti, A. (2002). *Categorical data analysis*, Second Edition, Wiley Series in Probability and Statistics.

Benzécri, J. P. (1973). *L'analyse des données : T. 2, l'analyse des correspondances*, Paris : Dunod.

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, **36**(1), 111-150.

de Leeuw, J., Pruzansky, S. (1978). A new computational method to fit the weighted Euclidean distance model. *Psychometrika*, **43**, 479-490.

Eckart, C., Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, **1**, 211-218.

Fraenkel, E. (1984). Variants of the varimax rotation method. *Biometrical journal*, **26**(7), 741-748.

Gifi, A. (1990). *Nonlinear Multivariate Analysis*, John Wiley & Sons.

Gower, J.C., Hand, D.J. (1996). *Biplots*, London : Chapman & Hall.

Greenacre, M.J. (1984). *Theory and Applications of Correspondence Analysis*, London : Academic Press.

Greenacre, M.J. (1993). Biplots in Correspondence Analysis. *Journal of Applied Statistics*, **20**(2), 251-269.

Greenacre, M.J. (2006). Tying up the loose ends in simple correspondence analysis. `http://www.econ.upf.es/docs/papers/downloads/940.pdf`.

Greenacre, M.J., Blasius, J. (2006). *Multiple Correspondence Analysis and Related Methods*, Chapman & Hall/CRC Press, London.

Hayashi, C. (1954). Multidimensional quantification–with applications to analysis of social phenomena. *Annals of the Institute of Statistical Mathematics*, **5**(2), 121-143.

Kaiser, H.F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, **23**(3), 187-200.

Kiers, H.A.L. (1991). Simple structure in Component Analysis Techniques for mixtures of qualitative and quantitative variables. *Psychometrika*, **56**, 197-212.

Lebart, L., Morineau, A., Warwick, K. M. (1984). *Multivariate descriptive analysis : Correspondence analysis and related techniques for large matrics*, New York, Wiley-Interscience.

Neudecker, H. (1981). On the matrix formulation of Kaiser's Varimax criterion. *Psychometrika*, **46**, 343-345.

Nishisato, S. (1980). *Analysis of categorical data : Dual Scaling and its applications*, Toronto : University of Toronto Press.

Nishisato, S. (1994). *Elements of Dual Scaling : An Introduction to Practical Data Analysis*, Hillsdale, NJ : Lawrence Erlbaum.

SAS (1990). User's guide, Version6, Vol 2., SAS Institute Inc. : Cary, North Caroline, *Psychometrika*, **31**(4), 535-538.

Sherin, R.J. (1966). A matrix formulation of Kaiser's varimax criterion, *Psychometrika*, **31**(4), 535-538.

ten Berge, J.M.F. (1984). A joint treatment of VARIMAX rotation and the problem of diagonalizing symmetric matrices simultaneously in the least-squares sense. *Psychometrika*, **49**, 347-358.

ten Berge, J.M.F. (1995). Suppressing permutations or rigid planar rotations : A remedy against nonoptimal varimax rotations. *Psychometrika*, **60**, 437-446.

Tenenhaus, M., Young, F. W. (1985). An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. *Psychometrika*, **50**, 91-119.

Van de Velden, M., Kiers, H. A. L. (2003). An application of rotation in correspondence analysis. *In H. Yanai, A. Okada, K. Shigemasu, Y. Kano, and J.J. Meulman, (Eds.), New Developments in Psychometrics*, Tokyo : Springer Verlag, 471-478.

Van de Velden, M., Kiers, H. A. L. (2005). Rotation in correspondence analysis. *Journal of Classification*, **22**, 251-271.

2.3 Clustering of categorical variables around latent variables

Abstract. Clustering of variables is studied as a way to arrange variables into homogeneous clusters, thereby organizing data into meaningful structures. Once the variables are clustered into groups such that variables are similar to the other variables belonging to their cluster, the selection of a subset of variables is possible. Several specific methods have been developed for the clustering of numerical variables. However concerning categorical variables, much less methods have been proposed. In this paper we extend the criterion used by Vigneau and Qannari (2003) in their Clustering around Latent Variables approach for numerical variables to the case of categorical data. The homogeneity criterion of a cluster of categorical variables is defined as the sum of the correlation ratio between the categorical variables and a latent variable, which is in this case a numerical variable. We show that the latent variable maximizing the homogeneity of a cluster can be obtained with Multiple Correspondence Analysis. Different algorithms for the clustering of categorical variables are proposed : iterative relocation algorithm, ascendant and divisive hierarchical clustering. The proposed methodology is illustrated by a real data application to satisfaction of pleasure craft operators.

Keywords : clustering of categorical variables, correlation ratio, iterative relocation algorithm, hierarchical clustering.

2.3.1 Introduction

Principal Component Analysis (PCA) and Multiple Correspondence Analysis (MCA) are appealing statistical tools for multivariate description of respectively numerical and categorical data. Rotated principal components fulfill the need to get more interpretable components. Clustering of variables is an alternative since it makes it possible to arrange variables into homogeneous clusters and thus to obtain meaningful structures. From a general point of view, variable clustering lumps together variables which are strongly related to each other and thus bring the same information. Once the variables are clustered into groups such that attributes in each group reflect the same aspect, the practicioner may be spurred on to select one variable from each group. One may also want to construct a synthetic variable. For instance in the case of quantitative variables, a solution is to realize a PCA (see Jolliffe, 2002) in each cluster and to retain the first principal component as the synthetic variable of the cluster. Another advantage that may be gained from the clustering of variables relates to the selection of a subset of variables. It is an alternative to procedures for discarding or selecting variables based on a statistical criterion that have been proposed by Jolliffe (1972), Mc Cabe (1984), Krzanowski (1987), Al-Kandari and Jolliffe (2001) or Guo et al. (2002) among others. The selection of a subset of variables is the aim of a lot of research in several areas of application. For instance in descriptive sensory profiling, this strategy of analysis can be used to reduce a list of attributes by selecting relevant and non redundant attributes. In biochemistry clustering genes based upon their expression patterns allows to predict

gene function. For preference studies when putting on the market new products, clustering of variables is also helpful to detect the existence of segments among the panel of consumers. Variable clustering can also be useful for association rules mining. Plasse et al. (2007) illustrate on an industrial application from the automotive industry the help of building homogeneous clusters of binary attributes for the discovering of relevant association rules mining. A conjoint use of variable clustering and Partial Least Squares (PLS) structural equations modeling is presented in Stan and Saporta (2005) in which clustering of variables is used to fulfill at best the underlying hypothesis in PLS approach of unidimensionality of the blocks of variables.

A simple and frequently used approach for variable clustering is to construct first a matrix of dissimilarities between the variables and then to apply classical cluster analysis methodology devoted to objects (units) which are able to deal with dissimilarity matrices (single, complete, average linkage hierarchical clustering or distance-based k-means). Partitioning Around Medoids can also deal with dissimilarity as input data (see Kaufman and Rousseeuw, 1990). Methods dealing only with numerical data like Ward or k-means among others can also be applied on the numerical coordinates obtained from Multidimensional Scaling of a previously built dissimilarity matrix.

Concerning quantitative variables, many authors have proposed different dissimilarity measures. Let us remind here some of these coefficients. Correlation coefficients (parametric or nonparametric) can be converted to different dissimilarities depending if the aim is to lump together correlated variables regardless of the sign of the correlation or if a negative correlation coeffcient between two variables shows disagreement between them. Soffritti (1999) defines a monotonous multivariate association measure that takes into account the within correlation and the number of variables of each group. A distance based on Escoufier's operator which takes the correlations as well as the variances of the variables into consideration has also been developped by Qannari et al. (1998). Note that this distance is also extended to the case of categorical variables and to a mixture of both types of data.

For categorical variables, many association measures can be used as χ^2, Rand, Belson, Jaccard, Sokal and Jordan among others. Some transformations are then in order to bring the coefficients into dissimilarity or distance measures. We can cite for instance the work of Abdallah and Saporta (1998) who consider various association measures and give the definition of a threshold beyond which two variables can be considered as linked.

Some specific approaches have also been developed for the clustering of variables. Once again for quantitative data, several specific methods have been proposed. We can cite among others the approach of Hastie et al. (2000) in genome biology or the recent work of Vichi and Saporta (2009), which aims at a simultaneous clustering of objects and a partitioning of variables. However the most famous one remains the VARCLUS procedure of SAS software. Two other interesting approaches that were independently proposed are Clustering around Latent Variables (CLV), introduced by Vigneau and Qannari (2003), and Diametrical Clustering of Dhillon et al. (2003). When the aim

is to lump together correlated variables regardless of the sign of the correlation, both methods aim at maximizing the sum over all clusters of the squared correlations between the variables and a latent variable.

Let us now tackle the issue of specific methods developed in view of clustering of categorical variables. Surprisingly, it has received much less attention than the numerical case. We can cite for instance the Likelihood Linkage Analysis proposed by Lerman (1993) which can deal with both numerical and categorical data.

In this paper we propose specific methods for the clustering of categorical variables. The homogeneity criterion of a cluster is not simply a distance based criterion but an extension of that used in CLV (Vigneau and Qannari, 2003). It is equal to the sum of the correlation ratio between the categorical variables and a latent variable, which is in this case a numerical variable. We show that the latent variable maximizing the homogeneity of a cluster is the first principal component obtained by MCA of the data of the cluster.

The overview of the paper is as follows. In Section 2.3.2, a specific measure of the homogeneity of a cluster of categorical variables is given and a partitioning criterion is defined. Section 2.3.3 is devoted to different clustering algorithms optimizing this specific criterion : iterative relocation algorithm, ascendant and divisive hierarchical clustering. In Section 2.3.4, a real data application relative to satisfactory of pleasure craft operators is treated. First the proposed hierarchical clustering algorithm is applied on a real data set. Then an empirical comparison of the performances of the different proposed algorithms is presented. Finally in Section 2.3.5, some concluding remarks and perspectives are given.

2.3.2 A correlation ratio based partitioning criterion for categorical variables

Let $\mathbf{X} = (x_{ij})$ be a data matrix of dimension (n, p) where a set of n objects are described on a set of p categorical variables. Let $\mathcal{V} = \{\mathbf{x}_1, \ldots, \mathbf{x}_p\}$ be the set of the p columns of $\mathbf{X}$, called for seek of simplicity categorical variables.

Homogeneity criterion of a cluster. Let $\mathcal{C} \subset \mathcal{V}$ be a cluster of categorical variables and $\mathbf{y}$ be a vector of $\mathbb{R}^n$ called latent variable. The homogeneity criterion of $\mathcal{C}$ measures the adequacy between the variables in $\mathcal{C}$ and $\mathbf{y}$:

$$S(\mathcal{C}) = \sum_{\mathbf{x}_j \in \mathcal{C}} \eta^2(\mathbf{x}_j, \mathbf{y}), \tag{2.31}$$

where $\eta^2(\mathbf{x}_j, \mathbf{y})$ stands for the correlation ratio between the categorical variable $\mathbf{x}_j$ and a numerical latent variable $\mathbf{y}$. This ratio is equal to the between group sum of squares of $\mathbf{y}$ in the groups defined by the categories of $\mathbf{x}_j$, divided by the total sum of squares of $\mathbf{y}$: $\eta^2(\mathbf{x}_j, \mathbf{y}) = \dfrac{\sum_{s \in \mathcal{M}_j} n_s (\bar{y}_s - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$, with n_s the frequency of category s, $\mathcal{M}_j$ the set of categories of $\mathbf{x}_j$ and $\bar{y}_s$ the mean value of $\mathbf{y}$ calculated on the objects belonging to

category s. The correlation ratio belongs to $[0,1]$ and measures the link between the categorical variable $\mathbf{x}_j$ and a numerical latent variable $\mathbf{y}$.

Definition of the latent variable of a cluster. In cluster $\mathcal{C}$, the latent variable $\mathbf{y}$ is defined to maximize the homogeneity criterion $S(\mathcal{C})$:

$$\mathbf{y} = \arg\max_{\mathbf{u}\in\mathbb{R}^n} \sum_{\mathbf{x}_j\in\mathcal{C}} \eta^2(\mathbf{x}_j, \mathbf{u}). \tag{2.32}$$

Result 1. The latent variable $\mathbf{y}$ of $\mathcal{C}$ is the first normalized eigenvector of $\widetilde{\mathbf{F}}\widetilde{\mathbf{F}}^t$, with $\widetilde{\mathbf{F}}$ defined in (2.33).

Proof. As $\eta^2(\mathbf{x}_j, \mathbf{u}) = \eta^2(\mathbf{x}_j, \alpha\mathbf{u})$, for any nonnull real α, the optimization problem (2.32) has an infinite set of solutions. We choose here to add the constraint $\mathbf{u}^t\mathbf{u} = 1$. To define the matrix $\widetilde{\mathbf{F}}$ we need to introduce usual notations from the theory of MCA. We can code the data of cluster $\mathcal{C}$ using indicator matrix $\mathbf{G}$ of dimension $n \times q$, with q the number of categories of the variables in $\mathcal{C}$, in which each category level is given a separate column and an entry of 1 indicates the relevant level of the category. The indicator matrix $\mathbf{G}$ is divided by its grand total $np_{\mathcal{C}}$, where $p_{\mathcal{C}}$ designates the number of variables in $\mathcal{C}$, to obtain the so-called "correspondence matrix" $\mathbf{F} = \frac{1}{np_{\mathcal{C}}}\mathbf{G}$, so that $\mathbf{1}_n^t\mathbf{F}\mathbf{1}_q = 1$, where, generically, $\mathbf{1}_i$ is an $i \times 1$ vector of ones. Furthermore, the row and column marginals define respectively the vectors of row and column masses $\mathbf{r} = \mathbf{F}\mathbf{1}_q$ and $\mathbf{c} = \mathbf{F}^t\mathbf{1}_n$. Let $\mathbf{D_r} = \mathrm{diag}(\mathbf{r})$ and $\mathbf{D_c} = \mathrm{diag}(\mathbf{c})$ be the diagonal matrices of these masses. In this particular case, the ith element of $\mathbf{r}$ is $f_{i.} = \frac{1}{n}$ and the sth element of $\mathbf{c}$ is $f_{.s} = \frac{n_s}{np_{\mathcal{C}}}$. We can now define the matrix

$$\widetilde{\mathbf{F}} = \mathbf{D_r}^{-1/2}(\mathbf{F} - \mathbf{r}\mathbf{c}^t)\mathbf{D_c}^{-1/2}. \tag{2.33}$$

Let us first show that if $\bar{u} = 0$ and $\mathrm{var}(\mathbf{u}) = \frac{1}{n}$, we have $\mathbf{u}^t\widetilde{\mathbf{F}}\widetilde{\mathbf{F}}^t\mathbf{u} = \frac{1}{p_{\mathcal{C}}}\sum_{\mathbf{x}_j\in\mathcal{C}}\eta^2(\mathbf{x}_j, \mathbf{u})$. Remembering from the definition of $\mathbf{F}$ that $f_{is} = \frac{g_{is}}{np_{\mathcal{C}}}$, the general term of $\mathbf{D_r}^{-1/2}(\mathbf{F} - \mathbf{r}\mathbf{c}^t)\mathbf{D_c}^{-1/2}$ is then $\tilde{f}_{is} = \frac{\sqrt{n_s}\sqrt{p_{\mathcal{C}}}}{n_s}(\frac{g_{is}}{p_{\mathcal{C}}} - \frac{n_s}{np_{\mathcal{C}}})$. It follows that $\sum_{i=1}^n \tilde{f}_{is}u_i = \frac{\sqrt{n_s}}{\sqrt{p_{\mathcal{C}}}}\bar{u}_s$, where $\bar{u}_s$ is the mean value of $\mathbf{u}$ calculated on the objects belonging to category s. Then we get

$$\mathbf{u}^t\widetilde{\mathbf{F}}\widetilde{\mathbf{F}}^t\mathbf{u} = \frac{1}{p_{\mathcal{C}}}\sum_{\mathbf{x}_j\in\mathcal{C}}\sum_{s\in\mathcal{M}_j} n_s\bar{u}_s^2 = \frac{\frac{1}{p_{\mathcal{C}}}\sum_{\mathbf{x}_j\in\mathcal{C}}\sum_{s\in\mathcal{M}_j}\frac{n_s}{n}(\bar{u}_s - 0)^2}{\frac{1}{n}} = \frac{1}{p_{\mathcal{C}}}\sum_{\mathbf{x}_j\in\mathcal{C}}\eta^2(\mathbf{x}_j, \mathbf{u}). \tag{2.34}$$

Moreover as the first normalized eigenvector of $\widetilde{\mathbf{F}}\widetilde{\mathbf{F}}^t$ maximizes $\mathbf{u}^t\widetilde{\mathbf{F}}\widetilde{\mathbf{F}}^t\mathbf{u}$ with respect to $\mathbf{u} \in \mathbb{R}^n$ under the constraint $\mathbf{u}^t\mathbf{u} = 1$, it is a solution of (2.32). Since it is normalized, its variance is equal to $\frac{1}{n}$. Then we have to check that it is centered. If $\widetilde{\mathbf{F}}$ is supposed to be of rank r, the Singular Value Decomposition (SVD) of $\widetilde{\mathbf{F}}$ is $\widetilde{\mathbf{F}} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^t$, where $\mathbf{\Lambda}$ contains the r nonnull singular values of $\widetilde{\mathbf{F}}^t\widetilde{\mathbf{F}}$ and $\widetilde{\mathbf{F}}\widetilde{\mathbf{F}}^t$ sorted in decreasing order, $\mathbf{U}$ (resp. $\mathbf{V}$) is the matrix whose columns are the normalized eigenvectors of $\widetilde{\mathbf{F}}\widetilde{\mathbf{F}}^t$ (resp.

$\widetilde{\mathbf{F}}^t\widetilde{\mathbf{F}}$) associated with the nonnull eigenvalues. Thus $\mathbf{U} = \widetilde{\mathbf{F}}\mathbf{V}\mathbf{\Lambda}^{-1}$ and then the first normalized eigenvector of $\widetilde{\mathbf{F}}\widetilde{\mathbf{F}}^t$, as a linear combination of the columns of $\widetilde{\mathbf{F}}$ which are centered, is in turn centered, which completes the proof.

Result 2. The latent variable $\mathbf{y}$ is colinear with the first principal component issued from MCA of the row profiles of the data matrix of $\mathcal{C}$.

Proof. MCA is defined here as the application of weighted PCA to the centered row profiles matrix $\mathbf{D_r^{-1}}(\mathbf{F} - \mathbf{rc}^t)$ with distances between profiles measured by the chi-squared metric defined by $\mathbf{D_c^{-1}}$. The $n \times r$ matrix Ψ of row principal coordinates is then defined by $\Psi = \mathbf{D_r^{-1/2}}\widetilde{\mathbf{F}}\mathbf{V}$, with the expression of $\widetilde{\mathbf{F}}$ given in (2.33). From the SVD of $\widetilde{\mathbf{F}}$, we get $\Psi = \mathbf{D_r^{-1/2}}\mathbf{U}\mathbf{\Lambda}$, thereby implying that the latent variable, defined as the first normalized eigenvector of $\widetilde{\mathbf{F}}\widetilde{\mathbf{F}}^t$, is colinear with the first principal component obtained with MCA.

Partitioning criterion. We denote by $\mathcal{P}_K = \{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$ a partition of $\mathcal{V}$ into K clusters and by $\mathcal{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_K\}$ a set of K latent variables. The paper addresses the problem of partitioning a set of p variables into K disjoint clusters in which variables are similar to the other variables belonging to their cluster and dissimilar to variables that belong to different clusters. The partitioning criterion concentrates on maximizing the cohesion (homogeneity) of the clusters in the partition :

$$H(\mathcal{P}_K) = \sum_{k=1}^{K} S(\mathcal{C}_k), \tag{2.35}$$

with $S(C_k)$ defined in (2.31). In the next section, we propose different clustering algorithms using this criterion.

2.3.3 Different clustering algorithms

Given criterion (2.35) measuring the homogeneity of a partition of a set of variables into K disjoint clusters, there are different possible clustering algorithms for maximizing this criterion. First we describe an iterative relocation algorithm, then two hierarchical algorithms are proposed : ascendant and divisive.

Iterative relocation algorithm. A first solution to search for optimal partitions of the variables is given by an iterative algorithm in the course of which the variables are allowed to move in and out of the groups at the different stages of the algorithm achieving at each stage an increase of criterion (2.35). This partitioning algorithm runs as follows :

(a) *Initialization step* : The specification of this step may be reached by different ways. The first solution consists in computing the first K principal components issued from MCA of the centered row profiles matrix of $\mathbf{X}$. As has been described in Section 2.3.2, each component can play the role of the latent variable of a cluster

with itself as single member. Then we go to step (c) for the allocation step. This initialization can be coupled with a rotation to start with a better partition as in the VARCLUS procedure. We can use for instance the planar rotation iterative procedure for rotation in MCA proposed by Chavent et al. (2009). By doing this, the values of the correlation ratio between the variables and the latent variables are either large or small and the allocation is easier and then may be better. Another solution is to select randomly K variables of $\mathcal{V}$ and to apply MCA on the row profiles obtained with the data provided by each single variable in order to get K latent variables. These latent variables define at the beginning K clusters each containing only one member. Then we go to step (c). As it is well-known that iterative relocation algorithms provide a local optimum, the proposed iterative relocation algorithm is run several times, with multiple random initializations and we retain the best partition in sense of our partitioning criterion (2.35).

(b) *Representation step* : For all k in $1, ..., K$, we compute the latent variable $\mathbf{y}_k$ of $\mathcal{C}_k$ as the first normalized eigenvector of $\widetilde{\mathbf{F}}_k\widetilde{\mathbf{F}}_k^t$, where $\widetilde{\mathbf{F}}_k$ is defined in (2.33) for a generic cluster.

(c) *Allocation step* : Each variable is then assigned to the cluster which latent variable is closest to it in sense of correlation ratio. For all j in $1, ..., p$, find ℓ such that $\ell = \arg\max_{k=1,...,K} \eta^2(\mathbf{x}_j, \mathbf{y}_k)$. Let $\mathcal{C}_k$ be the previous cluster of $\mathbf{x}_j$. Then if $\ell \neq k$, $\mathcal{C}_\ell \leftarrow \mathcal{C}_\ell \cup \{\mathbf{x}_j\}$ and $\mathcal{C}_k \leftarrow \mathcal{C}_k \backslash \{\mathbf{x}_j\}$.

(d) If nothing changes in step (c) then *stop*, else return to step (b).

An empirical comparison of the efficiency of the iterative relocation algorithm according to the initialization step (a) is provided in Section 2.3.4.

Ascendant hierarchical approach. We propose herein a hierarchical clustering strategy based on the same criterion (2.35). First from Result 1, this criterion can be rewritten as follows :

$$H(\mathcal{P}_K) = \sum_{k=1}^{K} p_k \lambda_k,$$

where p_k is the number of variables in $\mathcal{C}_k$ and λ_k is the largest eigenvalue of matrix $\widetilde{\mathbf{F}}_k\widetilde{\mathbf{F}}_k^t$, with $\widetilde{\mathbf{F}}_k$ defined in (2.33) for a generic cluster.

In the ascendant hierarchical clustering algorithm, one recursively merges two clusters, starting from the stage in which each variable is considered to form a cluster by itself to the stage where there is a single cluster containing all variables. Given the current partition $\mathcal{P}_K = \{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$, two clusters are merged in order to find a partition $\mathcal{P}_{K-1}$ which contains $K-1$ clusters and optimizes the chosen cohesion measure (2.35). More precisely because

$$H(\mathcal{P}_{K-1}) = H(\mathcal{P}_K) - \underbrace{(S(\mathcal{C}_l) + S(\mathcal{C}_m) - S(\mathcal{C}_l \cup \mathcal{C}_m))}_{h(\mathcal{C}_l \cup \mathcal{C}_m)}, \tag{2.36}$$

the merging of two clusters $\mathcal{C}_l$ and $\mathcal{C}_m$ results in a variation of criterion (2.35) given by :

$$h(\mathcal{C}_l \cup \mathcal{C}_m) = \lambda_l + \lambda_m - \lambda_{l \cup m}.$$

We can prove (see Appendix) that :

$$\lambda_{l\cup m} \leqslant \lambda_l + \lambda_m, \tag{2.37}$$

which implies that the merging of two clusters at each step results in a decrease in criterion (2.35). Therefore the strategy consists in merging the two clusters that result in the smallest decrease in the cohesion measure.

Divisive hierarchical approach. Divisive hierarchical clustering reverses the process of agglomerative hierarchical clustering, by starting with all variables in one cluster, and successively dividing each cluster into two sub-clusters. Given the current partition $\mathcal{P}_K = \{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$, one cluster $\mathcal{C}_l$ is split in order to find a partition $\mathcal{P}_{K+1}$ which contains $K+1$ clusters and optimizes the chosen adequacy measure (2.35). More precisely, at each stage, the divisive hierarchical clustering method

- splits a cluster $\mathcal{C}_l$ into a bipartition $(\mathcal{A}_l, \bar{\mathcal{A}}_l)$.
- chooses in the partition $\mathcal{P}_K$ the cluster $\mathcal{C}_l$ to be split in such a way that the new partition $\mathcal{P}_{K+1}$ has a maximum cohesion measure.

The problem of how to split a cluster. In order to split optimally a cluster $\mathcal{C}_l$ one has to choose the bipartition $(\mathcal{A}_l, \bar{\mathcal{A}}_l)$, amongst the $2^{p_l-1}-1$ possible bipartitions of this cluster of p_l variables (with p_l the number of variables in $\mathcal{C}_l$), which maximizes criterion (2.35). It is clear that such complete enumeration provides a global optimum but is computationally prohibitive. The iterative relocation algorithm proposed above (with $K=2$) provides at least one locally optimal division.

Selecting the cluster to be split. In divisive clustering, the set of clusters obtained after $K-1$ divisions is a hierarchy $\mathcal{H}_K$ whose singletons are the K clusters of the partition $\mathcal{P}_K$ obtained in the last stage of the procedure. Because the resulting hierarchy can be considered as a partial hierarchy halfway between the top and bottom levels, it is referred to as an upper hierarchy (Mirkin, 2005). This upper hierarchy is then indexed by h so that in the dendrogram the height of a cluster $\mathcal{C}_l$ split into two sub-clusters $\mathcal{A}_l$ and $\bar{\mathcal{A}}_l$ is :

$$h(\mathcal{C}_l) = S(\bar{\mathcal{A}}_l) + S(\mathcal{A}_l) - S(\mathcal{C}_l).$$

When the divisions are continued until giving singleton clusters, all of the clusters can be systematically split and the full hierarchy $\mathcal{H}_n$ can be indexed by h. When the divisions are not continued down to $\mathcal{H}_n$, the clusters are not systematically split : in order to have the dendrogram of the upper hierarchy $\mathcal{H}_K$ built at the "top" (the $K-1$ largest) levels of the dendrogram of $\mathcal{H}_n$, a cluster represented higher in the dendrogram of $\mathcal{H}_n$ has to be split before the others. The proposed procedure then chooses to split the cluster $\mathcal{C}_l$ with the maximum value $h(\mathcal{C}_l)$. Consequently because

$$H(\mathcal{P}_{K+1}) = H(\mathcal{P}_K) + h(\mathcal{C}_l)$$

maximizing $h(\mathcal{C}_l)$ ensures that the new partition $\mathcal{P}_{K+1} = \mathcal{P}_K \cup \{\mathcal{A}_l, \bar{\mathcal{A}}_l\} - \{\mathcal{C}_l\}$ has a maximum cohesion measure.

Remark. The index h of the hierarchy is well positive (see Appendix for the proof)

but we have not yet demonstrated that it is a monotone increasing function, that is $\forall \mathcal{A}, \mathcal{B} \in \mathcal{H}$, if $\mathcal{A} \subset \mathcal{B}$, then $h(\mathcal{A}) \leq h(\mathcal{B})$. Note that in practice, we have never observed inversion phenomenom.

2.3.4 Real data application

In the subsequent clustering of categorical variables is applied to a real data set. A user satisfaction survey of pleasure craft operators on the "Canal des Deux Mers", located in South of France, was carried out by the public corporation "Voies Navigables de France" responsible for managing and developing the largest network of navigable waterways in Europe. This study was realized from June to December 2008. Pleasure craft operators were asked their opinion about numerous questions with categorical answers, thus providing $p = 85$ categorical variables, each having two or three categories of response. The objective of the present case study is to examine the redundancy among variables in order to select a subset of attributes to be used in further studies saving time for the respondents, money for the edition of the questionnaires and the statistical treatment of the data.

First an application is reached on a reduced[1] data set to illustrate the interpretation of the results obtained with the proposed ascendant hierarchical clustering algorithm. Then the different algorithms of clustering (iterative relocation algorithm and its various initializations, ascendant and divisive hierarchical clustering) are applied on the complete data set to compare empirically the advantages of each approach.

2.3.4.1 Illustration on a reduced data set

We focus here on fourteen categorical variables described in Table 2.1. After removal of individuals with missing values for some of the questions, the sample size is $n = 709$ pleasure craft operators.

The ascendant hierarchical approach described in Section 2.3.3 is applied. Figure 2.6 shows the resulting dendrogram. The evolution of the aggregation criterion h is given in Figure 2.7. This figure should be read as a scree-graph. The aggregation criterion jumped when passing from 5 clusters to 4 clusters. This should suggest that "different" clusters are being merged and therefore the partition into 5 clusters is retained. The choice of the number of clusters can also be based on practical considerations such as the easiness of interpretation. Here the partition into 5 clusters provides satisfactory interpretable results. In a subsequent stage, the iterative relocation algorithm is performed with $K = 5$ clusters with as initial partition the one derived from the hierarchical procedure. In this case study, this complement stage leads to no improvement of criterion (2.35) as no variable changes membership.

Table 2.2 describes the 5-clusters partition of the 14 categorical variables. For instance cluster $\mathcal{C}_4$ contains variables dealing with the information on the use of the canal : sites worth visiting, leisure activity and historical canal sites. The value in brackets

[1]We only consider here a subset of 14 variables over the 85 categorical variables.

Name of the variable	**Description of the variable**	**Categories**
$\mathbf{x}_1$="sites worth visiting"	*What do you think about information you were provided with concerning sites worth visiting ?*	satisfactory, unsatisfactory, no opinion
$\mathbf{x}_2$="leisure activity"	*How would you rate the information given on leisure activity ?*	
$\mathbf{x}_3$="historical canal sites"	*What is your opinion concerning tourist information on historical canal sites (locks, bridges, etc.) ?*	
$\mathbf{x}_4$="manoeuvres"	*At the start of your cruise, were you sufficiently aware of manoeuvres at locks ?*	yes, no
$\mathbf{x}_5$="authorized mooring"	*At the start of your cruise, were you sufficiently aware of authorized mooring ?*	
$\mathbf{x}_6$="safety regulations"	*At the start of your cruise, were you sufficiently aware of safety regulations ?*	
$\mathbf{x}_7$="information on services"	*Please give us your opinion about signs you encountered along the way concerning information regarding services.*	satisfactory, unsatisfactory
$\mathbf{x}_8$="number of taps"	*What do you think about number of taps on your trip ?*	sufficient, unsufficient
$\mathbf{x}_9$="cost of water"	*The general cost of water is ...*	inexpensive, average, expensive
$\mathbf{x}_{10}$="cost of electricity"	*The general cost of electricity is ...*	
$\mathbf{x}_{11}$="visibility of electrical outlets"	*What is your opinion of visibility of electrical outlets ?*	sufficient, unsufficient
$\mathbf{x}_{12}$="number of electrical outlets"	*What do you think about number of electrical outlets on your trip ?*	
$\mathbf{x}_{13}$="cleanliness"	*How would you describe the canal's degree of cleanliness ?*	clean, average, dirty
$\mathbf{x}_{14}$="unpleasant odours"	*Were there unpleasant odours on the canal ?*	none, occasional, frequent

TAB. 2.1 – Description of the 14 categorical variables.

shows the correlation ratio between a variable of the cluster and the corresponding latent variable. We see that the variables in a cluster are highly related with their latent variable. Table 2.3 gives the values of the Tschuprow coefficient between the variables of cluster $\mathcal{C}_4 = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$ and the remaining ones. We see that the variables are more related to the other variables belonging to their cluster than to variables that belong to different clusters. Then an advantage which may be gained from the clustering of variables relates to the selection of a subset of variables. For instance in this case study we could reduce the number of questions in the survey by selecting one variable in each cluster using the correlation ratio values given in Table 2.2.

$\mathcal{C}_1$ **: environment**	$\mathcal{C}_2$ **: navigation rules**	$\mathcal{C}_3$ **: cost of services**
cleanliness (0.68) unpleasant odours (0.68)	manoeuvres (0.66) authorized mooring (0.71) safety regulations (0.69)	cost of water (0.84) cost of electicity (0.84)
$\mathcal{C}_4$ **: use of the canal**	$\mathcal{C}_5$ **: available services**	
sites worth visiting (0.71) leisure activity (0.69) historical canal sites (0.46)	information on services (0.40) number of taps (0.59) visibility of electrical outlets (0.65) number of electrical outlets (0.71)	

TAB. 2.2 – Partition of the 14 categorical variables into 5 clusters (correlation ratio between a variable of the cluster and the corresponding latent variable).

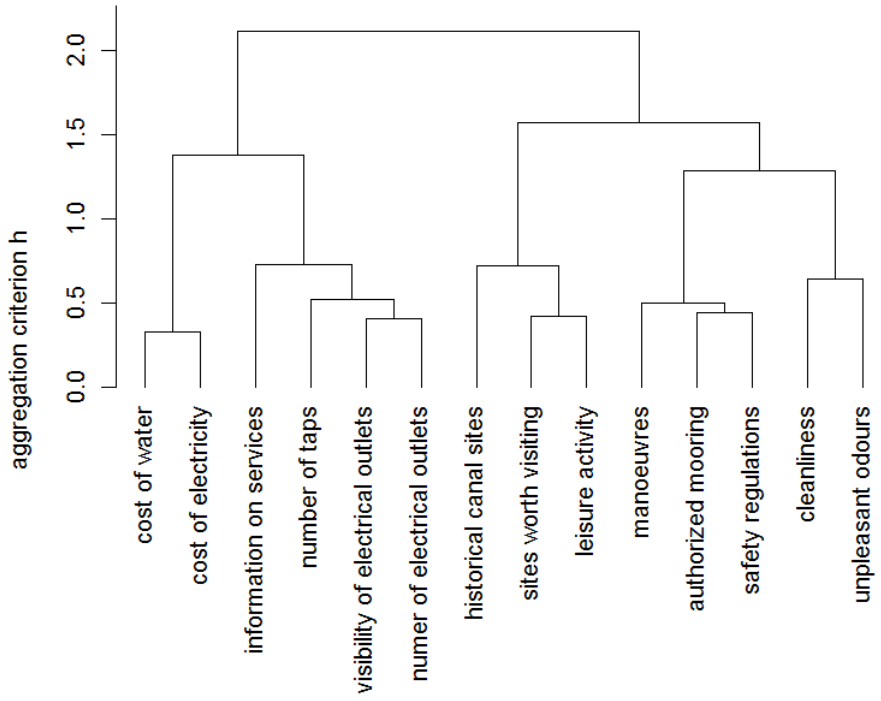

FIG. 2.6 – Dendrogram of the ascendant hierarchical clustering of the 14 categorical variables.

	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$	$\mathbf{x}_6$	$\mathbf{x}_7$	$\mathbf{x}_8$	...	$\mathbf{x}_{14}$
$\mathbf{x}_1$	**1.00**	**0.36**	**0.24**	0.09	0.10	0.11	0.08	0.06	...	0.05
$\mathbf{x}_2$	**0.36**	**1.00**	**0.20**	0.10	0.11	0.13	0.11	0.07	...	0.03
$\mathbf{x}_3$	**0.24**	**0.20**	**1.00**	0.02	0.04	0.05	0.11	0.08	...	0.05

TAB. 2.3 – Values of the Tschuprow coefficient between the variables of $\mathcal{C}_4$ and the remaining ones.

2.3.4.2 Empirical study and comparison of the different proposed clustering algorithms

We focus here on all the $p = 85$ categorical variables from the survey.

The proportion of explained cohesion. The clustering objective is formally expressed as the maximization of criterion (2.35) which can be perceived as a cohesion measure of the clusters in the partition. The cohesion criterion of a given partition $\mathcal{P}_K$ is given by $\mathcal{H}(\mathcal{P}_K) = \sum_{k=1}^{K} \sum_{\mathbf{x}_j \in \mathcal{C}_k} \eta^2(\mathbf{x}_j, \mathbf{y}_k)$, with $\mathbf{y}_k$ the latent variable of cluster $\mathcal{C}_k$. Similarly the total cohesion of a set $\mathcal{V}$ of p variables can be measured by $\mathcal{H}(\mathcal{V}) = \sum_{j=1}^{p} \eta^2(\mathbf{x}_j, \mathbf{y})$ with $\mathbf{y}$ the latent variable (or total representative) of $\mathcal{V}$. The cohesion measure is equal to $\mathcal{H}(\mathcal{V})$ for the single cluster $(\mathcal{V})$ and to p for the singleton partition. Hence the quality of the partitions $\mathcal{P}_K$ built by the three methods from the

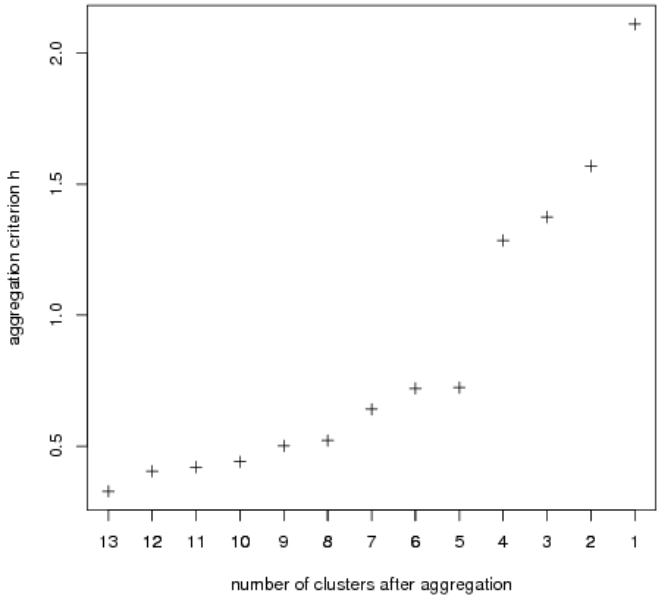

FIG. 2.7 – Evolution of the aggregation criterion h of the ascendant hierarchical clustering of the 14 categorical variables.

same set of variables $\mathcal{V}$, can be ranked using the proportion of gain in cohesion, that is the ratio of the gain obtained with $\mathcal{P}_K$ to the maximum gain that can be reached with the singleton partition :

$$E(\mathcal{P}_K) = \frac{\mathcal{H}(\mathcal{P}_K) - \mathcal{H}(\mathcal{V})}{p - \mathcal{H}(\mathcal{V})}.$$

This lies between 0% for the single cluster ($\mathcal{V}$) and 100% for the singleton partition. Because E increases with the number K of clusters of the partition, it can be used only to compare partitions having the same number of clusters. In the following, we assume that a partition $\mathcal{P}_K$ is better than a partition $\mathcal{P}'_K$ if $E(\mathcal{P}_K) > E(\mathcal{P}'_K)$. We will call $E(\mathcal{P}_K)$ the proportion of explained cohesion by the partition $\mathcal{P}_K$.

Different initializations of the iterative relocation algorithm. As has already been pointed, the iterative relocation algorithm involves an initialization step that can be specified for instance by the three techniques proposed in Section 2.3.3. The aim of the following is to study the impact of the initialization on the quality of the obtained partition. Table 2.4 gives the proportion $E(\mathcal{P}_K)$ of explained cohesion for partitions from $K = 2$ to 20 clusters. Each column displays this proportion obtained respectively with the initialization via the first K principal components, the first K rotated principal components and the best of $N = 30$ random initializations.

The partitions obtained with the initialization via the rotated principal components

K	K principal components	K rotated principal components	$N = 30$ random initializations
2	3.19	**3.39**	1.48
3	5.95	**6.40**	5.25
4	8.03	**8.87**	8.57
5	10.55	10.13	**11.60**
6	11.86	12.48	**14.62**
7	14.29	14.94	**17.13**
8	15.70	17.74	**18.87**
9	17.85	18.24	**21.22**
10	19.67	20.87	**23.83**
11	21.26	22.18	**25.80**
12	22.46	24.66	**27.76**
13	23.69	26.31	**29.16**
14	24.89	27.68	**31.41**
15	26.47	28.21	**33.51**
16	27.66	29.71	**35.33**
17	29.46	31.16	**37.05**
18	29.92	32.56	**38.21**
19	31.46	34.16	**40.53**
20	32.68	35.74	**42.39**

TAB. 2.4 – Iterative relocation algorithm : comparison of the proportion $E(\mathcal{P}_K)$ of explained cohesion with various initializations.

are always better (except for $K = 5$ where it is almost equal) than those obtained with the principal components. Thus the complement step of rotation seems to be efficient. For the third column, the iterative relocation algorithm is executed $N = 30$ times with different random initial seeds and the best solution in sense of the partitioning criterion (2.35) is retained. The partitions obtained with the rotated principal components are better up to 4 clusters and the iterative relocation algorithm with random initializations takes the lead from 5 clusters onwards. Moreover the gain in the proportion of explained cohesion increases as the number of clusters increases (18.6%=(42.39-35.74)/35.74 for 20 clusters versus 14.5%=(11.60-10.13)/10.13 for 5 clusters). Note that one possible explanation for the worse results of the multiple random initializations is probably that there is no strong structure in the data for a small number K of clusters so that the draw of some random initial seeds does not provide good partitions. As a rule concerning the iterative relocation methodology, running the algorithm several times with different initial partition in each run seems to be a satisfactory strategy.

Comparison of the different approaches. Now, we compare the results of the iterative relocation algorithm with multiple random initializations, which provides the best partitions in sense of $E(\mathcal{P}_K)$, with ascendant and divisive hierarchical clustering.

Comparing the first two columns of Table 2.5, we see that the ascendant hierarchical clustering is more efficient than the divisive one. A possible explanation is that the agglomerative algorithm is "stepwise optimal" : at each step, the amalgamation chosen is the best (in terms of the specified clustering criterion) that can be made at that time. However one reason for having worse results for the divisive approach is probably the way of splitting a cluster into two sub-clusters. This is reached by iterative relocation algorithm (with $N = 30$ multiple random initializations) and thus the bipartition ob-

K	ascendant hier. clust.	divisive hier. clust.	iterative reloc. algo. ($N = 30$ random init.)	ascendant hier. algo. + iterative relocation
2	3.01	2.58	1.48	**3.26**
3	5.73	4.51	5.25	**6.18**
4	8.19	7.31	8.57	**9.05**
5	10.63	9.31	11.60	**11.62**
6	12.92	10.95	**14.62**	13.99
7	15.13	12.36	**17.13**	15.99
8	17.19	13.61	**18.87**	17.98
9	19.23	14.92	**21.22**	19.83
10	21.24	16.62	**23.83**	21.88
11	23.09	18.62	**25.80**	23.67
12	24.93	19.72	**27.76**	25.45
13	26.72	21.14	**29.16**	27.35
14	28.48	22.61	**31.41**	29.07
15	30.16	23.87	**33.51**	30.73
16	31.78	25.40	**35.33**	32.03
17	33.38	26.73	**37.05**	33.63
18	34.92	28.09	**38.21**	35.05
19	36.45	29.38	**40.53**	36.54
20	37.94	30.95	**42.39**	38.03

TAB. 2.5 – Comparison of the proportion $E(\mathcal{P}_K)$ of explained cohesion with different algorithms of clustering.

tained may not be optimal, thus altering the quality of the hierarchy built with the divisive clustering.

Then we compare the results obtained with the ascendant hierarchical procedure with those reached with the iterative relocation algorithm (with $N = 30$ random initial seeds). The latter always provides better partitions in sense of the cohesion measure (2.35), except as seen previously for a small number of clusters ($K = 2, 3$). Once again the gain in the proportion of explained cohesion increases as the number of clusters increases (11.2% for 20 clusters versus 4.6% for 4 clusters). However one may prefer the hierarchical technique which has the advantage to build a hierarchy of nested partitions of the variables and then may be beneficial for the interpretation of the results and the choice of a number K of clusters.

We also propose in the fourth column of Table 2.5 to complement the ascendant hierarchical clustering by the iterative relocation algorithm with as initial partition the one derived from the hierarchical procedure. For a given partition $\mathcal{P}_K$ this step aims at improving criterion (2.35) by allowing variables to change membership. Thus for each number of clusters $K = 2, \ldots, 20$, we see that the new partitions obtained are better than the initial ones (first column). However the iterative relocation algorithm (with $N = 30$ random initializations) takes the lead from $K = 6$ clusters onwards.

2.3.5 Concluding remarks

This paper proposes an extension of an existing criterion for the clustering of numerical variables to the case of categorical data. The partitioning criterion measuring the cohesion of the clusters in the partition is based on correlation ratio between the

categorical variables of the cluster and a numerical latent variable. The latent variable of a cluster which optimizes the homogeneity criterion of a cluster is computed from MCA. Several algorithms for the clustering of categorical variables using the proposed partitioning criterion are described (iterative relocation algorithm, ascendant and divisive hierarchical clustering).

The results obtained with the proposed approach are illustrated and interpretated on a real data set. An empirical comparison of the different clustering approaches is also derived on this data set. We see on the proposed case study that the partitioning criterion may have several local optima. Then concerning the iterative relocation algorithm, the multiple random intitializations provides the best partitions in sense of proportion of explained cohesion. The divisive hierarchical clustering suffers from multiple local optima of the iterative relocation algorithm when splitting a cluster into two sub-clusters and then provides worse results than the ascendant hierarchial clustering or iterative relocation algorithm. Surprisingly the iterative relocation algorithm provides better results than the ascendant hierarchical clustering complemented by an iterative relocation of the variables. However one advantage of the hierarchical procedure is the easier interpretability of the results since it produces a hierarchy of nested partitions of the variables. The proposed algorithms have been implemented in $\mathcal{R}$ and source codes are available from the authors.

Furthermore a classical approach in data mining consists in carrying out a MCA and subsequently applying a clustering algorithm on the component scores of the objects, thereby using the first few components only. However DeSarbo et al. (1990), De Soete and Caroll (1994) and Vichi and Kiers (2001) warn against this approach, called "tandem analysis", because MCA may identify dimensions that do not necessarily contribute much to perceiving the clustering structure in the data and that, on the contrary, may obscure or mask the taxonomic information. Cluster analysis of variables is then an alternative technique as it makes it possible to organize the data into meaningful structures. Therefore the construction of latent variables may be more efficient that the classical MCA step.

One remaining point to study is the monotony of the proposed partitioning criterion. Another interesting aspect would be to compare the computational complexity of the different proposed algorithms. Concerning future prospects, the choice of the number of clusters with a bootstrap approach, consisting in generating multiple data replications of the data set and examining if the partition is stable, is currently under study. Research will also be undertaken on the treatment of missing values to avoid, as has been made in the presented real data application, deleting individuals who have returned questionnaires with the answers to some questions not completed.

Acknowledgements

The authors are very greatful to the public corporation Voies Navigables de France and the private firm Enform for providing the real data set.

Appendix : Proof of inequality (2.37)

We have

$$\begin{aligned}
\lambda_{l\cup m} &= \max_{\substack{\mathbf{u}\in\mathbb{R}^n \\ \mathbf{u}^t\mathbf{u}=1}} \{\mathbf{u}^t\widetilde{\mathbf{F}}_{l\cup m}\widetilde{\mathbf{F}}^t_{l\cup m}\mathbf{u}\} \\
&= \max_{\substack{\mathbf{u}\in\mathbb{R}^n \\ \mathbf{u}^t\mathbf{u}=1}} \{\mathbf{u}^t\widetilde{\mathbf{F}}_l\widetilde{\mathbf{F}}^t_l\mathbf{u} + \mathbf{u}^t\widetilde{\mathbf{F}}_m\widetilde{\mathbf{F}}^t_m\mathbf{u}\} \\
&\leq \max_{\substack{\mathbf{u}\in\mathbb{R}^n \\ \mathbf{u}^t\mathbf{u}=1}} \{\mathbf{u}^t\widetilde{\mathbf{F}}_l\widetilde{\mathbf{F}}^t_l\mathbf{u}\} + \max_{\substack{\mathbf{u}\in\mathbb{R}^n \\ \mathbf{u}^t\mathbf{u}=1}} \{\mathbf{u}^t\widetilde{\mathbf{F}}_m\widetilde{\mathbf{F}}^t_m\mathbf{u}\} \\
&= \lambda_l + \lambda_m,
\end{aligned}$$

where the definition of $\widetilde{\mathbf{F}}$ is given in (2.33) for a generic cluster.

References

Abdallah, H., Saporta, G. (1998). Classification d'un ensemble de variables qualitatives. *Revue de Statistique Appliquée*, **46**(4), 5-26.

Al-Kandari, N.M., Jolliffe, I.T. (2001). Variable selection and interpretation of covariance principal components. *Communications in Statistics - Simulation and Computation*, **30**, 339-354.

Chavent, M., Kuentz, V., Saracco, J. (2009). Rotation in Multiple Correspondence Analysis : a planar rotation iterative procedure. *Submitted paper*.

DeSarbo, W.S., Jedidi, K., Cool, K., Schendel, D. (1990). Simultaneous multidimensional unfolding and cluster analysis : An investigation of strategic groups. *Marketing Letters*, **2**, 129-146.

De Soete, G. and Carroll, J.D. (1994). K-means clustering in a low-dimensional Euclidean space. *In : Diday, E., et al. (Eds.), New Approaches in Classification and Data Analysis. Springer, Heidelberg*, 212-219.

Dhillon, I.S, Marcotte, E.M., Roshan, U. (2003). Diametrical clustering for identifying anti-correlated gene clusters. *Bioinformatics*, **19**(13), 1612-1619.

Guo, Q., Wu, W., Massart, D.L., Boucon, C., de Jong, S. (2002). Feature selection in principal component analysis of analytical data. *Chemometrics and Intelligent Laboratory Systems*, **61**, 123-132.

Hastie, T., Tibshirani, R., Eisen, M.B., Alizadeh, A., Levy, R., Staudt, L., Chan, W.C., Botstein, D., Brown, P. (2000). 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biology*, **1**(2), 1-21.

Jolliffe, I.T. (1972). Discarding variables in a principal component analysis. I. Artificial data. *Journal of the Royal Statistical Society. Series C. Applied Statistics* , **21**, 160-173.

Jolliffe, I.T. (2002). *Principal Component Analysis*, Second Edition, Springer-Verlag, New York.

Kaufman, L., Rousseeuw P.J. (1990), *Finding groups in data : an introduction to cluster analysis*, Wiley Series in probability and mathematical statistics, New York.

Krzanowski, W.J. (1987). Selection of variables to preserve multivariate data structure, using principal components. *Journal of the Royal Statistical Society. Series C. Applied Statistics*, **36**, 22-33.

Lerman, I.C. (1993). Likelihood linkage analysis (LLA) classification method : An example treated by hand. *Biochimie*, **75**,(5) 379-397.

McCabe, G.P. (1984). Principal variables. *Technometrics*, **26**(2), 137-144.

Mirkin, B. (2005). *Clustering for Data Mining. A Data Recovery Approach.*, Chapman & Hall, CRC Press, London, Boca Raton, FL.

Qannari, E.M., Vigneau, E., Courcoux PH. (1998). Une nouvelle distance entre variables. Application en classification. *Revue de Statistique Appliquée*, **46**(2), 21-32.

Plasse, M., Niang, N., Saporta, G., Villeminot, A., Leblond, L. (2007). Combined use of association rules mining and clustering methods to find relevant links between binary rare attributes in a large data set. *Computational Statistics and Data Analysis*, **52**, 596-613.

Soffritti, G. (1999). Hierarchical clustering of variables : a comparison among strategies of analysis. *Communications in Statistics - Simulation and Computation*, **28**(4), 977-999.

Stan, V., Saporta, G. (2005). Conjoint use of variables clustering and PLS structural equations modelling. *In PLS05, 2005. 4th International Symposium on PLS and related methods, Barcelone, 7-9 septembre 2005.*

Vichi, M., Kiers, H.A.L. (2001). Factorial k-means analysis for two way data. *Computational Statistics and Data Analysis*, **37**, 49-64.

Vichi, M., Saporta, G. (2009). Clustering and disjoint principal component analysis. *Computational Statistics and Data Analysis*, **53**, 3194-3208.

Vigneau, E., Qannari, E.M. (2003). Clustering of Variables Around Latent Components. *Communications in Statistics - Simulation and Computation*, **32**(4), 1131-1150.

Chapitre 3

Réduction de dimension via la méthode SIR (Sliced Inverse Regression)

Dans ce chapitre, la problématique de la réduction de dimension est abordée sous un angle de "modélisation" en considérant un modèle semiparamétrique de régression. Plus précisément, je me suis intéressée à la méthode de régression inverse par tranchage (SIR pour Sliced Inverse Regression). Tout d'abord je présente une synthèse des travaux réalisés dans ce cadre semiparamétrique. Puis les articles publiés qui ont découlé de ces recherches sont disponibles dans les deux dernières sections.

3.1 Synthèse des travaux

Je présente tout d'abord le modèle de référence de régression semiparamétrique dans le cadre duquel j'ai travaillé. Je décris la méthode SIR ainsi que la condition cruciale sur la distribution de la variable explicative. Je présente ensuite l'adaptation de SIR, que nous avons proposée sous le nom de "Cluster-based SIR", pour dépasser les restrictions pratiques de la condition fondamentale à l'utilisation de SIR. Enfin je décris l'approche "Bagging SIR" qui permet d'améliorer l'estimation de l'espace de réduction de dimension effective.

3.1.1 La méthode SIR

Un modèle paramétrique de régression décrit les relations entre une variable à expliquer $y \in \mathbb{R}$ et une variable explicative $\mathbf{x} \in \mathbb{R}^p$ $(p \geqslant 1)$, avec $\mathbb{E}(\mathbf{x}) = \mu$ et $\mathbb{V}(\mathbf{x}) = \Sigma$, de la forme :

$$y = f_\theta(\mathbf{x}) + \varepsilon,$$

où f_θ appartient à une famille de fonctions paramétrées par θ (vecteur de paramètres réels) et où ε est un terme d'erreur aléatoire. Dans ce type de modèle, l'objectif est d'estimer le paramètre θ. Les techniques d'estimation paramétrique (méthodes du maximum

de vraisemblance ou moindres carrés par exemple) sont efficaces lorsque la famille de f_θ est correctement spécifiée. Cependant dans de nombreuses applications, la mise en évidence d'un modèle paramétrique adéquat n'est pas simple. Aussi, quand un modèle paramétrique n'est pas disponible, les techniques nonparamétriques de régression apparaissent comme une alternative, offrant la flexibilité souhaitée dans la modélisation :

$$y = f(\mathbf{x}) + \varepsilon.$$

La régression fonctionnelle est basée sur un lissage local qui utilise les propriétés de continuité et de dérivabilité de la fonction de régression f. La qualité du lissage local en un point dépend alors de la présence de suffisamment de données dans le voisinage de ce point. Lorsque la variable est unidimensionnelle ($p = 1$), nous pouvons citer entre autres la méthode des noyaux ou les splines de lissage. Cependant lorsque la dimension de $\mathbf{x}$ devient grande, le nombre d'observations nécessaires pour le lissage local croît de manière exponentielle avec cette dimension. Ainsi, à moins de disposer d'un échantillon de taille gigantesque, ces méthodes nonparamétriques ne sont plus adaptées du fait du faible nombre de points dans la région d'intérêt.

Pour surmonter ce problème connu sous le nom de "fléau de la dimension", certaines méthodes de réduction de dimension supposent qu'on peut remplacer $\mathbf{x}$ par un vecteur de dimension K, strictement inférieure à p, sans perdre d'information sur la distribution conditionnelle de y sachant $\mathbf{x}$. Le modèle correspondant suppose que la dépendance entre les prédicteurs et la variable réponse y est décrite par des combinaisons linéaires des prédicteurs. Le modèle semiparamétrique sous-jacent s'écrit :

$$y = f(\mathbf{x}'\gamma_1, \ldots, \mathbf{x}'\gamma_K, \varepsilon), \tag{3.1}$$

où f est une fonction de lien inconnue, ε est une erreur aléatoire inconnue et indépendante de $\mathbf{x}$ et $\gamma_1, \ldots, \gamma_K$ sont K vecteurs inconnus de $\mathbb{R}^p$, supposés linéairement indépendants. Comme aucune condition sur la forme de f n'est imposée, les vecteurs γ_k ne sont pas identifiables. Seul l'espace engendré par ces vecteurs peut être estimé, c'est l'espace de réduction de dimension effective, appelé espace e.d.r. pour "effective dimension reduction", et noté E. Duan et Li (1987) et Li (1991) ont proposé une méthode d'estimation de cette base e.d.r. appelée Sliced Inverse Regression (SIR), traduit en français par régression inverse par tranchage. Ainsi nous nous intéressons à la partie paramétrique de ce modèle. Remarquons qu'après avoir obtenu une estimation de la partie paramétrique, nous pouvons utiliser les techniques usuelles de lissage pour estimer la fonction de lien f, techniques d'autant plus efficaces que la dimension K est faible.

Le principe des méthodes SIR est d'échanger le rôle de y et $\mathbf{x}$ et d'étudier les moments conditionnels de $\mathbf{x}$ sachant y. Ainsi nous nous ramenons à un problème plus simple de dimension inférieure qui consiste à régresser $\mathbf{x}$ sur y. Dans ce chapitre, nous nous intéressons à la méthode SIR-I, dénoté ici SIR, qui est basée sur le premier moment conditionnel inverse. Pour faciliter l'estimation de l'espérance conditionnelle inverse, un tranchage est réalisé sur la variable réponse y. Notons T cette transformation de y. Sous la condition de linéarité ci-dessous :

$$\mathbb{E}(\mathbf{x}'b|\mathbf{x}'\gamma_1, \ldots, \mathbf{x}'\gamma_K) \text{ est linéaire en } \mathbf{x}'\gamma_1, \ldots, \mathbf{x}'\gamma_K \text{ pour tout } b, \tag{3.2}$$

Li (1991) a montré la propriété géométrique suivante : la courbe de régression inverse centrée $\mathbb{E}(\mathbf{x}|T(y)) - \mathbb{E}(\mathbf{x})$ lorsque y varie, est contenue dans le sous-espace linéaire de $\mathbb{R}^p$ engendré par les vecteurs $\Sigma\gamma_1, \ldots, \Sigma\gamma_K$. Une conséquence directe est que la matrice de covariance $M = \mathbb{V}(\mathbb{E}(\mathbf{x}|T(y)))$ est dégénérée dans toute direction Σ-orthogonale aux $\gamma_k, k = 1, \ldots, K$. Il s'en suit que les vecteurs propres associés aux K valeurs propres non nulles de la matrice $\Sigma^{-1}M$ sont des directions e.d.r., c'est-à-dire appartiennent à l'espace e.d.r.

Il est cependant très difficile de vérifier a priori si la condition (3.2) est satisfaite étant donné qu'elle dépend des vraies directions e.d.r. inconnues. Remarquons que (3.2) est vérifiée lorsque $\mathbf{x}$ suit une distribution elliptique symétrique, condition plus forte en théorie mais plus facile à vérifier en pratique. Un cas particulier est la multinormalité de la variable explicative. Ces hypothèses entrainent des restrictions pratiques car la plupart des données collectées ne suivent pas une distribution elliptique. Notons que si les données n'ont pas encore été collectées, il est possible grâce à un plan d'expériences convenable d'atteindre l'objectif recherché. D'autres méthodes sont disponibles pour obtenir l'ellipticité de l'échantillon sélectionné. Lorsque les données ont déjà été collectées et que la dimension de $\mathbf{x}$ est faible, différentes techniques existent pour forcer les données à se comporter comme si elles étaient issues d'une distribution elliptique. La méthode la plus simple est celle du rééchantillonnage normal de Brillinger (1983). L'idée est de simuler un échantillon normal de même taille que le jeu de données original. Ces points seront les "attracteurs". Ensuite le principe est de sélectionner pour chaque attracteur le point le plus proche dans l'échantillon original. Remarquons que certains points des données seront sélectionnés plusieurs fois alors que d'autres ne le seront jamais. Ainsi les points sélectionnés sont supposés suivre une distribution normale (ou du moins plus normale que la distribution initiale). Pour le choix des paramètres de la distribution normale, Brillinger propose de prendre comme espérance et matrice de covariance les deux moments empiriques correspondants de l'échantillon initial. La seconde possibilité est la méthode proposée par Cook et Nachtsheim (1994). Elle permet de calculer des poids pour chaque point de l'échantillon afin de disposer de variables explicatives proches de l'ellipticité. L'idée est de construire une mesure discrète de probabilité qui va affecter des masses sur les points et telle qu'elle soit proche d'une distribution de probabilité elliptique cible. Le choix de cette distribution elliptique cible est basé sur l'ellipsoïde de volume minimal qui supprime une proportion des points de l'échantillon. En ce qui concerne la construction d'une mesure discrète qui est une approximation d'une distribution continue spécifiée, une solution naturelle est l'utilisation des mosaïques de Voronoi (ou cellules de Dirichlet).

Mais ces techniques présentent quelques inconvénients. Elles modifient l'échantillon original, l'une d'elles diminue le nombre d'observations. De plus, lorsque le nombre de variables explicatives devient très grand, il devient très difficile de forcer les données à être elliptiques avec des méthodes du type de celles présentées ci-dessus. Cependant Hall et Li (1993) ont montré avec un argument bayésien que la condition (3.2) est approximativement vérifiée pour des données en grande dimension. Ainsi pour des ensembles de données lorsque p est très grand, une application "aveugle" de SIR peut être utile et fournir une réponse à peu près correcte pour l'estimation des directions e.d.r.

Pour des covariables de dimension modérée, nous avons proposé une extension de SIR qui est utilisable lorsque la condition fondamentale de linéarité n'est pas vérifiée.

Remarque. Il sera pratique dans la description de l'approche Bagging SIR de considérer la version standardisée de SIR qui utilise $\mathbf{z} = \Sigma^{-1/2}(\mathbf{x} - \mu)$ de $\mathbf{x}$. On peut alors réécrire le modèle (3.1) sous la forme :

$$y = h(\mathbf{z}'\eta_1, \ldots, \mathbf{z}'\eta_K, \varepsilon),$$

où $\eta_k = \Sigma^{1/2}\gamma_k, k = 1 \ldots, K$ sont des directions e.d.r. standardisées. L'espace engendré par ces directions est alors appelé espace e.d.r. standardisé et noté E_s. La matrice de covariance de la courbe de régression inverse standardisée est $M_s = \mathbb{V}(\mathbb{E}(\mathbf{z}|T(y)))$. Les vecteurs propres associés aux K valeurs propres non nulles de M_s, dénotés $v_1, \ldots, v_K$, sont des directions e.d.r. standardisées. Ainsi l'espace e.d.r. peut être déduit de cet espace e.d.r. standardisé : il est engendré par les vecteurs $\Sigma^{-1/2}v_k, k = 1 \ldots, K$.

3.1.2 L'approche "Cluster-based SIR"

L'approche développée dans ce chapitre a donné lieu à l'article intitulé "Cluster-based Sliced Inverse Regression" écrit en collaboration avec Jérôme Saracco. Cet article va paraître dans *Journal of the Korean Statistical Society* et est disponible dans la section 3.2.

3.1.2.1 Principe

L'idée de cette approche que nous avons appelée "Cluster-based SIR", est inspirée du travail de Li et al. (2004) qui ont proposé une approche "Cluster-based Ordinary Least Squares" pour des modèles à un seul indice ($K = 1$). Le principe de Cluster-based SIR est de partitionner l'espace des prédicteurs de sorte que la condition de linéarité soit vérifiée dans chaque classe. En pratique, un algorithme des k-means est utilisé pour construire une partition de l'espace des prédicteurs en classes disjointes approximativement elliptiques. Au sein de chaque classe, puisque la condition de linéarité est vérifiée, nous estimons la direction e.d.r. avec la méthode SIR. Finalement nous combinons ces directions pour estimer l'espace e.d.r. du modèle (3.1), en utilisant l'ensemble des données disponibles. Nous considérons les modèles à un seul indice ($K = 1$) et les modèles multi-indices ($K > 1$). L'approche sur population puis celle sur échantillon sont décrites. La convergence en probabilité est obtenue et la distribution asymptotique de l'estimateur est donnée. Une étude sur simulations montre la bonne performance numérique de cette approche lorsque la condition de linéarité est violée. De plus, les résultats ne sont pas altérés lorsque cette condition est vérifiée. Une application sur des données réelles montrent la supériorité de Cluster-based SIR face à SIR en terme de prédiction de la variable réponse.

3.1.2.2 Modèle à un seul indice

Nous considérons un modèle à un seul indice :

$$y = f(\mathbf{x}'\gamma, \epsilon), \tag{3.3}$$

et nous nous intéressons à l'estimation d'une direction e.d.r. (c'est-à-dire colinéaire à γ).

Version sur population. Soit c un nombre de classes fixé et supposons que $\mathbf{x}$ est partitionné en c classes. Selon le schéma de partitionnement de $\mathbf{x}$, on obtient la partition $(\mathbf{x}^{(j)}, y^{(j)}), j = 1, \ldots, c$ de $(\mathbf{x}, y)$. On suppose que la condition de linéarité est vérifiée dans chaque classe :

(LC) Pour tout $j = 1, \ldots, c$, $\mathbb{E}(\mathbf{x}^{(j)\prime}b|\mathbf{x}^{(j)\prime}\gamma)$ est linéaire en $\mathbf{x}^{(j)\prime}\gamma$ pour tout b.

Dans chaque classe j, $T^{(j)}$ désigne le tranchage de $y^{(j)}$ en $H^{(j)}$ tranches fixes, $s_1^{(j)}, \ldots, s_{H^{(j)}}^{(j)}$, avec $H^{(j)} > 1$. Avec ce tranchage, $M_I^{(j)}$ s'écrit : $M_I^{(j)} = \sum_{h=1}^{H^{(j)}} p_h^{(j)}(m_h^{(j)} - \mu^{(j)})(m_h^{(j)} - \mu^{(j)})'$, où $p_h^{(j)} = P(y^{(j)} \in s_h^{(j)})$, $m_h^{(j)} = \mathbb{E}(\mathbf{x}^{(j)}|y^{(j)} \in s_h^{(j)})$ et $\mu^{(j)} = \mathbb{E}(\mathbf{x}^{(j)})$. Soit $\Sigma^{(j)} = \mathbb{V}(\mathbf{x}^{(j)})$. Le vecteur propre $g^{(j)}$ associé à la plus grande valeur propre de $(\Sigma^{(j)})^{-1}M_I^{(j)}$ est une direction e.d.r. Nous définissons la matrice $G = [g^{(1)}, \ldots, g^{(c)}]$ et nous notons g le premier vecteur propre de GG'. Nous avons obtenu le théorème suivant qui montre que g est une direction e.d.r. du modèle (3.3).

Théorème 1 *Sous la condition de linéarité (LC) et le modèle* (3.3)*, le vecteur propre g associé à la plus grande valeur propre de GG' est colinéaire à γ.*

La preuve de ce théorème est donné dans la section 3.2.

Version sur échantillon. Soit $S = \{(y_i, \mathbf{x}_i'), i = 1, \ldots, n\}$ un échantillon du modèle de référence (3.3). Une partition de ces observations est construite à l'aide d'un algorithme des k-means modifié. Notons tout d'abord que 100 initialisations au hasard sont générées et que la meilleure partition, au sens de la somme des distances des points aux centres des classes, est retenue. Par ailleurs, l'algorithme proposé évite les classes trop petites afin que le tranchage de SIR puisse avoir lieu. Si le nombre d'observations obtenues dans une classe est insuffisant, celle-ci est agrégée à la classe la plus proche, au sens du critère de Ward. Enfin, un nombre de classes maximum est préconisé en fonction de la taille de l'échantillon, du nombre de tranches voulu et du nombre minimal d'observations par tranche souhaité. Pour $j = 1, \ldots, c$, on a $S^{(j)} = \{(y_i^{(j)}, \mathbf{x}_i^{(j)\prime}), i = 1, \ldots, n^{(j)}\}$, avec $n^{(j)}$ le nombre d'observations dans $S^{(j)}$. Dans chaque classe, la moyenne et la covariance empirique sont respectivement données par $\overline{\mathbf{x}}^{(j)} = \frac{1}{n^{(j)}} \sum_{i=1}^{n^{(j)}} \mathbf{x}_i^{(j)}$ et $\widehat{\Sigma}^{(j)} = \frac{1}{n^{(j)}} \sum_{i=1}^{n^{(j)}} (\mathbf{x}_i^{(j)} - \overline{\mathbf{x}}^{(j)})(\mathbf{x}_i^{(j)} - \overline{\mathbf{x}}^{(j)})'$. La matrice $M_I^{(j)}$ est estimée par $\widehat{M_I}^{(j)} = \sum_{h=1}^{H^{(j)}} \hat{p}_h^{(j)}(\hat{m}_h^{(j)} - \overline{\mathbf{x}}^{(j)})(\hat{m}_h^{(j)} - \overline{\mathbf{x}}^{(j)})'$ avec $\hat{p}_h^{(j)} = \frac{1}{n^{(j)}} \sum_{i=1}^{n^{(j)}} \mathbb{I}_{[y_i \in s_h^{(j)}]}$ et $\hat{m}_h^{(j)} = \frac{1}{n^{(j)}\hat{p}_h^{(j)}} \sum_{i=1}^{n^{(j)}} \mathbf{x}_i^{(j)} \mathbb{I}_{[y_i \in s_h^{(j)}]}$, où $\mathbb{I}$ est la fonction indicatrice. Ainsi le vecteur

propre $\hat{g}^{(j)}$ associé à la plus grande valeur propre de $(\widehat{\Sigma}^{(j)})^{-1}\widehat{M_I}^{(j)}$ est la direction e.d.r. estimée dans la classe j. Nous construisons la matrice $\widehat{G} = [\hat{g}^{(1)}, \ldots, \hat{g}^{(c)}]$. Le plus grand vecteur propre $\hat{g}$ de $\widehat{G}\widehat{G}'$ est la direction e.d.r. estimée du modèle (3.3).

Théorie asymptotique. Les hypothèses suivantes sont nécessaires pour établir la convergence en probabilité et la normalité asymptotique de l'estimateur proposé.
(A1) S est un échantillon d'observations indépendantes issues de (3.1) ou (3.3).
(A2) $\mathbf{x}$ est partitionné en c classes fixes $\mathbf{x}^{(j)}, j = 1, \ldots, c$, tel que $\cup_{j=1}^{c} \mathcal{S}^{(j)} = \mathcal{S}$ et $\forall j \neq l, \mathcal{S}^{(j)} \cap \mathcal{S}^{(l)} = \emptyset$.
(A3) Le support de $y^{(j)}$ est partitionné en un nombre fixe $H^{(j)}$ de tranches telles que $p_h^{(j)} \neq 0, h = 1, \ldots, H^{(j)}$.
(A4) Pour $j = 1, \ldots, c$, $n^{(j)} \to \infty$ lorsque $n \to \infty$.

En utilisant le Théorème Central Limite, la méthode Delta ainsi que des résultats asymptotiques de SIR décrits dans Saracco (1997), nous montrons le théorème suivant.

Théorème 2 *Sous la condition de linéarité (LC) et les hypothèses* ***(A1)-(A4)****, on a :*

(a) $\hat{g} = g + O_p(n^{-1/2})$*, où* g *est une direction e.d.r. (colinéaire à* γ*).*

(b) $\sqrt{n}(\hat{g} - g) \longrightarrow_d U \sim \mathcal{N}(0, \Gamma_U)$*, où l'expression de* Γ_U *est donnée dans la section 3.2.*

La preuve de ce théorème est disponible dans la section 3.2.

Nombre optimal de classes. En pratique, le choix du nombre de classes est une étape cruciale dans l'approche proposée. Nous proposons de le choisir à l'aide du problème d'optimisation de type moindres carrés suivant :

$$\hat{c}^* = \arg \min_{c=1,\ldots,C} \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_{i,[c]})^2, \tag{3.4}$$

où $\hat{y}_{i,[c]} = \sum_{j=1}^{n} y_j \mathcal{K}((\mathbf{x}_i'\hat{g}_{[c]} - \mathbf{x}_j'\hat{g}_{[c]})/h_c) / \sum_{j=1}^{n} \mathcal{K}((\mathbf{x}_i'\hat{g}_{[c]} - \mathbf{x}_j'\hat{g}_{[c]})/h_c)$ est un estimateur à noyau de $\mathbb{E}(y|\mathbf{x}_i'\hat{g}_{[c]})$, pour lequel h_c est la largeur de fenêtre pour un partitionnement en c classes et $\mathcal{K}$ est un noyau, la densité de la loi normale centrée réduite par exemple.

3.1.2.3 Extension au cas d'un modèle multi-indices

Nous étendons l'approche proposée à des modèles multi-indices ($K > 1$). Le modèle correspondant est donné dans (3.1). Nous cherchons une base qui engendre l'espace e.d.r. $E = \text{Span}(\gamma_1, \ldots, \gamma_K)$. Seule l'approche sur population est décrite, la version sur échantillon est obtenue en remplaçant les moments théoriques par leur version empirique. La convergence en probabilité de l'estimateur est donnée. La normalité asymptotique peut être obtenue par une démonstration similaire au cas $K = 1$.

Version sur population. Comme pour le modèle à un seul indice, nous partitionnons l'espace prédicteur $\mathbf{x}$ en c classes. Selon ce partitionnement, nous obtenons la partition $(\mathbf{x}^{(j)}, y^{(j)}), j = 1, \ldots, c$ de $(\mathbf{x}, y)$. Nous supposons que la condition de linearité (LC*) est vérifiée dans chaque classe $j = 1, \ldots, c$:

(LC*) Pour $j = 1, \ldots, c$, $\mathbb{E}(\mathbf{x}^{(j)\prime} b | \mathbf{x}^{(j)\prime}\gamma_1, \ldots, \mathbf{x}^{(j)\prime}\gamma_K)$ est linéaire en $\mathbf{x}^{(j)\prime}\gamma_1, \ldots, \mathbf{x}^{(j)\prime}\gamma_K$ pour tout b.

Les vecteurs $g_1^{(j)}, \ldots, g_K^{(j)}$ associés aux K plus grandes valeurs propres de$(\Sigma^{(j)})^{-1} M_I^{(j)}$ sont des directions e.d.r. du modèle (3.1). Nous définissons la matrice $G^{(j)} = [g_1^{(j)}, \ldots, g_K^{(j)}]$ en juxtaposant en colonnes ces directions e.d.r., formant ainsi une base $\Sigma^{(j)}$-orthogonale de E. Avant de juxtaposer les matrices $G^{(j)}, j = 1, \ldots, c$ pour calquer la technique proposée dans le cas à un indice, une étape supplémentaire est nécessaire pour que ces matrices soient normalisées selon la même métrique. Les K plus grands vecteurs propres de la matrice $G^{(j)}G^{(j)\prime}$, dénotés $\tilde{g}_1^{(j)}, \ldots, \tilde{g}_K^{(j)}$, forment une base I_p-orthonormale de E. Ces vecteurs sont rangés dans la matrice $\tilde{G}^{(j)} = [\tilde{g}_1^{(j)}, \ldots, \tilde{g}_K^{(j)}]$. Nous combinons ces matrices $\tilde{G}^{(j)}$, $j = 1, \ldots, c$ dans la matrice $\mathbb{G}^{(c)} = [\tilde{G}^{(1)}, \ldots, \tilde{G}^{(c)}]$. Les K premiers vecteurs propres de $\mathbb{G}^{(c)}\mathbb{G}^{(c)\prime}$ sont notés $\tilde{\tilde{g}}_1, \ldots, \tilde{\tilde{g}}_K$. Le théorème ci-dessous prouve que ces vecteurs forment une base de E.

Théorème 3 *Sous la condition de linearité (LC*) et le modèle* (3.1)*, les vecteurs* $\tilde{\tilde{g}}_1, \ldots, \tilde{\tilde{b}}_K$ *forment une base* I_p*-orthogonale de l'espace e.d.r.*

La preuve de ce théorème est fourni dans la section 3.2.

Choix de la dimension K. Dans les applications, la dimension K est inconnue et doit être estimée à partir des données. Nous préconisons de la choisir avec les méthodes proposées pour SIR classique. Par exemple les méthodes de tests d'hypothèses sont basées sur la nullité des $(p - K)$ valeurs propres (Li, 1991, Schott, 1994 ou Barrios et Velilla, 2007, entre autres). Un autre type d'approche est basé sur une mesure de qualité entre le vrai espace e.d.r. et son estimation (voir par exemple Ferré, 1998 ou Liquet et Saracco, 2008).

3.1.2.4 Etude sur simulations

Une étude sur des données simulées à l'aide de modèles à un ou deux indices a montré le bon comportement de l'estimateur proposé. Lorsque la condition de linéarité est violée, l'estimateur construit avec Cluster-based SIR est meilleur que celui obtenu avec SIR. De plus lorsque les variables explicatives sont simulées à l'aide de lois normales (condition de linéarité vérifiée), la qualité obtenue avec Cluster-based SIR est toujours supérieure ou égale à celle de SIR. Ainsi même dans un cas favorable à SIR, la qualité de l'estimation n'est pas altérée avec l'étape supplémentaire de partitionnement. L'utilisation de Cluster-based SIR à la place de SIR produit donc des estimations de l'espace e.d.r. globalement meilleures. Le seul prix à payer est un léger surplus en temps de calculs.

3.1.2.5 Application sur de vraies données

L'approche proposée a été appliquée sur un jeu de données réelles décrit par exemple dans Camden (1989) ou Cook et Weisberg (1991). Les données décrivent les caractéristiques de $n = 201$ mollusques : une variable dépendante, la masse musculaire et une variable explicative de dimension $p = 4$ (longueur, largeur, hauteur, poids). Les nombreuses études réalisées précédemment sur ces données ont conduit à une structure uni-dimensionnelle ($K = 1$). Notre critère de minimisation nous a permis de choisir un nombre de classes $c = 5$ pour Cluster-based SIR. Pour ce choix de nombre de classes et de dimension, nous avons vérifié qu'il y a une structure pertinente dans le nuage de points $\{(y_i, \mathbf{x}_i'\hat{g}, i = 1, \ldots, n\}$, où $\hat{g}$ est la direction e.d.r. estimée avec Cluster-based SIR ($c = 5$). Les méthodes SIR et Cluster-based SIR sont alors comparées en fonction de leur capacité prédictive. Cluster-based SIR est plus performante que SIR. En particulier la moyenne absolue de l'erreur relative de reconstruction de la variable dépendante est diminuée de moitié avec Cluster-based SIR sur ce jeu de données.

3.1.2.6 Conclusion

Nous avons proposé une extension de SIR qui est utilisable lorsque la condition fondamentale n'est pas vérifiée. La normalité asymptotique de l'estimateur a été obtenue et le bon comportement de l'approche a été vérifié en simulations. Ainsi Cluster-based SIR est beaucoup moins sensible que SIR à la violation de cette hypothèse et produit globalement de meilleures estimations que SIR. L'approche a également été appliquée sur un jeu de données réelles et la supériorité de Cluster-based SIR par rapport à SIR a de nouveau été mise en évidence. Cependant une limitation de l'approche provient de l'utilisation d'un algorithme des k-means pour construire des classes elliptiques. En effet, si le jeu de données a une structure très particulière (forme enroulée, en "S", etc.), l'algorithme des k-means classique sera peu performant et une autre méthode de partitionnement serait préférable pour la construction de classes elliptiques.

3.1.3 Versions "Bagging" de SIR

L'approche proposée dans ce chapitre a donné lieu à l'article intitulé "Bagging versions of Sliced Inverse Regression" écrit en collaboration avec Benoît Liquet et Jérôme Saracco. Cet article va paraître dans *Communications in Statistics - Theory and Methods* et est disponible dans la section 3.3.

3.1.3.1 La méthode du bootstrap

Le bootstrap, proposé par Efron (1979), est une méthode de rééchantillonnage bien connue qui vise à obtenir de l'information sur les données dont on dispose en générant plusieurs versions du jeu de données original. Le lecteur peut se référer à Efron (1993) ou Shao et Tu (1995) pour une présentation complète de cette technique. Le bagging (voir par example Breiman, 1996 ou Bühlmann, 2004), acronyme pour "bootstrap" et

"aggregating", consiste à générer des réplications bootstrap du jeu de données, calculer la statistique d'intérêt dans chaque échantillon bootstrap et finalement agréger les différentes estimations. La façon d'agréger dépend de l'estimateur étudié, il s'agit souvent de la moyenne pour des estimateurs quantitatifs. Pour la prédiction d'une classe d'appartenance, on sélectionne celle qui est la plus fréquente.

3.1.3.2 Principe de l'approche Bagging SIR

L'idée de l'approche que nous avons appelée "Bagging SIR" est de générer plusieurs réplications bootstrap à partir des observations (par rééchantillonnage avec remise), puis de calculer un estimateur de type SIR dans chaque échantillon, et enfin d'agréger les différents estimateurs. Cette agrégation peut se faire de différentes façons conduisant ainsi à plusieurs versions de Bagging SIR. La théorie asymptotique est écrite sur une des versions, les autres peuvent être obtenues avec le même type de démonstration. Une étude sur simulations permet de mettre en parallèle les différentes versions de Bagging SIR et de les comparer à l'approche SIR classique. Les résultats numériques montrent que le gain de l'approche Bagging SIR est évident pour des modèles bruités ou des échantillons de faible taille.

3.1.3.3 Différentes versions de Bagging SIR

Nous considérons le modèle (3.1) et nous proposons quatre versions "Bagging" de SIR, qui diffèrent dans la façon d'agréger les matrices ou directions d'intérêt issues des réplications bootstrap.

Bagging-I. Soit B le nombre de réplications bootstrap et soit $S^{*(b)} = \{(y_i^{*(b)}, \mathbf{x}_i^{*(b)\prime}), i = 1, \ldots, n\}$, pour $b = 1, \ldots, B$, un échantillon bootstrap nonparamétrique. La moyenne et la covariance empiriques des $\mathbf{x}_i^{*(b)\prime}$ sont respectivement données par $\overline{\mathbf{x}}^{*(b)} = n^{-1}\sum_{i=1}^n \mathbf{x}_i^{*(b)}$ et $\widehat{\Sigma}^{*(b)} = n^{-1}\sum_{i=1}^n (\mathbf{x}_i^{*(b)} - \overline{\mathbf{x}}^{*(b)})(\mathbf{x}_i^{*(b)} - \overline{\mathbf{x}}^{*(b)})'$. La version standardisée des $\mathbf{x}_i^{*(b)}$ s'écrit $\mathbf{z}_i^{*(b)} = (\widehat{\Sigma}^{*(b)})^{-1/2}(\mathbf{x}_i^{*(b)} - \overline{\mathbf{x}}^{*(b)})$. Dans chaque échantillon bootstrap $S^{*(b)}$, soit $T^{(b)}$ le tranchage des $y_i^{*(b)}$ en H tranches fixes, $s_1^{(b)}, \ldots, s_H^{(b)}$. La matrice $\widehat{M}_s^{*(b)}$ est calculée de la manière suivante : $\widehat{M}_s^{*(b)} = \sum_{h=1}^H \hat{p}_h^{*(b)} (\hat{m}_{s,h}^{*(b)})(\hat{m}_{s,h}^{*(b)})'$, où $\hat{p}_h^{*(b)} = n^{-1}\sum_{i=1}^n \mathbb{I}_{[y_i^{*(b)} \in s_h^{(b)}]}$ et $\hat{m}_{s,h}^{*(b)} = \left(n\hat{p}_h^{*(b)}\right)^{-1} \sum_{i=1}^n \mathbf{z}_i^{*(b)} \mathbb{I}_{[y_i^{*(b)} \in s_h^{(b)}]}$. Nous calculons la moyenne de ces B matrices de covariance, $\widehat{M}_{s,B}^* = \sum_{b=1}^B \widehat{M}_s^{*(b)}/B$. Les vecteurs propres $\hat{g}_{s,k}^*, k = 1 \ldots, K$ associés aux K plus grandes valeurs propres de la matrice $\widehat{M}_{s,B}^*$ sont les estimateurs Bagging-I des directions e.d.r. standardisées. Enfin $\hat{g}_k^* = \widehat{\Sigma}^{-1/2}\hat{g}_{s,k}^*, k = 1 \ldots, K$ sont les directions e.d.r. estimées du modèle (3.1).

Bagging-II. Cette version peut être vue comme une version non standardisée de Bagging-I. Dans chaque échantillon bootstrap $S^{*(b)}$, nous calculons la matrice $\widehat{M}^{*(b)} = \sum_{h=1}^H \hat{p}_h^{*(b)} (\hat{m}_h^{*(b)} - \overline{\mathbf{x}}^{*(b)})(\hat{m}_h^{*(b)} - \overline{\mathbf{x}}^{*(b)})'$ où $\hat{p}_h^{*(b)} = n^{-1}\sum_{i=1}^n \mathbb{I}_{[y_i^{*(b)} \in s_h^{(b)}]}$ et $\hat{m}_h^{*(b)} =$

$\left(n\hat{p}_h^{*(b)}\right)^{-1} \sum_{i=1}^{n} \mathbf{x}_i^{*(b)} \mathbb{I}_{[y_i^{*(b)} \in s_h^{(b)}]}$. Nous calculons la moyenne de ces B matrices de covariance $\widehat{M}_B^* = \sum_{b=1}^{B} \widehat{M}^{*(b)}/B$. Les vecteurs propres $\hat{g}_k^*, k = 1 \ldots, K$ associés aux K plus grandes valeurs propres de $\hat{\Sigma}^{-1}\widehat{M}_B^*$ sont les directions e.d.r. estimées.

Dans un souci de clarté et de simplicité, les deux versions suivantes de Bagging SIR sont seulement décrites pour des modèles à un seul indice ($K = 1$). Ces versions s'étendent facilement au cas multi-indices (voir dans l'article sur l'approche Bagging SIR disponible dans la section 3.3).

Bagging-III. Dans chaque échantillon bootstrap $S^{*(b)}$, nous calculons la matrice $\widehat{M}_s^{*(b)}$ définie dans Bagging-I. Le vecteur propre $\hat{g}_s^{*(b)}$ associé à la plus grande valeur propre de la matrice $\widehat{M}_s^{*(b)}$ est la direction e.d.r. standardisée estimée dans le $b$ème échantillon bootstrap. Nous construisons $\widehat{G}_s^* = [\hat{g}_s^{*(1)}, \ldots, \hat{g}_s^{*(B)}]$. Nous avons montré dans l'article sur Cluster-based SIR que si nous considérons la matrice $G_s = [g_s, \ldots, g_s]$, alors le plus grand vecteur propre de $G_s G_s'$ est g_s. Nous considérons donc le plus grand vecteur propre $\hat{g}_s^*$ de la matrice $\widehat{G}_s^*(\widehat{G}_s^*)'$ comme direction e.d.r. standardisée estimée. Finalement la direction e.d.r. estimée est donnée par $\hat{g}^* = \hat{\Sigma}^{-1/2}\hat{g}_s^*$.

Bagging-IV. Cette version peut être considérée comme une version non standardisée de Bagging-III. Dans chaque réplication bootstrap $S^{*(b)}$, nous calculons la matrice $\widehat{M}^{*(b)}$ definie dans Bagging-II. Le vecteur propre $\hat{g}^{*(b)}$ associé à la plus grande valeur propre de la matrice $(\widehat{\Sigma}^{*(b)})^{-1}\widehat{M}^{*(b)}$ est la direction e.d.r. estimée dans l'échantillon $S^{*(b)}$. Nous construisons la matrice $\widehat{G}^* = [\hat{g}^{*(1)}, \ldots, \hat{g}^{*(B)}]$. Nous considérons alors le plus grand vecteur propre $\hat{g}^*$ de la matrice $\widehat{G}^*(\widehat{G}^*)'$ comme la direction e.d.r. estimée dans le modèle (3.1).

3.1.3.4 Théorie asymptotique

La théorie asymptotique est seulement écrite pour l'estimateur obtenu avec l'approche Bagging-I. Le même type de raisonnement permettrait d'obtenir la normalité asymptotique des trois autres estimateurs. La notation $Z_n \longrightarrow_d Z$ signifie que Z_n converge en distribution vers Z lorsque $n \to \infty$ et la notation $Z^* \longrightarrow_{D*} \widehat{Z}$ désigne la convergence bootstrap. Enfin la notation N^+ représente l'inverse généralisé de Moore-Penrose de la matrice carrée N.

En utilisant le Théorème Central Limite, la théorie asymptotique de SIR décrite dans Saracco (1997) et les résultats sur la convergence bootstrap présentés dans Barrios et Velilla (2007), nous démontrons le théorème suivant.

Théorème 4 *Sous la condition de linéarité et dans le cadre du modèle* (3.1)*, on a*

$$\sqrt{n}(vec(\widehat{M}_{s,B}^* - \widehat{M}_s)) \to_{D^*} \mathcal{N}(0, \Gamma_V),$$

où l'expression de Γ_V *est donnée dans la section 3.3.2.2.*

3.1.3.5 Choix de la dimension K

Nous préconisons de choisir la dimension du modèle à partir de l'échantillon original à l'aide des méthodes proposées pour SIR, par exemple l'approche de Barrios et Velilla (2007), utilisant le bootstrap, ou l'approche graphique introduite par Liquet et Saracco (2008) faisant également intervenir du bootstrap.

3.1.3.6 Simulations

Des simulations sur des modèles à un et deux indices ont montré le bon comportement des approches de type "Bagging".

Tout d'abord, un des paramètres important de l'approche Bagging SIR est le choix du nombre B d'échantillons bootstrap. L'impact du nombre de réplications sur la qualité de l'estimateur obtenu a été étudié sur $N = 100$ jeux de données simulés. Ainsi, un nombre de réplications $B = 100$ ou 200 semble suffisant. Au-delà, le gain apporté est quasiment nul. De plus pour toutes les valeurs de B (excepté 10), Bagging SIR est meilleur que SIR.

Ensuite les quatres versions de Bagging SIR ont été comparées entre elles et confrontées à l'approche SIR classique sur différents modèles avec $B = 200$. Les simulations montrent que les quatres versions produisent des estimateurs très similaires d'une qualité supérieure à SIR. Le gain évident de Bagging SIR est la diminution de la largeur des boxplots des mesures de qualité de l'estimation. L'effet de la taille de l'échantillon sur la performance de l'approche Bagging SIR a ensuite été considéré. Les bénéfices de Bagging SIR sont d'autant plus grands que l'échantillon est petit (plus petit que 100). La variance du terme d'erreur dans le modèle est ensuite augmentée pour ajouter du bruit : les résultats montrent alors que l'approche Bagging SIR est beaucoup moins sensible que SIR, le gain obtenu grâce à Bagging SIR face à SIR est d'autant plus grand que la relation structurelle sous-jacente entre la variable dépendante et les indices $\mathbf{x}'\gamma_k$, $k = 1, \ldots, K$, est relativement bruitée.

3.1.3.7 Conclusion

Nous avons proposé plusieurs versions "Bagging" de SIR. Des résultats asymptotiques sont donnés pour l'une d'entre elles. La bonne performance numérique de ces approches est mise en exergue sur des données simulées : Bagging SIR produit de meilleurs estimations de l'espace e.d.r. que la méthode SIR classique, en particulier pour de petites tailles d'échantillons ou des échantillons particulièrement bruités.

Remarquons par ailleurs que Bagging SIR est une méthode d'estimation très rapide car elle repose sur plusieurs utilisations de SIR qui sont elles-mêmes très rapides en temps de calculs. En pratique, lors de l'analyse de données réelles, nous recommandons d'utiliser SIR et Bagging SIR, de comparer attentivement les estimateurs obtenus et de privilégier l'approche Bagging SIR si une différence est observée.

3.1.4 Perspectives de ces travaux

La méthode SIR ne permet pas de diagnostiquer une dépendance symétrique : $y = f(\gamma'\mathbf{x}) + \epsilon$, où f est une fonction symétrique de l'argument $\gamma'\mathbf{x}$, et $\mathbf{x}'\gamma$ est symétrique par rapport à $\mu'\gamma$. Ceci vient du fait que SIR est seulement basé sur l'estimation de l'espérance conditionnelle de $\mathbf{x}$ sachant y. Pour surmonter cette difficulté, une approche consiste à explorer des moments conditionnels de $\mathbf{x}$ sachant y d'ordre supérieur. Cook et Weisberg (1991) se sont intéressés aux moments conditionnels d'ordre 2 dans leur méthode SAVE (Sliced Average Variance Estimation). Li (1991) a proposé la méthode SIR-II qui prend en compte la courbe de la covariance conditionnelle. Li (1991) a également introduit la méthode SIR-α qui conjugue l'information fournie par SIR-I et SIR-II. Le choix du paramètre α permet de spécifier une combinaison convenable des matrices d'intérêt des deux approches et ainsi de bénéficier des avantages des deux méthodes pour estimer au mieux la direction e.d.r. Ainsi il serait intéressant d'étendre les méthodes Cluster-based SIR et Bagging SIR à SIR-II et SIR-α.

Une autre perspective de travail consiste à adapter l'approche proposée au cas d'une variable y multidimensionnelle, c'est-à-dire à SIR multivarié.

D'autre part, nous nous intéressons aussi à l'estimation des directions e.d.r. en présence d'un prédicteur qualitatif z à l'aide d'un modèle semiparamétrique de régression $y = f(\mathbf{x}'\beta, z, \epsilon)$.

3.2 Cluster-based Sliced Inverse Regression

ARTICLE IN PRESS

Journal of the Korean Statistical Society ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Journal of the Korean Statistical Society

journal homepage: www.elsevier.com/locate/jkss

Cluster-based Sliced Inverse Regression

Vanessa Kuentz [a,b], Jérôme Saracco [a,b,c,*]

[a] Université de Bordeaux, IMB, UMR CNRS 5251, 351 Cours de la Libération, 33405 Talence Cedex, France
[b] INRIA Bordeaux Sud-Ouest, CQFD team, France
[c] Université Montesquieu - Bordeaux 4, GREThA, UMR CNRS 5113, Avenue Léon Duguit, 33608 Pessac Cedex, France

ARTICLE INFO

Article history:
Received 20 February 2009
Accepted 24 August 2009
Available online xxxx

AMS 2000 subject classifications:
primary 62G05
secondary 62D05

Keywords:
Sliced Inverse Regression (SIR)
Effective dimension reduction (e.d.r.) space
Clustering
Linearity condition

ABSTRACT

In the theory of sufficient dimension reduction, Sliced Inverse Regression (SIR) is a famous technique that enables us to reduce the dimensionality of regression problems. This semiparametric regression method aims at determining linear combinations of a p-dimensional explanatory variable $\mathbf{x}$ related to a response variable y. However it is based on a crucial condition on the marginal distribution of the predictor $\mathbf{x}$, often called the linearity condition. From a theoretical and practical point of view, this condition appears to be a limitation. Using an idea of Li, Cook, and Nachtsheim (2004) in the Ordinary Least Squares framework, we propose in this article to cluster the predictor space so that the linearity condition approximately holds in the different partitions. Then we apply SIR in each cluster and finally estimate the dimension reduction subspace by combining these individual estimates. We give asymptotic properties of the corresponding estimator. We show with a simulation study that the proposed approach, referred as cluster-based SIR, improves the estimation of the e.d.r. basis. We also propose an iterative implementation of cluster-based SIR and show in simulations that it increases the quality of the estimator. Finally the methodology is applied on the horse mussel data and the comparison of the prediction reached on test samples shows the superiority of cluster-based SIR over SIR.

1. Introduction

Parametric regression models are used to highlight the relationship between one response variable y and a p-dimensional explanatory variable $\mathbf{x} = (x^1, \ldots, x^p)'$, with $\mathbb{E}(\mathbf{x}) = \mu$ and $\mathbb{V}(\mathbf{x}) = \Sigma$, via for instance the following model:

$$y = f_\theta(\mathbf{x}) + \varepsilon,$$

where f_θ belongs to a family of parametric functions described by θ (vector of real parameters) and ε is a random error. The aim is the estimation of θ, which can be reached for example by maximum likelihood or least squares methods. These techniques work well if the family of f_θ is well specified. However this identification can turn out very difficult in some applications. Nonparametric regression models are then a possible solution. They offer a larger flexibility since they do not formulate any parametric assumption on the link function f. For instance, the model can be written:

$$y = f(\mathbf{x}) + \varepsilon.$$

These methods are essentially based on the property of continuity and derivability of the unknown regression function f. However they only provide good numerical results with low-dimensional explanatory variable. Indeed with high-dimensional problems, the number of observations needed to get information about the local behaviour of f becomes enormous. This is the well-known curse of dimensionality that can be challenged by dimension reduction models.

* Corresponding author at: Université de Bordeaux, IMB, UMR CNRS 5251, 351 Cours de la Libération, 33405 Talence Cedex, France.
E-mail addresses: vanessa.kuentz@math.u-bordeaux1.fr (V. Kuentz), jerome.saracco@math.u-bordeaux1.fr (J. Saracco).

1226-3192/$ – see front matter
doi:10.1016/j.jkss.2009.08.004

Please cite this article in press as: Kuentz, V., & Saracco, J. Cluster-based Sliced Inverse Regression. Journal of the Korean Statistical Society (2009), doi:10.1016/j.jkss.2009.08.004

Many dimension reduction tools assume that the features of $\mathbf{x}$ can be captured in a lower K-dimensional projection subspace (with $K < p$), such as Sliced Inverse Regression (SIR) methods introduced by Li (1991). They enable us to estimate a basis of this linear subspace. The corresponding model assumes that the dependency between the predictors and the response variable is described by linear combinations of the predictors. The underlying semiparametric model is written:

$$y = f(\mathbf{x}'\beta_1, \ldots, \mathbf{x}'\beta_K, \varepsilon), \tag{1}$$

where f is an unknown function, ε is an unknown random error independent of $\mathbf{x}$, and $\beta_1, \ldots, \beta_K$ are K unknown vectors in $\mathbb{R}^p$, assumed to be linearly independent. As none condition on the form of f is imposed, the vectors β_k are not identifiable. It is only possible to estimate the space spanned by these vectors, called the effective dimension reduction (e.d.r.) space, which will be denoted by E. When K is small ($K \ll p$), the goal of reduction theory is achieved and we can project the p-dimensional regressor $\mathbf{x}$ onto this K-dimensional space without loss of information on the feature of y given $\mathbf{x}$. Then it will be easier to study the relationship between $\mathbf{x}$ and y via a nonparametric estimation of the regression of y on the corresponding K-dimensional variable.

The basic principle of SIR methods is to reverse the role of y and $\mathbf{x}$ and to study the property of the conditional moments of $\mathbf{x}$ given y. In this paper, we will only focus on the SIR-I method (denoted by SIR hereafter) which is based on the first conditional moment. To facilitate the estimation of the inverse conditional mean, a slicing on the response variable y is realized. Let us denote by T this transformation of y. SIR is a method based on geometrical properties. Indeed, Li (1991) has shown that the centered inverse regression curve, $\mathbb{E}(\mathbf{x}|T(y)) - \mathbb{E}(\mathbf{x})$ as y varies, is contained in the linear subspace of $\mathbb{R}^p$ spanned by the vectors $\Sigma\beta_1, \ldots, \Sigma\beta_K$. A straightforward consequence is that the covariance matrix $M_I = \mathbb{V}(\mathbb{E}(\mathbf{x}|T(y)))$ is degenerated in any direction Σ-orthogonal to the β_k's. Therefore the eigenvectors associated with the nonnull K eigenvalues of the matrix $\Sigma^{-1}M_I$ are e.d.r. directions, that is are in the e.d.r. space. One important point in SIR theory is the underlying crucial linearity condition:

$$\mathbb{E}(\mathbf{x}'b|\mathbf{x}'\beta_1, \ldots, \mathbf{x}'\beta_K) \text{ is linear in } \mathbf{x}'\beta_1, \ldots, \mathbf{x}'\beta_K \quad \text{for any } b. \tag{2}$$

This condition is hard to verify in practice since it involves the unknown directions of the e.d.r. space. However it can be proved that (2) is verified when $\mathbf{x}$ follows an elliptically symmetric distribution, condition which is stronger in theory but easier to verify in practice. A special case is the multinormality of $\mathbf{x}$.

If the collected data set does not follow an elliptically distribution, solutions exist to force data to behave as if they were issued from such a distribution. For instance, if the dimension of $\mathbf{x}$ is small, two alternatives appear. The first one is the normal resampling of the data proposed by Brillinger (1983). The idea is to simulate a normal sample of same size as the original data. These simulated points are called "attractors". Then the principle is to select for each attractor the nearest point of the original sample. Note that some points of the original data set can be chosen several times while others will never be selected. Then the distribution of the selected points is more "normal" than the one of the original observations. The second solution is the re-weighting and trimming scheme of Cook and Nachtsheim (1994). The principle is to define a discrete probability measure that will assign weights to the observations and that is close to one elliptically symmetric distribution. This target elliptical distribution is based on Minimum Volume Ellipsoid (MVE), which enables us to trim a specified proportion of extreme points. Then the choice of one discrete distribution that approaches the target one is reached by Voronoi weights and Dirichlet cells. A problem with this technique is that it can severely reduce the sample size. Moreover these two methods are difficult to put into practice if $\mathbf{x}$ is high-dimensional. However Hall and Li (1993) showed with a bayesian argument that (2) approximately holds for high-dimensional data sets. So they argued that it is not a severe restriction in practice, implying that a blind application of these methods is not dangerous and will produce "good" estimations of the e.d.r. directions, when the dimension p is large.

In this paper we propose to cluster the predictor space, which will force the linearity condition to hold approximately in each cluster. The idea is inspired by the work of Li et al. (2004), who proposed a cluster-based Ordinary Least Squares (OLS) approach for single index models ($K = 1$). It consists in partitioning the predictor space with a k-means algorithm, evaluating the OLS estimate of each cluster and finally pooling them so as to provide an efficient estimation of the central mean subspace. In our approach, we also partition the predictor space into disjoint clusters with a k-means algorithm, which aims at constructing approximately elliptical clusters. Then we estimate the e.d.r directions in each cluster and combine them to produce an efficient estimation of the e.d.r space of model (1). The proposed approach will be referred in the rest of the paper as cluster-based SIR.

Note that for some special data structure cases, k-means can have relative poor performance on elliptical clusters: for instance, a mixture of two elongated components or the some special data structures such as Swiss roll or S-curve, see Cheung (2003) or Everitt, Landau, and Leese (2001) for details and discussion.

In Section 2, we consider the case of single index model, we describe the population and sample approaches of the cluster-based SIR. We show the convergence in probability and the asymptotic distribution of the corresponding estimator of the e.d.r. direction. We extend this approach to multiple indices models in Section 3. A simulation study is carried out in Section 4 in order to show the numerical performance of the approach and to compare it with SIR and cluster-based OLS. We also propose an iterative implementation and show with simulations that the quality of the estimated e.d.r. basis is improved. A real data application is reported in Section 5 to show the predictive performance of cluster-based SIR versus SIR. Finally concluding remarks are given in Section 6.

Please cite this article in press as: Kuentz, V., & Saracco, J. Cluster-based Sliced Inverse Regression. Journal of the Korean Statistical Society (2009), doi:10.1016/j.jkss.2009.08.004

2. Approach for single index model

We consider in this section single index model ($K = 1$). The corresponding model is:

$$y = f(\mathbf{x}'\beta, \varepsilon). \tag{3}$$

So we focus on the estimation of only one e.d.r. direction b colinear to β. The idea of the proposed approach is to partition the predictor space into a fixed number c of clusters. By doing that, the linearity condition will approximately hold in each cluster. For each one, we compute the e.d.r. direction with SIR. Finally we combine these directions to find the e.d.r direction of model (3) taking into account the whole space.

2.1. Population version

Let us consider a fixed number c of clusters and let us assume that $\mathbf{x}$ is partitioned into c clusters $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(c)}$. Accordingly to the partitioning scheme of $\mathbf{x}$, we get the partition $(\mathbf{x}^{(j)}, y^{(j)}), j = 1, \ldots, c$ of $(\mathbf{x}, y)$. Let us assume that the linearity condition holds in each cluster:

(LC) For $j = 1, \ldots, c$, $\quad \mathbb{E}(\mathbf{x}^{(j)\prime} b | \mathbf{x}^{(j)\prime}\beta)$ is linear in $\mathbf{x}^{(j)\prime}\beta$ for any b.

In each cluster j, let $T^{(j)}$ be the slicing of $y^{(j)}$ into $H^{(j)}$ fixed slices, $s_1^{(j)}, \ldots, s_{H^{(j)}}^{(j)}$, with $H^{(j)} > 1$. From this slicing, the matrix $M_I^{(j)}$ can be written as $M_I^{(j)} = \sum_{h=1}^{H^{(j)}} p_h^{(j)}(m_h^{(j)} - \mu^{(j)})(m_h^{(j)} - \mu^{(j)})'$, where $p_h^{(j)} = P(y^{(j)} \in s_h^{(j)})$, $m_h^{(j)} = \mathbb{E}(\mathbf{x}^{(j)} | y^{(j)} \in s_h^{(j)})$ and $\mu^{(j)} = \mathbb{E}(\mathbf{x}^{(j)})$. Let $\Sigma^{(j)} = \mathbb{V}(\mathbf{x}^{(j)})$. The eigenvector $b^{(j)}$ associated with the largest eigenvalue of the matrix $(\Sigma^{(j)})^{-1} M_I^{(j)}$ is an e.d.r. direction. We define the matrix $B = [b^{(1)}, \ldots, b^{(c)}]$ and we note b the first left singular vector of this matrix. Then Theorem 1 guarantees that this vector is an e.d.r. direction.

Theorem 1. *Assuming the linearity condition* (LC) *and model* (3)*, the major eigenvector b of the matrix BB' is colinear with β.*

Proof. For each $j = 1, \ldots, c$, $b^{(j)}$ is colinear with β, i.e. $b^{(j)} = \alpha_j \beta$, where α_j is a nonnull real. As $B = [\alpha_1\beta, \ldots, \alpha_c\beta]$, we have:

$$BB' = \sum_{j=1}^{c} \alpha_j^2 \beta\beta' = \|\alpha\|^2 \beta\beta',$$

where $\alpha = (\alpha_1, \ldots, \alpha_c)'$ and $\|.\|$ is the norm associated to usual scalar product. Then the eigenvector b associated with the strictly positive eigenvalue of BB' is colinear with β. □

2.2. Sample version

Let $S = \{(y_i, \mathbf{x}_i'), i = 1, \ldots, n\}$ be a sample from the reference model (3). We partition these observations into c clusters using a k-means approach. Note that one hundred initial random sets are chosen and the best partitioning is retained, i.e. the one that provides the minimum sum of squares from points to the assigned cluster centers. By this way, it stabilizes the clustering step and then cluster-based SIR approach. Moreover our k-means algorithm is constrained to avoid sparse clusters. We set the minimum number of points in a slice at $n_{h,\min}$, which implies that the minimum number of observations in the jth cluster is $n_{\min}^{(j)} = n_{h,\min} \times H^{(j)}$. If one cluster obtained with classical k-means contains less than $n_{\min}^{(j)}$ observations, it is merged with the nearest cluster, in sense of Ward criterion. Finally we advise a maximum number of clusters for cluster-based SIR defined as $C_{\max}^n = \frac{n}{H \times n_{h,\min}}$. So for $j = 1, \ldots, c$, we get samples $S^{(j)} = \{(y_i^{(j)}, \mathbf{x}_i^{(j)\prime}), i = 1, \ldots, n^{(j)}\}$. In each cluster, the empirical mean and covariance matrix of the $\mathbf{x}_i$'s are respectively given by $\overline{\mathbf{x}}^{(j)} = \frac{1}{n^{(j)}} \sum_{i=1}^{n^{(j)}} \mathbf{x}_i^{(j)}$ and $\widehat{\Sigma}^{(j)} = \frac{1}{n^{(j)}} \sum_{i=1}^{n^{(j)}} (\mathbf{x}_i^{(j)} - \overline{\mathbf{x}}^{(j)})(\mathbf{x}_i^{(j)} - \overline{\mathbf{x}}^{(j)})'$. The matrix $M_I^{(j)}$ is estimated by $\widehat{M_I}^{(j)} = \sum_{h=1}^{H^{(j)}} \hat{p}_h^{(j)} (\hat{m}_h^{(j)} - \overline{\mathbf{x}}^{(j)})(\hat{m}_h^{(j)} - \overline{\mathbf{x}}^{(j)})'$ with $\hat{p}_h^{(j)} = \frac{1}{n^{(j)}} \sum_{i=1}^{n^{(j)}} \mathbb{I}_{[y_i \in s_h^{(j)}]}$ and $\hat{m}_h^{(j)} = \frac{1}{n^{(j)} \hat{p}_h^{(j)}} \sum_{i=1}^{n^{(j)}} \mathbf{x}_i^{(j)} \mathbb{I}_{[y_i \in s_h^{(j)}]}$, where the notation $\mathbb{I}$ designates the indicator function. Then the eigenvector $\hat{b}^{(j)}$ associated with the largest eigenvalue of $(\widehat{\Sigma}^{(j)})^{-1} \widehat{M_I}^{(j)}$ is the estimated e.d.r. direction in the jth cluster. We construct the matrix $\widehat{B} = [\hat{b}^{(1)}, \ldots, \hat{b}^{(c)}]$. The major eigenvector $\hat{b}$ of the matrix $\widehat{B}\widehat{B}'$ is then the e.d.r. estimated direction in model (3).

2.3. Asymptotic theory

In what follows, the notation $Z_n \to_d Z$ means that Z_n converges in distribution to Z as $n \to \infty$. The assumptions that are necessary to state our results are gathered below for easy reference.

(A1) The sample S is a sample of independent observations from the single index model (3) or the multiple indices model (1).
(A2) $\mathbf{x}$ is partitioned into c fixed clusters $\mathbf{x}^{(j)}$, $j = 1, \ldots, c$, such that $\cup_{j=1}^{c} \mathcal{S}^{(j)} = \mathcal{S}$ and $\forall j \neq l$, $\mathcal{S}^{(j)} \cap \mathcal{S}^{(l)} = \emptyset$.

Please cite this article in press as: Kuentz, V., & Saracco, J. Cluster-based Sliced Inverse Regression. Journal of the Korean Statistical Society (2009), doi:10.1016/j.jkss.2009.08.004

(A3) The support of $y^{(j)}$ is partitioned into a fixed number $H^{(j)}$ of slices such that $p_h^{(j)} \neq 0$, $h = 1, \ldots, H^{(j)}$.
(A4) For $j = 1, \ldots, c$, $n^{(j)} \to \infty$ as $n \to \infty$.

Comment on (A4). With the proposed k-means step avoiding sparse clusters, we have for each cluster $j = 1, \ldots, c$, $n_h^{(j)} \geq n_{h,\min}$ and then $n^{(j)} \geq n_{\min}^{(j)} = H^{(j)} \times n_{h,\min}$, where $n_{h,\min} = \frac{n}{C_{\max}^n \times H^{(j)}}$. In order to get $n^{(j)} \to \infty$ as $n \to \infty$, we can choose for instance $C_{\max}^n = O\left((n/H^{(j)})^{1/2}\right)$.

We show in Theorem 2 the convergence in probability of the cluster-based SIR estimator and give its asymptotic distribution in Theorem 3.

Theorem 2. *Under the linearity condition* (LC) *and the assumptions* (A1)–(A4), *we have* $\hat{b} = b + O_p(n^{-1/2})$, *where b is an e.d.r. direction (colinear with* β*).*

Proof. Li (1991) has shown for SIR that the estimated e.d.r. direction converges to an e.d.r. direction at root n rate. So under the assumptions of the theorem, for each $\mathcal{S}^{(j)}$, $j = 1, \ldots, c$, we have

$$\hat{b}^{(j)} = b^{(j)} + O_p(n^{-1/2}).$$

Then we get $\widehat{B} = B + O_p(n^{-1/2})$ and $\widehat{B}\widehat{B}' = BB' + O_p(n^{-1/2})$. Thus the major eigenvector of $\widehat{B}\widehat{B}'$ converges to the corresponding one of BB' at the same rate: $\hat{b} = b + O_p(n^{-1/2})$. From Theorem 1, b is colinear with β. So the estimated e.d.r. direction obtained with cluster-based SIR converges to an e.d.r. direction at root n rate. □

Theorem 3. *Under the linearity condition* (LC) *and the assumptions* (A1)–(A4), *we have:*

$$\sqrt{n}(\hat{b} - b) \longrightarrow_d U \sim \mathcal{N}(0, \Gamma_U),$$

where the expression of Γ_U *is given in* (18).

The proof of Theorem 3 is given in the Appendix.

2.4. Optimal number of clusters

In practice, a crucial step in the proposed method is the choice of the number c of clusters for the partitioning of the predictor space. The choice of an optimal number c^* of clusters can be defined through the following optimization problem:

$$c^* = \arg \min_{c=1,\ldots,C} \mathbb{E}((y - \mathbb{E}(y|\mathbf{x}'\hat{b}_{[c]}))^2), \tag{4}$$

where $\hat{b}_{[c]}$ denotes the estimator of the e.d.r. direction when the number of clusters is c.
From a practical point of view, we consider an empirical smoothed version of this minimization problem:

$$\hat{c}^* = \arg \min_{c=1,\ldots,C} \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_{i,[c]})^2, \tag{5}$$

where $\hat{y}_{i,[c]} = \sum_{j=1}^n y_j \mathcal{K}((\mathbf{x}_i'\hat{b}_{[c]} - \mathbf{x}_j'\hat{b}_{[c]})/h_c) / \sum_{j=1}^n \mathcal{K}((\mathbf{x}_i'\hat{b}_{[c]} - \mathbf{x}_j'\hat{b}_{[c]})/h_c)$ is a kernel estimation of $\mathbb{E}(y|\mathbf{x}_i'\hat{b}_{[c]})$, for which h_c is the bandwidth parameter for a partitioning into c clusters and $\mathcal{K}$ is a kernel (the density of the standard univariate normal distribution for instance). The bandwidth parameters h_c, $c = 1, \ldots, C$, can be chosen by cross validation.

3. Extension to multiple indices model

In this section, we extend the proposed approach to multiple indices model ($K > 1$). The corresponding model is given in (1). We search for a basis that spans the e.d.r. space $E = \text{Span}(\beta_1, \ldots, \beta_K)$.

3.1. Population version

As for the single index model, we partition the predictor space $\mathbf{x}$ into c clusters. We get the partitions $(\mathbf{x}^{(j)}, y^{(j)})$, $j = 1, \ldots, c$. For each cluster, we seek with SIR a basis of the e.d.r. space. Let us assume that the following linearity condition (LC*) holds:

(LC*) For $j = 1, \ldots, c$, $\quad \mathbb{E}(\mathbf{x}^{(j)\prime} b | \mathbf{x}^{(j)\prime}\beta_1, \ldots, \mathbf{x}^{(j)\prime}\beta_K)$ is linear in $\mathbf{x}^{(j)\prime}\beta_1, \ldots, \mathbf{x}^{(j)\prime}\beta_K$ for any b.

The eigenvectors $b_1^{(j)}, \ldots, b_K^{(j)}$ associated with the largest K eigenvalues of the matrix $(\Sigma^{(j)})^{-1} M_I^{(j)}$ are e.d.r. directions, where matrices $\Sigma^{(j)}$ and $M_I^{(j)}$ have been defined in Section 2. We define the matrix $B^{(j)} = [b_1^{(j)}, \ldots, b_K^{(j)}]$ containing these e.d.r. directions, which form a $\Sigma^{(j)}$-orthogonal basis of E. Then the first K eigenvectors of the matrix $B^{(j)}B^{(j)'}$, denoted by $\tilde{b}_1^{(j)}, \ldots, \tilde{b}_K^{(j)}$, form an I_p-orthonormal basis of E. We store these vectors in the matrix $\tilde{B}^{(j)} = [\tilde{b}_1^{(j)}, \ldots, \tilde{b}_K^{(j)}]$. We can now pool the matrices $\tilde{B}^{(j)}$ in the matrix $\mathbb{B}^{(c)} = [\tilde{B}^{(1)}, \ldots, \tilde{B}^{(c)}]$. The first K eigenvectors of the matrix $\mathbb{B}^{(c)}\mathbb{B}^{(c)'}$ are denoted by $\tilde{\tilde{b}}_1, \ldots, \tilde{\tilde{b}}_K$.

Please cite this article in press as: Kuentz, V., & Saracco, J. Cluster-based Sliced Inverse Regression. Journal of the Korean Statistical Society (2009), doi:10.1016/j.jkss.2009.08.004

Theorem 4. *Assuming the linearity condition* (LC*) *and model* (1)*, the vectors* $\tilde{\tilde{b}}_1, \ldots, \tilde{\tilde{b}}_K$ *form an* I_p*-orthogonal basis of the e.d.r. space* E*.*

Proof. Since $\tilde{b}_1^{(j)}, \ldots, \tilde{b}_K^{(j)}$ form an I_p-orthonormal basis of E, we have $\mathrm{Span}(\mathbb{B}^{(c)}) = E$. Then the eigenvectors associated with the K largest eigenvalues of $\mathbb{B}^{(c)}\mathbb{B}^{(c)'}$ form an I_p-orthonormal basis of E. □

3.2. Sample version

As for the single index model, we estimate in each cluster a basis of the e.d.r space: the first K eigenvectors of the matrix $(\widehat{\Sigma}^{(j)})^{-1}\widehat{M}_I^{(j)}$ defined in Section 2. These vectors form a $\widehat{\Sigma}^{(j)}$-orthogonal basis of the estimated e.d.r. space. We store them in the matrix $\widehat{B}^{(j)} = [\hat{b}_1^{(j)}, \ldots, \hat{b}_K^{(j)}]$. Then the first K eigenvectors of the matrix $\widehat{B}^{(j)}\widehat{B}^{(j)'}$, denoted by $\hat{\tilde{b}}_1^{(j)}, \ldots, \hat{\tilde{b}}_K^{(j)}$, form an I_p-orthogonal basis of the estimated e.d.r. space. We store them in the matrix $\hat{\tilde{B}}^{(j)} = [\hat{\tilde{b}}_1^{(j)}, \ldots, \hat{\tilde{b}}_K^{(j)}]$. Let $\hat{\mathbb{B}}^{(c)} = [\hat{\tilde{B}}^{(1)}, \ldots, \hat{\tilde{B}}^{(c)}]$. Finally the first K eigenvectors of the matrix $\hat{\mathbb{B}}^{(c)}\hat{\mathbb{B}}^{(c)'}$, denoted by $\hat{\hat{\tilde{b}}}_1, \ldots, \hat{\hat{\tilde{b}}}_K$, form an I_p-basis of the estimated e.d.r. space.

3.3. Asymptotic theory

Under the linearity condition (LC*) and the assumptions (A1)–(A4), SIR theory provides $\widehat{B}^{(j)} = B^{(j)} + O_p(n^{-1/2})$. Then the first K eigenvectors of the matrix $\widehat{B}^{(j)}\widehat{B}^{(j)'}$ converge at same rate to the corresponding ones of $B^{(j)}B^{(j)'}$. Analogously $\hat{\mathbb{B}}^{(c)} = \mathbb{B}^{(c)} + O_p(n^{-1/2})$ and $\hat{\mathbb{B}}^{(c)}\hat{\mathbb{B}}^{(c)\prime} = \mathbb{B}^{(c)}\mathbb{B}^{(c)\prime} + O_p(n^{-1/2})$. Finally $\hat{\hat{\tilde{b}}}_k = \tilde{\tilde{b}}_k + O_p(n^{-1/2})$, $k = 1, \ldots, K$, then the estimated e.d.r. basis converges to an e.d.r. basis at root n rate.
As for the single index model, using Delta method and asymptotic results of Saracco (1997) and Tyler (1981), the asymptotic normality of the eigenprojector onto the estimated e.d.r. space can be obtained, as well as the asymptotic distribution of the estimated e.d.r. direction, associated with eigenvalues assumed to be different.

3.4. Choice of dimension K and number of clusters

Until now we have supposed that the dimension K of the reduction model is known. However in most applications this dimension is a priori unknown and hence must be estimated from the data. From a practical point of view, we recommend to choose the dimension K using classical SIR. Several approaches have been proposed in the literature for SIR. The first type of approaches are hypothesis tests based on the nullity of the last $(p-K)$ eigenvalues, see Li (1991), Schott (1994) or Barrios and Velilla (2007). Another approach relies on a quality measure based on the square trace correlation between the true e.d.r. space and its estimate, see for instance Ferré (1998) or Liquet and Saracco (2008).

With this choice $\hat{K}$ of K, we can determine the optimal number of clusters using the kernel estimator method proposed in (5). From a theoretical point of view, when $\hat{K} > 1$ the kernel $\mathcal{K}$ can be replaced by a multidimensional one. From a practical point of view, as soon as $\hat{K} > 2$, this method is no more appropriate due to the curse of dimensionality. However in real applications, this dimension is seldom larger than 2.

Finally for the chosen couple of parameters $(\hat{K}, \hat{c}^*)$, it is important to check if there is a relevant structure in the scatter plot $\{(y_i, \mathbf{x}_i'\hat{\hat{\tilde{b}}}_1, \ldots, \mathbf{x}_i'\hat{\hat{\tilde{b}}}_{\hat{K}}), i = 1, \ldots, n\}$ and to verify that there is no structure in the scatter plot $\{(y_i, \mathbf{x}_i'\hat{\hat{\tilde{b}}}_{\hat{K}+1}), i = 1, \ldots, n\}$ if we assume that the dimensions is $\hat{K} + 1$. This methodology is that used for the real data application in Section 5.

4. Simulation study

A simulation study is carried out to evaluate the numerical performance of the proposed method. First we recall the definition of the efficiency measure. Then, in a first stage we consider a single index model and compare the results obtained with cluster-based SIR with those provided by classical SIR and cluster-based OLS (Li et al., 2004). In a second stage, we compare the results obtained with SIR and cluster-based SIR on a multiple indices model (with $K = 2$). Finally we present an iterative version of the cluster-based SIR approach.

In the simulation study, we set $H = H^{(j)} = 4$ for SIR and cluster-based SIR. Moreover according to comment on (A4) we used for instance for sample size $n = 200$ (resp. 500), $C^n_{\max} = 7$ (resp. 12) and $n_{h,\min} = 6$ (resp. 11) in the selection of the number of clusters and in our modified k-means approach.

4.1. Efficiency measure

Let $\check{b}_1, \ldots, \check{b}_K$ be the K estimated e.d.r. directions. We note $\check{B} = [\check{b}_1, \ldots, \check{b}_K]$ and $\check{E} = \mathrm{Span}(\check{B})$ the linear subspace spanned by the $\check{b}_k$'s. Let $B = [\beta_1, \ldots, \beta_K]$ be the matrix of the true directions and let $E = \mathrm{Span}(B)$. Let P_E (resp. $P_{\check{E}}$) be the I_p-orthogonal projector onto E (resp. $\check{E}$) defined as follows: $P_E = B(B'B)^{-1}B'$ and $P_{\check{E}} = \check{B}(\check{B}'\check{B})^{-1}\check{B}'$. Since with cluster-based

Please cite this article in press as: Kuentz, V., & Saracco, J. Cluster-based Sliced Inverse Regression. Journal of the Korean Statistical Society (2009), doi:10.1016/j.jkss.2009.08.004

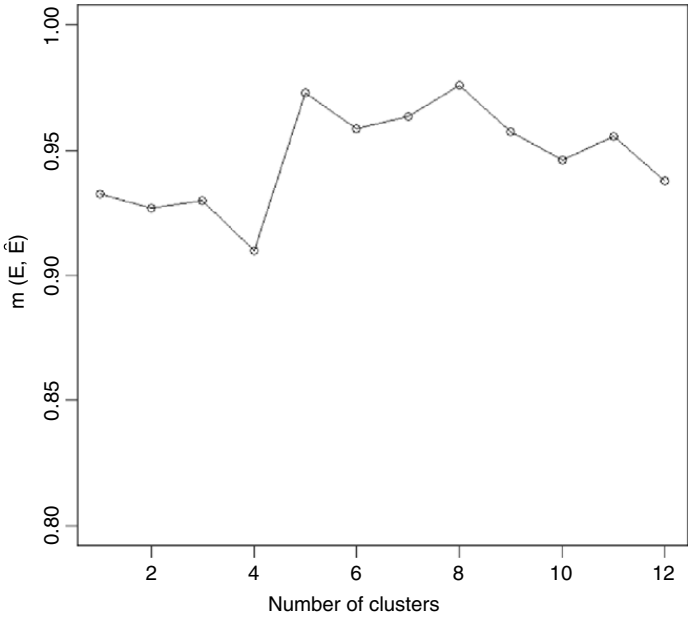

Fig. 1. Efficiency measure (6) of the estimation obtained with SIR ($c = 1$) and cluster-based SIR ($c > 1$) for model (7) with $n = 500$ and $\theta = 0$.

SIR approach we construct an I_p-orthonormal basis of E (that is $\check{B}'\check{B} = I_K$) and if we choose the β_k's such that $B'B = I_K$, the expression of the projectors reduces to: $P_E = BB'$ and $P_{\check{E}} = \check{B}\check{B}'$.
The quality of the estimate $\check{E}$ of E is measured by:

$$m(E, \check{E}) = \text{Trace}(P_E P_{\check{E}})/K. \tag{6}$$

This measure belongs to [0, 1] with $m(E, \check{E}) = 0$ if $\check{E} \perp E$ and $m(E, \check{E}) = 1$ if $\check{E} = E$. Therefore the closer this value is to one, the better is the estimation. When $K = 1$ (single index model), this measure is the squared cosine of the angle formed by the vectors β and $\check{b}$.

4.2. Single index model

First we define the simulated model, then we describe our approach on a sample when the linearity condition is not verified. Finally we generate multiple data replications for which the linearity condition may be seriously violated or not.

4.2.1. Simulated model

We consider the following regression model:

$$y = \exp(x_1 - x_2) + \varepsilon, \tag{7}$$

with $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8)'$, where $x_j \sim (1 - \theta) \times \text{Exp}(1) + \theta \times \mathcal{N}(0, 1)$ and $\varepsilon \sim \mathcal{N}(0, 0.5^2)$. The variables x_j are mutually independent and the error term ε is independent of $\mathbf{x}$. In this model, the true normalized direction is $\beta = (\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}, 0, 0, 0, 0, 0, 0)'$. In our simulations, the parameter θ will belong to [0, 1]. The value 0 corresponds to non-elliptical distribution (in this case the linearity condition is not verified) and 1 to multinormal distribution.

4.2.2. Single data replication

We exhibit a sample of size $n = 500$ of model (7) with $\theta = 0$. Fig. 1 shows the evolution of the quality criterion (6) as the number of clusters varies between 1 and 12. The case with one cluster matches with classical SIR. The maximum is reached at 8 clusters with an efficiency measure of 0.98, against 0.93 for classical SIR. The estimation of the e.d.r. direction is then improved by the use of cluster-based SIR. Fig. 2 shows the evolution of the empirical criterion (5) as the number of clusters increases. With this measure, we would choose the optimal number of clusters in sense of the quality measure (6). However we have observed in the simulation study that we do not always choose the same number of clusters as with (6). But the chosen number of clusters always provides a measure $m(E, \widehat{E})$ equal or very close to the maximum. Clearly criterion (5) is the only one appropriate criterion from a practical point of view, since criterion (6) requires knowledge of the true basis of the e.d.r. space, which is unknown in real applications. By contrast, criterion (5) is always estimable in practice but it involves a kernel estimation which is computationally expensive because it needs to introduce a tuning parameter (the bandwidth) chosen by cross validation. To reduce the computational cost of the rest of the simulation study (as we know the true e.d.r. direction) the optimal number of clusters will be chosen with the efficiency measure (6).

Please cite this article in press as: Kuentz, V., & Saracco, J. Cluster-based Sliced Inverse Regression. Journal of the Korean Statistical Society (2009), doi:10.1016/j.jkss.2009.08.004

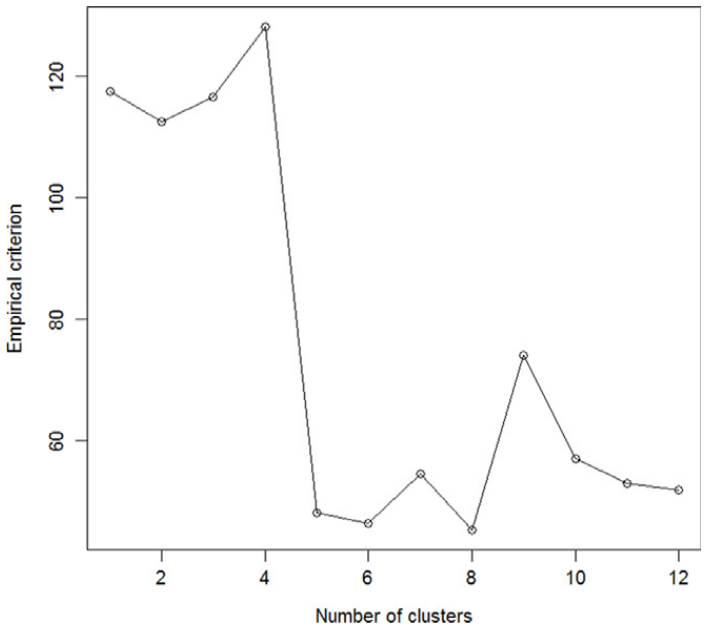

Fig. 2. Empirical criterion (5) obtained with SIR ($c = 1$) and cluster-based SIR ($c > 1$) for model (7) with $n = 500$ and $\theta = 0$.

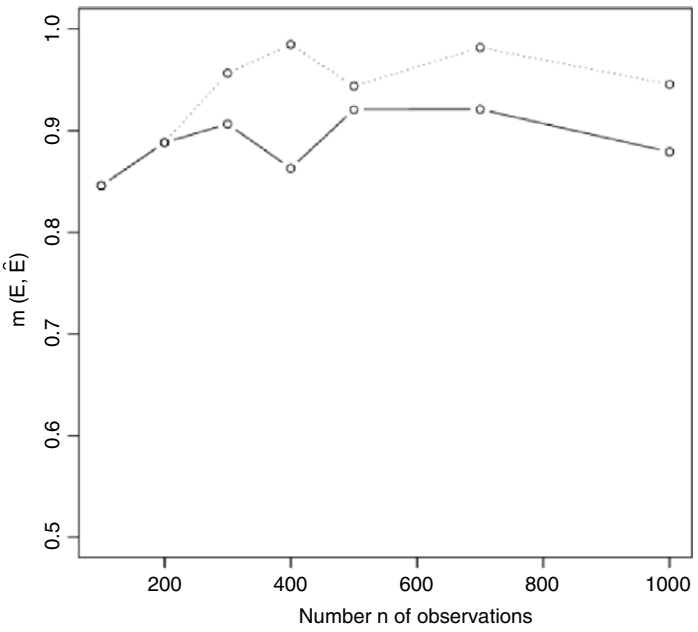

Fig. 3. Quality measure for model (7) with $\theta = 0$ as the number of observations increases (solid line: SIR, dotted line: cluster-based SIR).

Fig. 3 shows the quality of the estimations obtained respectively with SIR and cluster-based SIR for model (7) when $\theta = 0$ and the sample size n takes values 100, 200, 300, 400, 500, 700 and 1000. For each sample size, we have determined the optimal number of clusters. Not surprisingly, we observe that the performances of the two methods globally tend to increase as the number of observations becomes higher. We also observe that in this simulation cluster-based SIR is always better than SIR, except for $n = 100$ and $n = 200$ observations, where the two methods are similar. Indeed for these small sample sizes, cluster-based SIR does not always improve classical SIR because of the small number of observations in some clusters. The cluster-based SIR approach chooses in this case an optimal number of cluster equal to 1, corresponding then to classical SIR. Remark that we have not generated samples with a smaller size than 100, because the use of SIR is then inappropriate. In this case, one should replace the slicing step with pooled-slicing one, see Aragon and Saracco (1997) or Saracco (2001) for details.

4.2.3. Multiple data replications

In this section, we compare SIR and cluster-based SIR on $N = 100$ data replications of model (7). The parameter θ will belong to the set {0, 0.2, 0.4, 0.6, 0.8, 1} and the number n of observations will be 100, 200, 500 and 1000. For each

Please cite this article in press as: Kuentz, V., & Saracco, J. Cluster-based Sliced Inverse Regression. Journal of the Korean Statistical Society (2009), doi:10.1016/j.jkss.2009.08.004

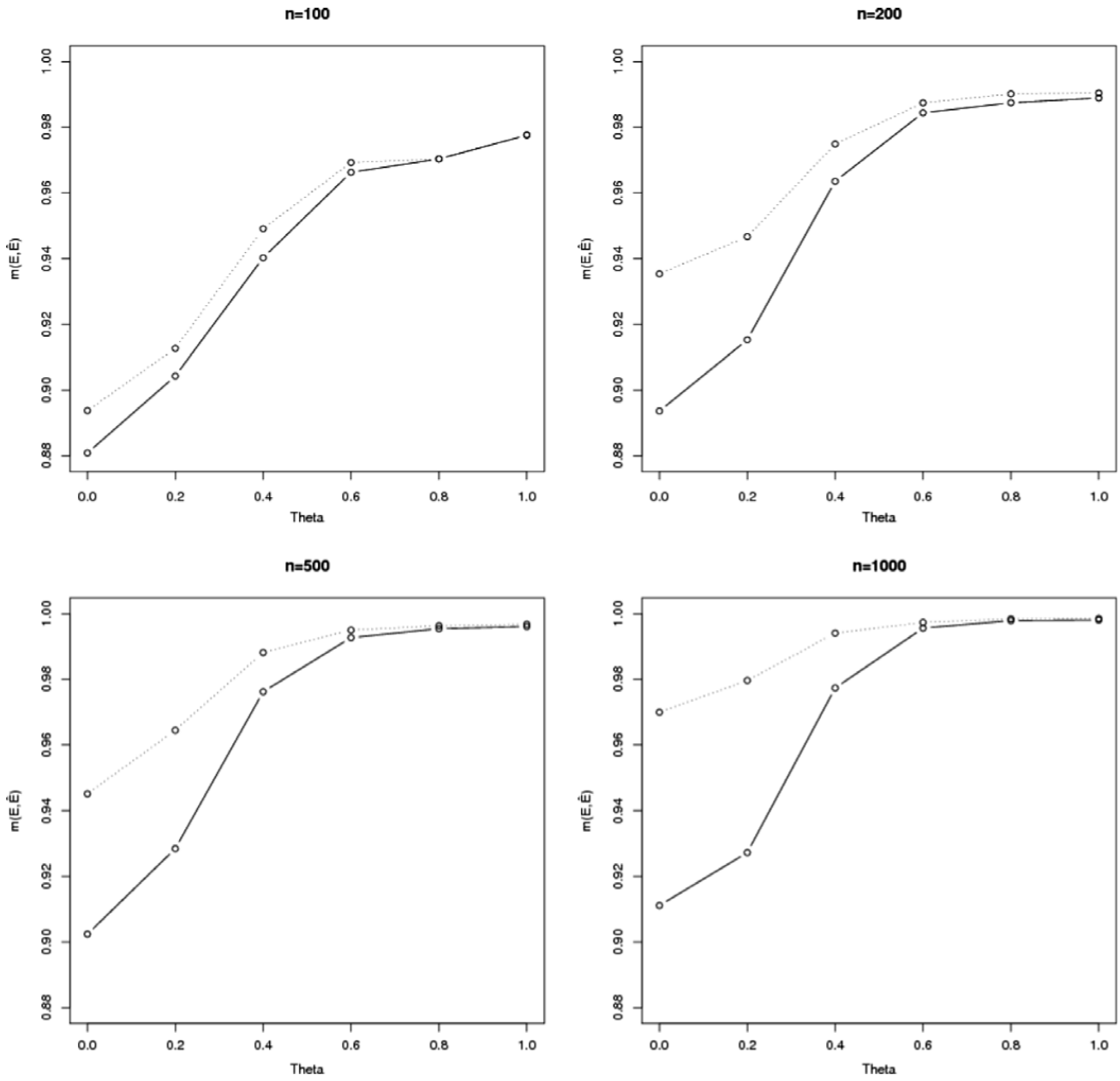

Fig. 4. Plots of the mean of the squared cosines for model (7) with different values of θ and n (solid line: SIR, dotted line: cluster-based SIR).

simulated sample, the e.d.r. direction is estimated with SIR and cluster-based SIR. Cluster-based SIR was implemented with a number of clusters c varying from 1 to 10 (or 20 for $n = 1000$). In this simulation study, the optimal number of clusters was chosen according to criterion (6). We could also use the criterion (5) which gives very similar results but is computationally expensive. The quality measure presented for cluster-based SIR is the one obtained with the optimal number of clusters. Note that the best number may sometimes be equal to 1 (especially for $n = 100$), corresponding then to classical SIR.

Comments on the plots of the means of the squared cosines. Fig. 4 shows the mean of the $N = 100$ squared cosines obtained for each value of θ (from 0 to 1) with SIR and cluster-based SIR estimation methods.

- In each case, both methods give reliable results.
- For the four sample sizes, the performances of both methods increase as θ increases, that is as the data are close to be elliptically distributed ($\theta = 1$). For instance, for cluster-based SIR with $n = 500$, we obtain a mean squared cosine of 0.94 with $\theta = 0$ and 0.99 with $\theta = 1$. For classical SIR, we observe of course the same phenomenon, the mean squared cosine is 0.90 with $\theta = 0$ and 0.99 with $\theta = 1$. This shows that cluster-based SIR is above all helpful in the case of non-ellipticity, which was the aim of the proposed work. Moreover nothing is lost in the case of elliptical distribution. The quality of the results of cluster-based SIR are as good as the ones obtained with SIR.
- As already seen in Fig. 3, we have here a confirmation that the performances of both methods increase as the sample size gets higher. Larger the sample size is, better are the cluster-based SIR results. Indeed, with $n = 1000$ observations, the mean of the squared cosine increases from 0.91 with classical SIR to 0.97 with cluster-based SIR. However with a small

Please cite this article in press as: Kuentz, V., & Saracco, J. Cluster-based Sliced Inverse Regression. Journal of the Korean Statistical Society (2009), doi:10.1016/j.jkss.2009.08.004

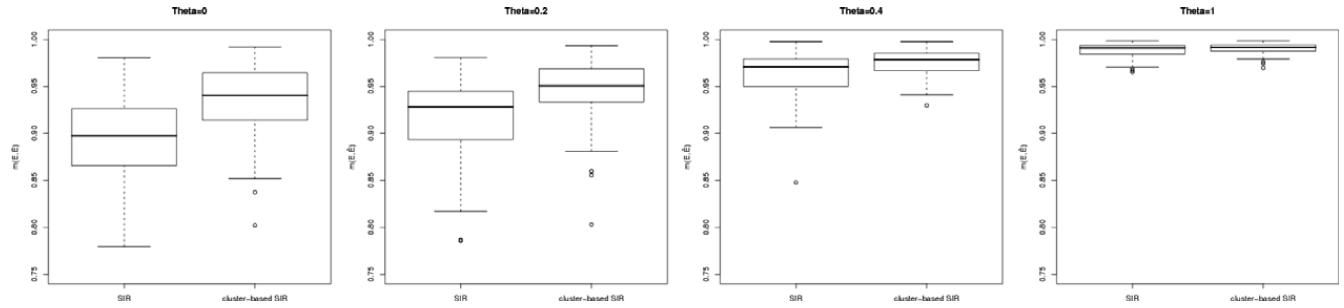

Fig. 5. Boxplots of the squared cosines for model (7) with different values of θ and $n = 200$.

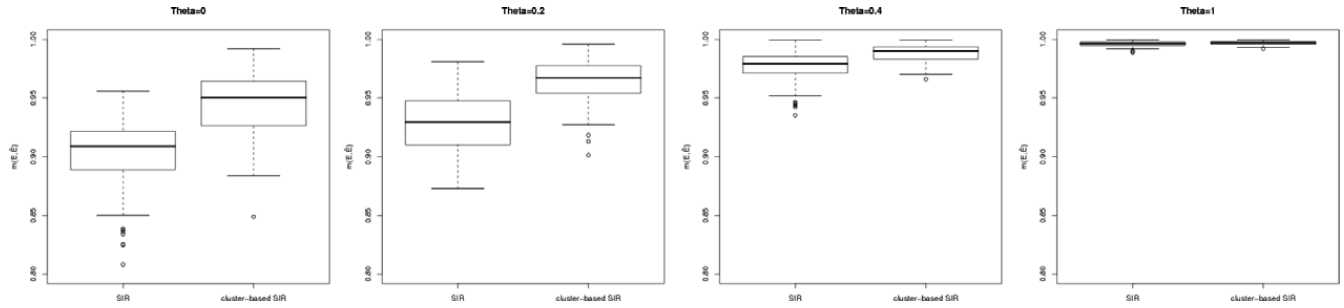

Fig. 6. Boxplots of the squared cosines for model (7) with different values of θ and $n = 500$.

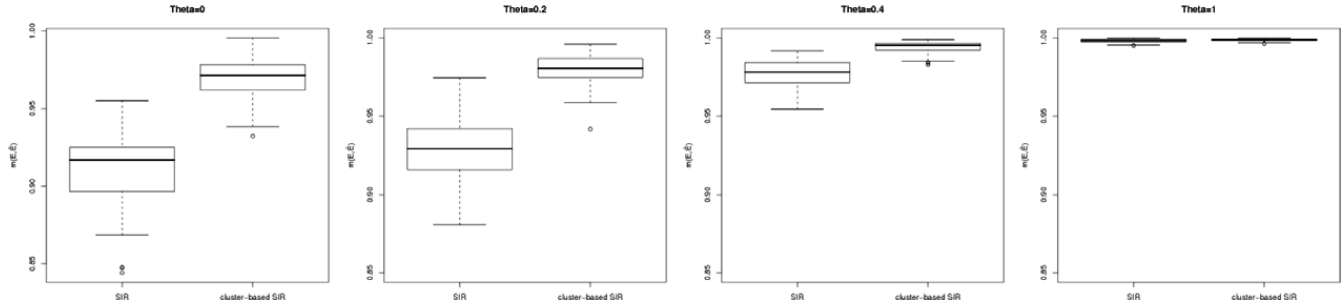

Fig. 7. Boxplots of the squared cosines for model (7) with different values of θ and $n = 1000$.

sample ($n = 100$), the improvement is not so high, the mean squared cosine is 0.88 with SIR and a little more than 0.89 with cluster-based SIR. This comes from the fact that the proposed approach partitions the predictor space. Indeed with large samples, the clustering is better: clusters are better defined and bigger. Therefore the slicing in SIR step occurs on a large number of observations. On the contrary with a small number of observations, the clustering is not so clear and provides sometimes clusters too small for the slicing to be computed. Then cluster-based SIR chooses one cluster, that is the partitioning does not improve the results.

Comments on the boxplots of the squared cosines. Figs. 5–7 show the boxplots of the squared cosines based on the $N = 100$ data replications for $n = 200, 500, 1000$ (with $\theta = 0, 0.2, 0.4, 1$). Both methods give reliable results with a quality measure increasing as the sample size gets higher. For $\theta = 0$ (non-elliptical distribution) and the three sample sizes, cluster-based SIR is better than classical SIR. However in the case of elliptical distribution, the two methods are as effective. The benefits of the use of cluster-based SIR approach is obvious in the case of non-elliptical distribution and big sample size ($n = 1000$).

4.2.4. Comparison with cluster-based OLS

In this section, we compare SIR, cluster-based SIR and cluster-based OLS introduced by Li et al. (2004), on $N = 100$ data replications of model (7) with $\theta = 0$ and $n = 500$. For cluster-based SIR, the optimal number of clusters is chosen according to criterion (5) and for cluster-based OLS, the optimal number of clusters is obtained with a kernel estimator proposed by the authors in their paper. Fig. 8 shows the boxplots of the efficiency measures obtained with SIR, cluster-based SIR and cluster-based OLS. We see that cluster-based SIR is more efficient than SIR and cluster-based OLS. The width of the boxplots of the quality measures is smaller with cluster-based SIR than with the other two methods.

Please cite this article in press as: Kuentz, V., & Saracco, J. Cluster-based Sliced Inverse Regression. Journal of the Korean Statistical Society (2009), doi:10.1016/j.jkss.2009.08.004

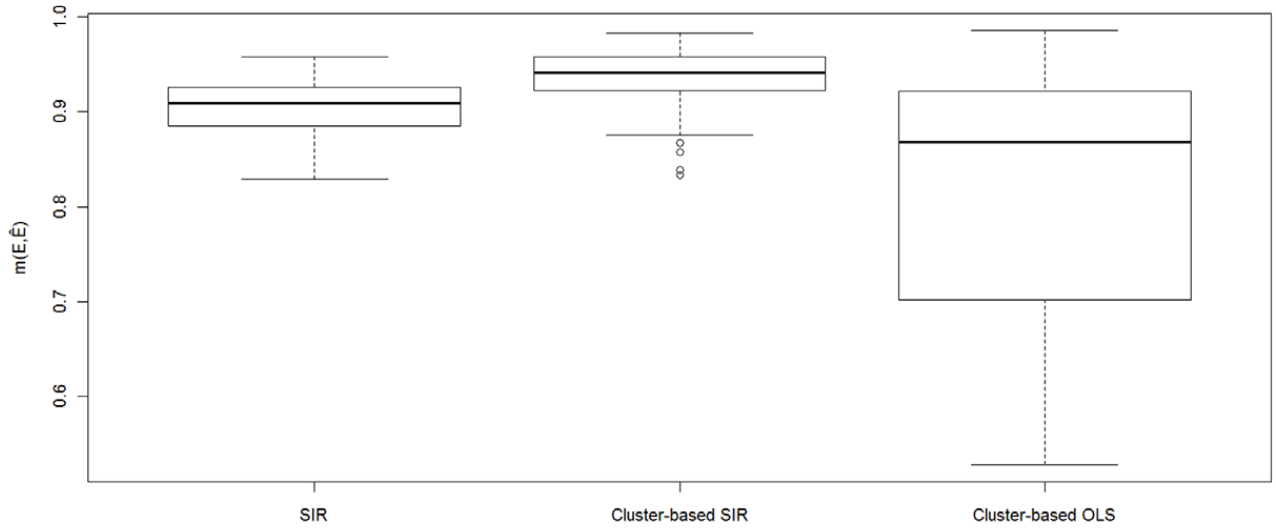

Fig. 8. Boxplots of the efficiency measures obtained with SIR, cluster-based SIR and cluster-based OLS for model (7) where $\theta = 0$ and $n = 500$.

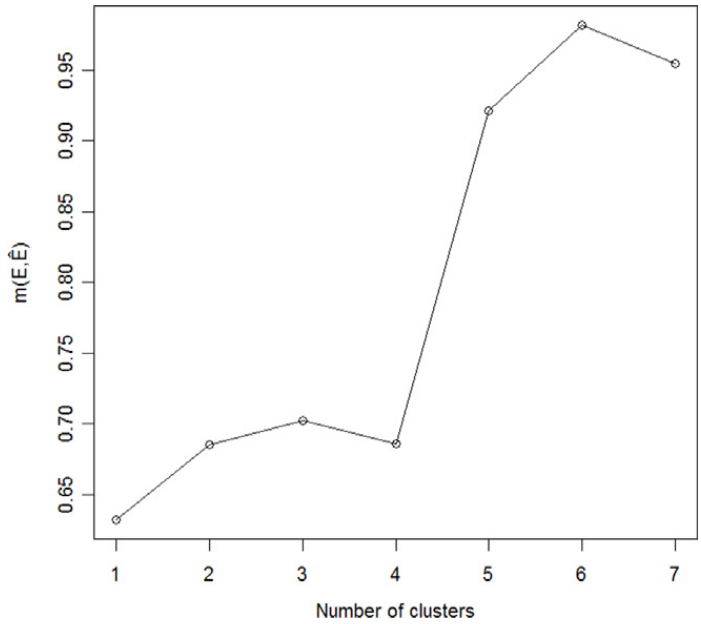

Fig. 9. Efficiency measure of the estimation obtained with SIR ($c = 1$) and cluster-based SIR ($c > 1$) for model (8) where $n = 200$ and $\theta = 0$.

4.3. Two indices model

4.3.1. Simulated model

We consider the following two indices model:

$$y = (\mathbf{x}'\beta_1 + \varepsilon_1)\mathbb{I}_{[\mathbf{x}'\beta_2+\varepsilon_2>0]}, \tag{8}$$

with $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5)'$, where $x_j \sim (1-\theta)\times \text{Exp}(1)+\theta\times\mathcal{N}(0, 1)$, $\varepsilon_1 \sim \mathcal{N}(0, 0.1^2)$ and $\varepsilon_2 \sim \mathcal{N}(0, 0.1^2)$. The variables x_j are independent, the error terms ε_1 and ε_1 are independent from each other and from $\mathbf{x}$. We choose as true normalized e.d.r. directions $\beta_1 = (\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}, 0, 0, 0)'$ and $\beta_2 = (0, 0, 0, \frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}})'$. This model is known as sample selection model. With SIR or cluster-based SIR, the slope parameters β_1 for the observation part of y and β_2 for the selection part (that is the state of y: non-observed (0) or observed) are not individually identifiable. In Chavent, Liquet, and Saracco (2009), a method based on SIR and canonical analysis provides estimates of the directions of β_1 and β_2 for multivariate semiparametric sample selection model, but the price to pay is to add identifiability conditions. Here only the e.d.r. space $E = \text{Span}(\beta_1, \beta_2)$ is identifiable and the quality can only be measured in simulation by $m(E, \hat{E})$ and by $\cos^2(\hat{\tilde{\beta}}_1, \beta_1)$ and $\cos^2(\hat{\tilde{\beta}}_2, \beta_2)$.

Please cite this article in press as: Kuentz, V., & Saracco, J. Cluster-based Sliced Inverse Regression. Journal of the Korean Statistical Society (2009), doi:10.1016/j.jkss.2009.08.004

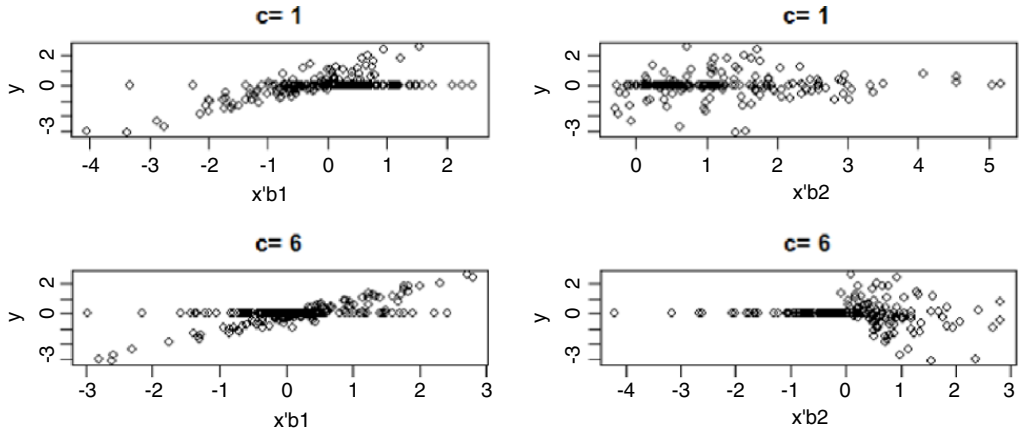

Fig. 10. Scatter plots of $\{(y_i, \mathbf{x}_i'\hat{b}_1), i = 1, \ldots, n\}$ and $\{(y_i, \mathbf{x}_i'\hat{b}_2), i = 1, \ldots, n\}$ for SIR ($c = 1$) at the top and scatter plots of $\{(y_i, \mathbf{x}_i'\hat{\tilde{b}}_1), i = 1, \ldots, n\}$ and $\{(y_i, \mathbf{x}_i'\hat{\tilde{b}}_2), i = 1, \ldots, n\}$ for cluster-based SIR ($\hat{c}^* = 6$) at the bottom.

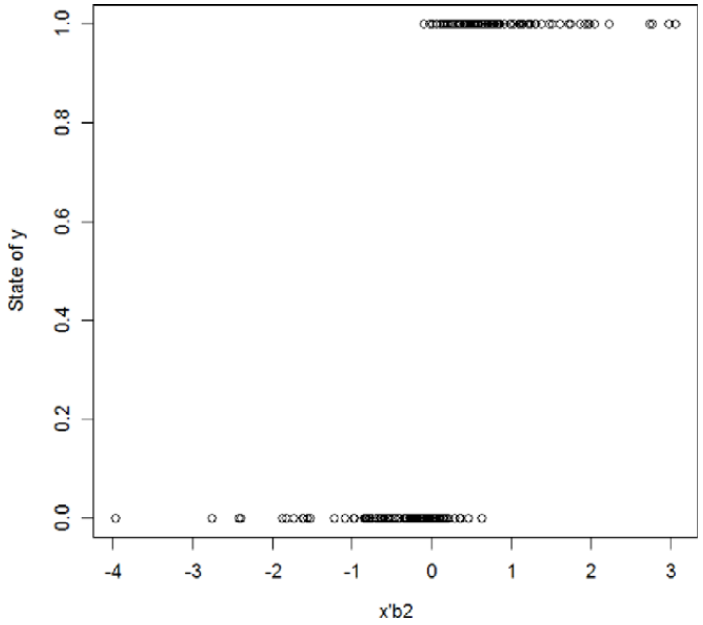

Fig. 11. Scatter plot of the state of y_i (0 when y_i is null and 1 when y_i is nonnull) versus $\mathbf{x}_i'\hat{\tilde{b}}_2$, $i = 1, \ldots, n$.

4.3.2. Single data replication

We exhibit a sample of size $n = 200$ of model (8) with $\theta = 0$. Fig. 9 shows the evolution of the quality criterion (6) as the number of clusters varies between 1 and 7. The case with one cluster matches with classical SIR. We see that the estimation of the e.d.r. space is improved by clustering the predictor variable $\mathbf{x}$. For cluster-based SIR, the maximum of the quality measure is reached at 6 clusters with an efficiency measure of 0.98 versus 0.63 with classical SIR.

Fig. 10 plots the response variable versus the two estimated e.d.r. directions for SIR and cluster-based SIR (with $\hat{c}^* = 6$). Note that with this simulated sample and the corresponding estimates, the indices $x'\hat{\tilde{b}}_1$ and $x'\beta_1$ (resp. $x'\hat{\tilde{b}}_2$ and $x'\beta_2$) are highly correlated, this is only due to chance (since β_1 and β_2 are not theoretically individually identifiable). Thereby we can observe that the structure is more relevant with cluster-based SIR than with SIR, which emphasizes that the partitioning of the predictor variable is helpful. The bottom left of Fig. 10 shows that the first true e.d.r. direction of model (8) is well recovered. At the bottom right of Fig. 10 we see that the second true e.d.r. direction is also found. Fig. 11 of the scatter plot of the state of y_i (null or nonnull) versus $\mathbf{x}_i'\hat{\tilde{b}}_2$, $i = 1, \ldots, n$ confirms that the second estimated e.d.r. direction differentiates between null and nonnull values of the response variable.

4.3.3. Multiple data replications

In this section, we compare SIR and cluster-based SIR on $N = 100$ data replications of model (8).

Please cite this article in press as: Kuentz, V., & Saracco, J. Cluster-based Sliced Inverse Regression. Journal of the Korean Statistical Society (2009), doi:10.1016/j.jkss.2009.08.004

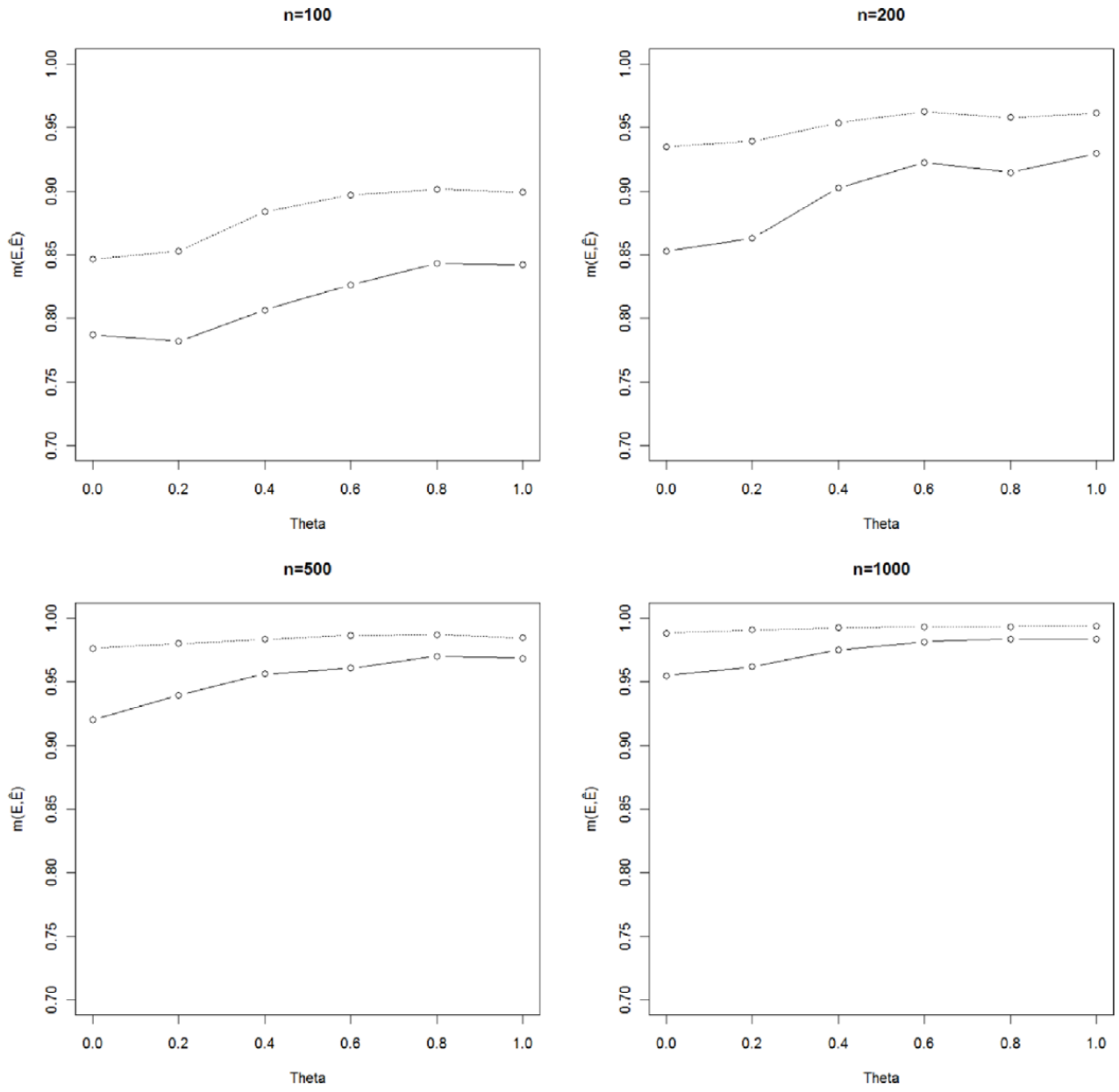

Fig. 12. Plots of the mean of the squared cosines for model (8) with different values of θ and n (solid line: SIR, dotted line: cluster-based SIR).

Comments on the plots of the means of the efficiency measures. Fig. 12 shows the mean of the $N = 100$ quality measures obtained for each value of θ (from 0 to 1) with SIR and cluster-based SIR. Both methods provide very good results and the quality of the estimation increases as the data are close to be elliptical ($\theta = 1$). Cluster-based SIR is always better than classical SIR, especially for sample size of 200 or 500. Indeed with small samples ($n = 100$), the improvement due to clustering is not so high because clusters are sometimes bad defined or too small, preventing then the use of SIR. With large samples ($n = 1000$), the performances of the two methods are similar (with a slight advantage for cluster-based SIR).

Comments on the boxplots of the efficiency measures. Figs. 13–15 show the boxplots of the $N = 100$ efficiency measures obtained for $\theta = 0, 0.2, 0.4, 1$ with SIR and cluster-based SIR estimation methods. The efficiency of both methods is improved when the sample size n increases. Cluster-based SIR always provides better estimations than classical SIR. For both methods, the quality measure increases as θ increases. Compared to SIR, cluster-based SIR is less sensitive to violation of the linearity condition (or elliptical distribution). This is true for any sample size.

4.4. An iterative implementation

For simplicity in the presentation of the iterative implementation we consider the single index model (3). The iterative cluster-based SIR is based on the following implementation. We compute with cluster-based SIR the estimated e.d.r. direction $\hat{b}^{(0)}$. Then we cluster the sample $\{\mathbf{x}_i'\hat{b}^{(0)}, i = 1, \ldots, n\}$, and according to this partition, we compute, with cluster-

Please cite this article in press as: Kuentz, V., & Saracco, J. Cluster-based Sliced Inverse Regression. Journal of the Korean Statistical Society (2009), doi:10.1016/j.jkss.2009.08.004

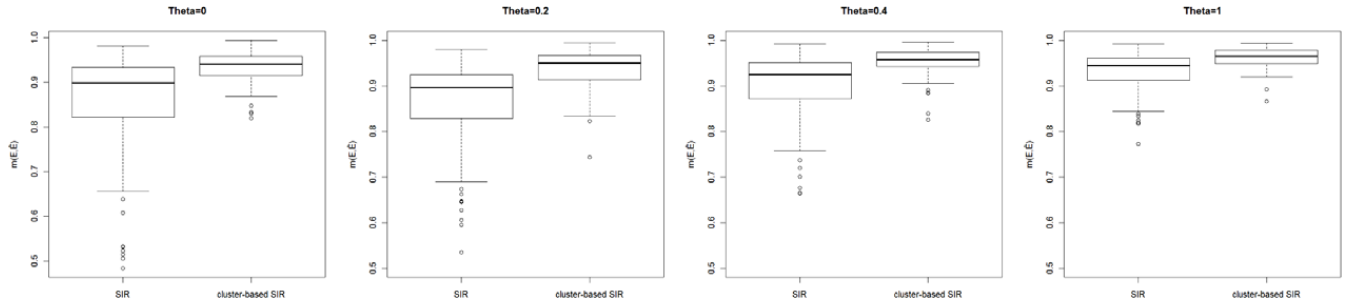

Fig. 13. Boxplots of the efficiency measures for model (8) with different values of θ and $n = 200$.

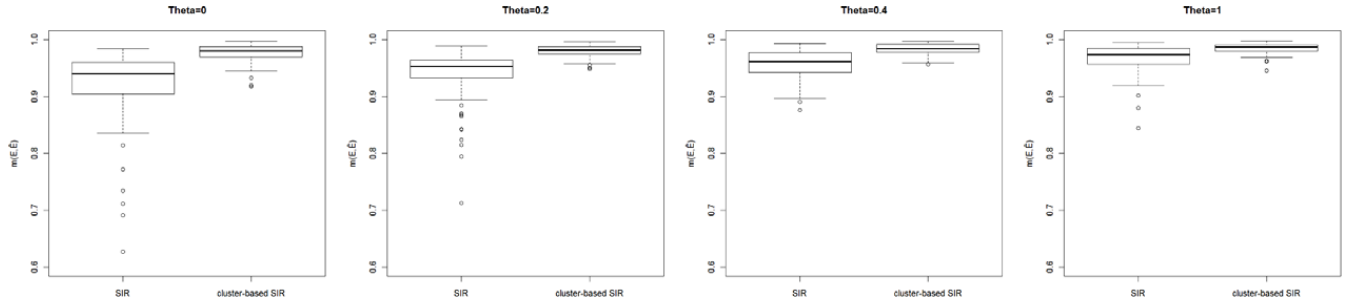

Fig. 14. Boxplots of the efficiency measures for model (8) with different values of θ and $n = 500$.

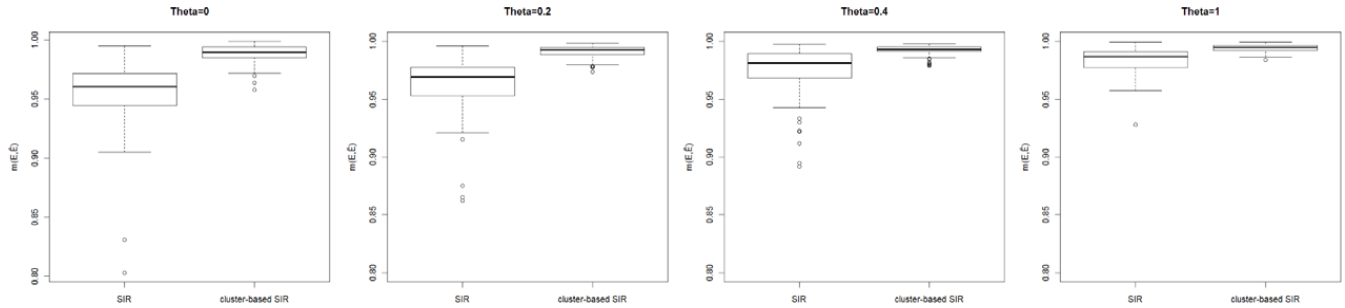

Fig. 15. Boxplots of the efficiency measures for model (8) with different values of θ and $n = 1000$.

based SIR, a new estimation $\hat{b}^{(1)}$ of the e.d.r. direction. As the clustering occurs in a lower-dimensional space (here in $\mathbb{R}$), the partitioning may be better defined and the estimation of the new e.d.r. direction may be improved. We iterate this principle until a stopping criterion, based on the work of Li et al. (2004), is reached: the iteration procedure stops when the correlation between $\mathbf{x}'\hat{b}^{(m)}$ and $\mathbf{x}\hat{b}^{(m+1)}$ reaches a specified threshold (fixed at 0.9 in our simulations). Note that in practice we have often observed that only one iteration is necessary.

In the following simulation results we consider the simulated model (7) with the same parameters given in Section 4.2. We estimate the e.d.r. direction with the cluster-based SIR method and its iterative implementation version on $N = 100$ data replications with $n = 200, 500, 1000$ and $\theta = 0$ (which corresponds to non-elliptical distribution). Fig. 16 shows the boxplots of the quality measure for the different sample sizes. We can clearly observe the improvement of the quality of the estimated e.d.r. basis with the iterative cluster-based SIR approach, particularly for large sample size. The same phenomenon was observed for the other values of θ (0.2,0.4,0.6,0.8,1). We did not report the corresponding results here.

5. Real data application

We consider the data example of horse mussels described in Camden (1989) or Cook and Weisberg (1999). The observations correspond to $n = 201$ horse mussels captured in the Marlborough Sounds at the Northeast of New Zealand's South Island. The response variable y is the muscle mass, the edible portion of the mussel, in grams. The predictor $\mathbf{x}$ is of dimension $p = 4$ and measures numerical characteristics of the shell: length, width, height, each in mm, and mass in grams. In this problem, as the response is discrete, it is slightly transformed as follows $y = y + \epsilon, \epsilon \sim \mathcal{N}(0, 0.01^2)$. Thus we get a continuous variable, which improves the slicing step of SIR. Note that the number of slices used for SIR and cluster-based SIR is $H^{(j)} = 4$ for each cluster $j = 1, \ldots, c$. We used $C^n_{\max} = 8$ and $n_{h,\min} = 6$ in the selection of the number of clusters and in our modified k-means approach. The various studies on this data example have reached to a one-dimensional structure.

Please cite this article in press as: Kuentz, V., & Saracco, J. Cluster-based Sliced Inverse Regression. Journal of the Korean Statistical Society (2009), doi:10.1016/j.jkss.2009.08.004

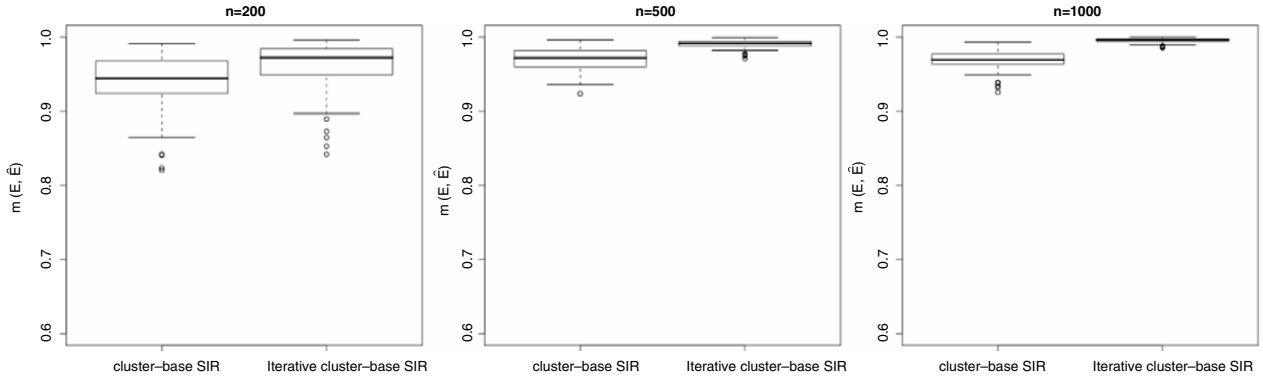

Fig. 16. Boxplots of the efficiency measures for model (7) with cluster-based SIR and iterative cluster-based SIR, for different values of n and $\theta = 0$.

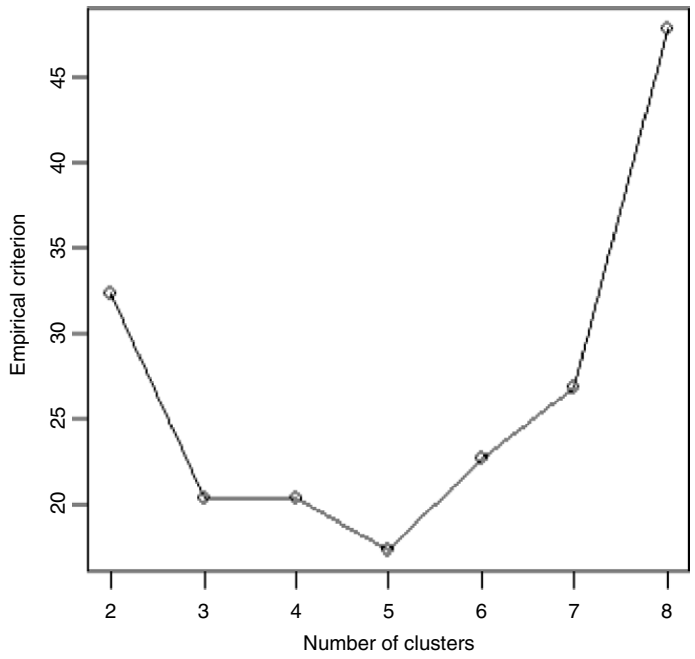

Fig. 17. Empirical criterion defined in (5) with cluster-based SIR ($c > 1$).

Fig. 17 plots empirical criterion (5) and shows that the choice $\hat{c}^* = 5$ of the number of clusters seems to be the best adapted to iterative cluster-based SIR.

For the chosen couple of parameters ($\hat{K} = 1, \hat{c}^* = 5$) we check at the top of Fig. 18 that there is a relevant structure in the scatter plot $\{(y_i, \mathbf{x}_i'\hat{b}_1), i = 1, \ldots, n\}$. On the contrary for the couple ($\hat{K} = 2, \hat{c}^* = 5$), there is no structure in the scatter plot $\{(y_i, \mathbf{x}_i'\hat{\hat{b}}_2), i = 1, \ldots, n\}$, see bottom of Fig. 18.

Then we compare the prediction reached on a test sample with SIR and iterative cluster-based SIR using the following algorithm.

Step 1. We split the data into two subsets: $S_J = \{(y_j, \mathbf{x}_j'), j \in J\}$ the training sample containing almost 80% of the total number of observations, and $S_I = \{(y_i, \mathbf{x}_i'), i \in I\}$ the test sample of the remaining observations. Let $n_I = \text{card}(I)$.

Step 2. We use the training sample S_J to compute the estimated e.d.r. direction with SIR, denoted $\hat{b}_{[1]}$, and with iterative cluster-based SIR for $\hat{c}^* = 5$ clusters, denoted $\hat{b}_{[5]}$.

Step 3. We compute the kernel estimate $\hat{y}_{i,[c]}$ of $\mathbb{E}(y|\mathbf{x}_i'\hat{b}_{[c]})$ for $i \in I$ using the sample $\{(y_i, \mathbf{x}_i'\hat{b}_{[c]}), i \in J\}$. We get $\hat{y}_{i,[1]}, i \in I$ for SIR and $\hat{y}_{i,[5]}, i \in I$ for iterative cluster-based SIR.

Step 4. We compute the Mean Absolute Relative Error (MARE) for both SIR and cluster-based SIR estimates as follows:

$$\text{MARE} = \frac{1}{n_I} \sum_{i \in S_I} \left| \frac{y_i - \hat{y}_{i,[c]}}{y_i} \right|.$$

The previous algorithm is repeated $N = 100$ times. Fig. 19 shows the boxplots of the MARE values obtained with SIR and iterative cluster-based SIR. Iterative cluster-based SIR is clearly more efficient than SIR. The range of the boxplot is smaller

Please cite this article in press as: Kuentz, V., & Saracco, J. Cluster-based Sliced Inverse Regression. Journal of the Korean Statistical Society (2009), doi:10.1016/j.jkss.2009.08.004

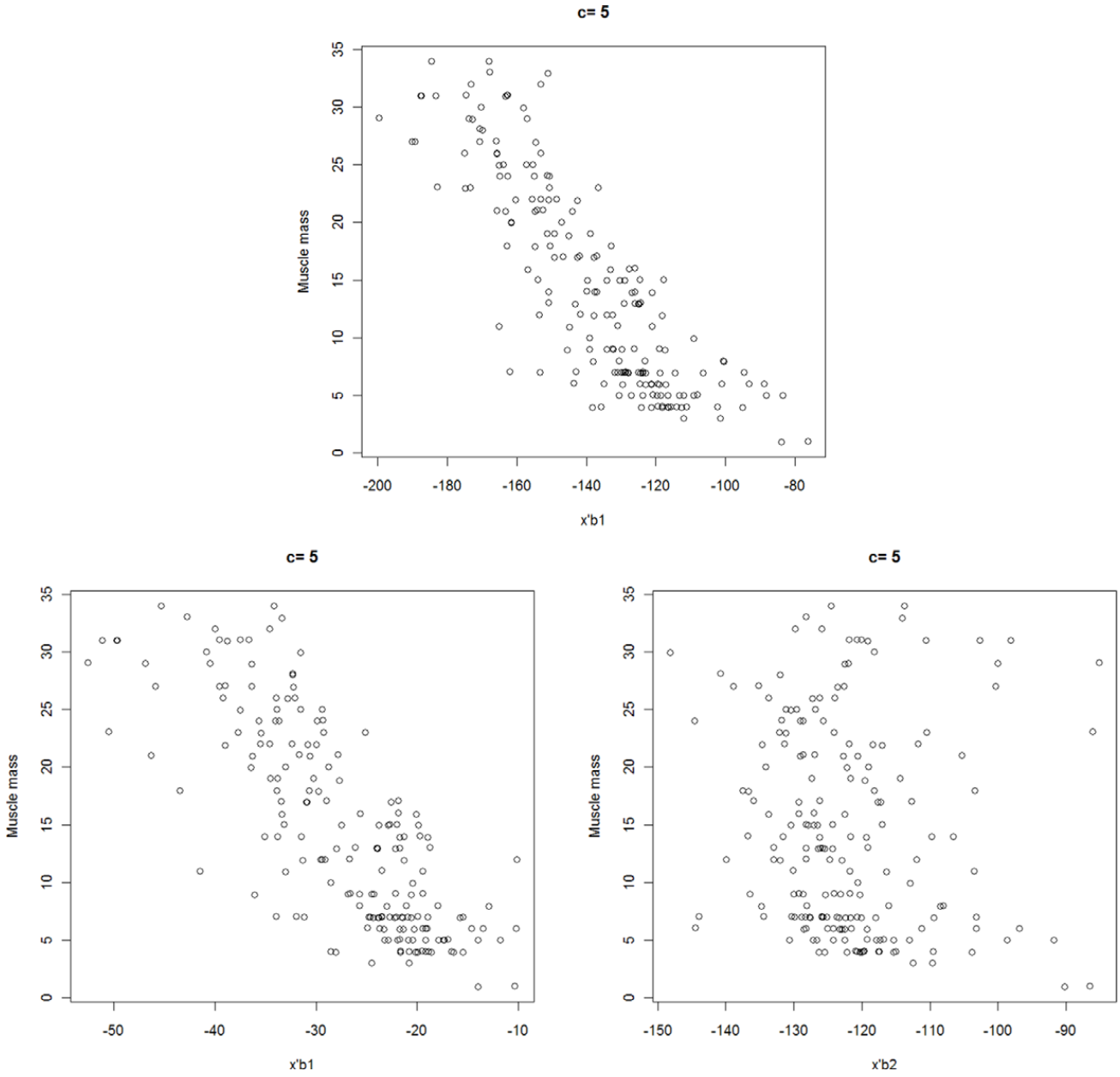

Fig. 18. Scatter plot $\{(y_i, \mathbf{x}_i'\hat{b}_1), i = 1, \ldots, n\}$ at the top and scatter plot $\{(y_i, \mathbf{x}_i'\hat{\tilde{b}}_1, \mathbf{x}_i'\hat{\tilde{b}}_2), i = 1, \ldots, n\}$ at the bottom.

with the use of clustering the predictor space. The median MARE obtained with cluster-based SIR is decreased by half (0.2 versus 0.4 with classical SIR).

6. Concluding remarks

In this article, we have proposed an extension of the well-known dimension reduction method SIR, called cluster-based SIR, which can be used when the crucial linearity condition is not verified. The idea is to partition the predictor space so that the linearity condition approximately holds in each cluster. The optimal number of clusters can be computed from a minimization criterion. Asymptotic properties of the estimator have been obtained. A simulation study has shown the good numerical behaviour of the proposed approach. A real data application has shown the better predictive performance of cluster-based SIR over SIR. Note that cluster-based SIR is less sensitive than SIR to violation of the linearity condition. Thus it opens future prospects for a broader use of SIR. The method has been implemented in R and source codes are available from the authors. As we mentioned in the introduction, the k-means clustering does not always ensure to construct approximately elliptical clusters. However from our simulation results, the use of cluster-based SIR instead of classical SIR globally provides better estimation of the e.d.r. space. The small price to pay is that the cluster-based SIR method is relative much time consuming computationally. Finally our method can be extended to SIR-II and SIR$_\alpha$ (see Gannoun & Saracco, 2003; Li, 1991) or to multivariate SIR approach (see Barreda, Gannoun, & Sarraco, 2007).

Please cite this article in press as: Kuentz, V., & Saracco, J. Cluster-based Sliced Inverse Regression. Journal of the Korean Statistical Society (2009), doi:10.1016/j.jkss.2009.08.004

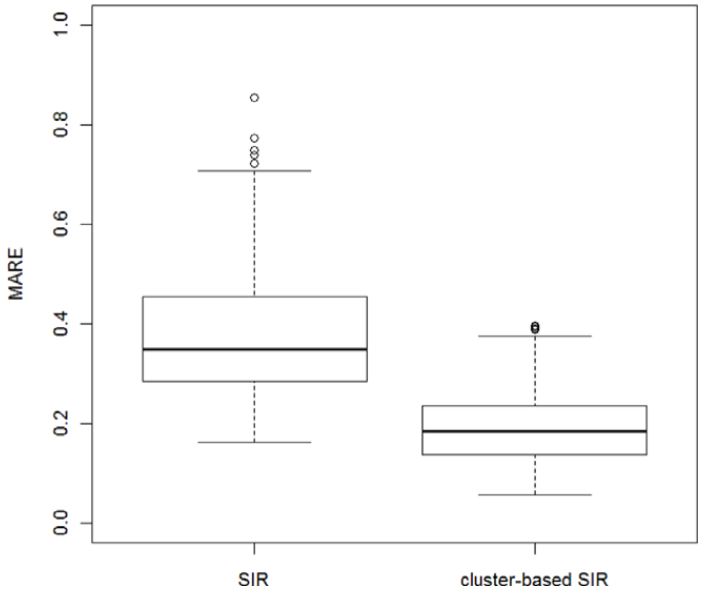

Fig. 19. Boxplots of the MARE values obtained with SIR and cluster-based SIR ($\hat{c}^* = 5$).

Appendix. Proof of Theorem 2

Let $D_1 \otimes D_2$ denote the Kronecker product of the matrices D_1 and D_2 (see for instance Harville (1999) for some useful properties of the Kronecker product). Let $D = [d_1, \ldots, d_q]$ be a $(p \times q)$ matrix, where the d_k's are p-dimensional column vectors. We note $\mathrm{vec}(D)$ the pq-dimensional column vector: $\mathrm{vec}(D) = \begin{pmatrix} d_1 \\ \vdots \\ d_q \end{pmatrix}$.

The proof of Theorem 2 is divided into four steps.

Step 1: Asymptotic distribution of $\hat{b}^{(j)}$

Classical asymptotic theory of SIR gives us the following result for each cluster $j = 1, \ldots, c$:

$$\sqrt{n}(\hat{b}^{(j)} - b^{(j)}) \longrightarrow_d W^{(j)} \sim \mathcal{N}(0, \Gamma^{(j)}), \tag{9}$$

where the expression of $\Gamma^{(j)}$ can be found in Saracco (1997).

Step 2: Asymptotic distribution of $\hat{B}$

Under conditions (A2) and (A3), we have:

$$\sqrt{n}(\mathrm{vec}(\hat{B}) - \mathrm{vec}(B)) \longrightarrow_d \mathrm{vec}\begin{pmatrix} W^{(1)} \\ \ldots \\ W^{(j)} \end{pmatrix} \sim \mathcal{N}(0, \Gamma), \tag{10}$$

where:

$$\Gamma = \begin{pmatrix} \Gamma^{(1)} & & 0 \\ & \ddots & \\ 0 & & \Gamma^{(c)} \end{pmatrix}. \tag{11}$$

Step 3: Asymptotic distribution of $\hat{B}\hat{B}'$

We use Delta method. For that, we have to write $\mathrm{vec}(BB')$ in terms of $\mathrm{vec}(B)$.

We have: $\mathrm{vec}(BB') = \mathrm{vec}(BI_cB') = (B \otimes B)\mathrm{vec}(I_c)$. As $\mathrm{vec}(\mathrm{vec}(BB')) = \mathrm{vec}(BB')$, we can write:

$$\begin{aligned} \mathrm{vec}(BB') &= \mathrm{vec}((B \otimes B)\mathrm{vec}(I_c)) \\ &= (\mathrm{vec}(I_c)' \otimes I_{p^2})\mathrm{vec}(B \otimes B) \\ &= (\mathrm{vec}(I_c)' \otimes I_{p^2})(I_c \otimes K_{cp} \otimes I_p)(\mathrm{vec}(B) \otimes \mathrm{vec}(B)), \end{aligned} \tag{12}$$

where the vec-permutation matrix K_{cp} is equal to $K_{cp} = \sum_{i=1}^{c}\sum_{i=1}^{p}(U_{ij} \otimes U_{ij}')$ with $U_{ij} = e_i u_j'$ and e_i is the ith column of I_c and u_j the jth column of I_p.

Please cite this article in press as: Kuentz, V., & Saracco, J. Cluster-based Sliced Inverse Regression. Journal of the Korean Statistical Society (2009), doi:10.1016/j.jkss.2009.08.004

Thus we define the following function:

$$f : \mathbb{R}^{pc} \to \mathbb{R}^{p^2}, \quad x \mapsto M_1 M_2 (x \otimes x), \tag{13}$$

with matrices $M_1 = (\mathrm{vec}(I_c)' \otimes I_{p^2})$ and $M_2 = (I_c \otimes K_{cp} \otimes I_p)$.

The Jacobian matrix D associated to f is then equal to:

$$\begin{aligned} D &= \frac{\partial f(x)}{\partial x'} = M_1 M_2 \frac{\partial (x \otimes x)}{\partial x'} = M_1 M_2 \frac{\partial \mathrm{vec}(x \otimes x)}{\partial x'} \\ &= M_1 M_2 (K_{1pc} \otimes I_{pc}) \left[x \otimes \frac{\partial x}{\partial x'} + \frac{\partial x}{\partial x'} \otimes x \right] \\ &= M_1 M_2 (K_{1pc} \otimes I_{pc}) [x \otimes I_{pc} + I_{pc} \otimes x]. \end{aligned} \tag{14}$$

Then applying Delta method with function f defined in (13) and Jacobian matrix D defined in (14), we get:

$$\sqrt{n}(\mathrm{vec}(\hat{B}\hat{B}') - \mathrm{vec}(BB')) \longrightarrow_d V \sim \mathcal{N}(0, \Gamma_V = D\Gamma D'), \tag{15}$$

with matrices Γ and D respectively defined in (11) and (14).

Step 4: Asymptotic distribution of $\hat{b}$

Recall that $\hat{b}$ (resp. b) is the eigenvector associated to the largest eigenvalue $\hat{\lambda}$ (resp. λ) of $\hat{B}\hat{B}'$ (resp. BB'). We will note N^+ the Moore–Penrose generalized inverse of the square matrix N.

Since $\hat{B}\hat{B}' = BB' + O_p(n^{-1/2})$ and using (15), according to Lemma 1 of Saracco (1997), we get that:

$$\sqrt{n}(\hat{b} - b) \longrightarrow_d (BB' - \lambda I_p)^+ V b, \tag{16}$$

where:

$$(BB' - \lambda I_p)^+ V b \sim \mathcal{N}(0, \Gamma_U), \tag{17}$$

with:

$$\Gamma_U = [b' \otimes (BB' - \lambda I_p)^+] \Gamma_V [b \otimes (BB' - \lambda I_p)^+]. \tag{18}$$

References

Aragon, Y., & Saracco, J. (1997). Sliced Inverse Regression (SIR): An appraisal of small sample alternatives to slicing. *Computational Statistics*, *12*, 109–130.
Barreda, L., Gannoun, A., & Sarraco, J. (2007). Some extensions of multivariate Sliced Inverse Regression. *Journal of Statistical Computation and Simulation*, *77*, 1–17.
Barrios, M. P., & Velilla, S. (2007). A bootstrap method for assessing the dimension of a general regression problem. *Statistics & Probability Letters*, *77*, 247–255.
Brillinger, D. R. (1983). A generalized linear model with "Gaussian" regressor variables. In P. J. Bickel, K. A. Doksum, & J. L. Hodges (Eds.), *Festschrift for Erich L. Lehmann in honor of his sixty-fifth birthday* (pp. 97–114). Belmont, Calif: Wadsworth.
Camden, M. (1989). *The data bundle*. Wellington: New Zealand Statistical Association.
Chavent, M., Liquet, B., & Saracco, J. (2009). A semiparametric approach for a multivariate sample selection model. *Statistica Sinica* (in press).
Cheung, Y. M. (2003). *K*-means: A new generalized *k*-means clustering algorithm. *Pattern Recognition Letters*, *24*(15), 2883–2893.
Cook, R. D., & Nachtsheim, C. J. (1994). Re-weighting to achieve elliptically contoured covariates in regression. *Journal of the American Statistical Association*, *89*, 592–599.
Cook, R. D., & Weisberg, S. (1999). *Applied regression including computing and graphics*. New York: Wiley.
Everitt, B. S., Landau, S., & Leese, M. (2001). *Cluster analysis* (4th ed.). A Hodder Arnold Publication.
Ferré, L. (1998). Determining the dimension in Sliced Inverse Regression and related methods. *Journal of the American Statistical Association*, *93*(441), 132–140.
Gannoun, A., & Saracco, J. (2003). An asymptotic theory for SIR_α method. *Statistica Sinica*, *13*(2), 297–310.
Hall, P., & Li, K. C. (1993). On almost linearity of low dimensional projections from high dimensional data. *The Annals of Statistics*, *21*, 867–889.
Harville, D. A. (1999). *Matrix algebra from a statistician's perspective*. Springer-Verlag.
Li, K. C. (1991). Sliced inverse regression for dimension reduction, with discussion. *Journal of the American Statistical Association*, *86*, 316–342.
Li, L., Cook, R. D., & Nachtsheim, C. J. (2004). Cluster-based estimation for sufficient dimension reduction. *Computational Statistics & Data Analysis*, *47*, 175–193.
Liquet, B., & Saracco, J. (2008). Application of the bootstrap approach to the choice of dimension and the α parameter in the SIR_α method. *Communications in Statistics — Simulation and Computation*, *37*(6), 1198–1218.
Saracco, J. (1997). An asymptotic theory for Sliced Inverse Regression. *Communications in Statistics — Theory and Methods*, *26*, 2141–2171.
Saracco, J. (2001). Pooled slicing methods versus slicing methods. *Communications in Statistics — Simulation and Computation*, *30*(3), 489–513.
Schott, J. R. (1994). Determining the dimensionality in Sliced Inverse Regression. *Journal of the American Statistical Association*, *89*, 141–148.
Tyler, D. E. (1981). Asymptotic inference for eigenvectors. *The Annals of Statistics*, *9*, 725–736.

Please cite this article in press as: Kuentz, V., & Saracco, J. Cluster-based Sliced Inverse Regression. Journal of the Korean Statistical Society (2009), doi:10.1016/j.jkss.2009.08.004

3.3 Bagging versions of Sliced Inverse Regression

Abstract. Sliced Inverse Regression (SIR) introduced by Li (1991) is a well-known dimension reduction method in semiparametric regression. In this paper we propose bagging versions of SIR which consist in using bootstrap replications of the data set and in aggregating the corresponding estimators. We give the asymptotic distribution of the Bagging-SIR estimator. A simulation study is used to compare the numerical performance of the proposed alternative bagging versions of SIR with the classical SIR approach. The benefits of these methods are significant for noisy models and when the sample size is small. The `R` codes are available from the authors.

Keywords : bootstrap, Bagging, Sliced Inverse Regression (SIR), effective dimension reduction (e.d.r.) space.

3.3.1 Introduction

The bootstrap method proposed by Efron (1979) is a well-known resampling method, which aims at getting information on a data set by generating multiple versions of the original data. A relevant presentation of the bootstrap approach can be found in Efron (1993) or Shao and Tu (1995). Beran and Srivastava (1985) focused on bootstrapping the sample covariance matrix and gave bootstrap tests and confidence regions for some eigenfunctions of the population covariance matrix. Bagging (see for instance Breiman, 1994 or Bühlmann, 2004), the acronym for bootstrap aggregating, consists in using bootstrap replications of the data set, then computing the statistic of interest in each bootstrap sample and finally aggregating the different estimations. This aggregation often consists in averaging the multiple estimations when working with a numerical outcome and voting among classifiers when predicting a class. The benefits of bagging in the performance of regression and classification trees were demonstrated in Breiman's pioneering paper (Breiman, 1996).

In the theory of sufficient dimension reduction, Sliced Inverse Regression (SIR) methods introduced by Li (1991) are well-known techniques. These semiparametric regression methods assume that the features of a p-dimensional explanatory variable $\mathbf{x} = (x_1, \ldots, x_p)'$, with $\mathbb{E}(\mathbf{x}) = \mu$ and $\mathbb{V}(\mathbf{x}) = \Sigma$, can be captured in a lower K-dimensional projection subspace (with $K < p$). They make it possible to estimate a basis of this linear subspace. The corresponding model assumes that the dependency between the predictors and the response variable y is described by linear combinations of the predictors. The underlying semiparametric model is written :

$$y = f(\mathbf{x}'\gamma_1, \ldots, \mathbf{x}'\gamma_K, \varepsilon), \tag{3.5}$$

where f is an unknown link function, ε is an unknown random error assumed to be independent of $\mathbf{x}$, and $\gamma_1, \ldots, \gamma_K$ are K unknown vectors in $\mathbb{R}^p$, assumed to be linearly independent. As none condition on the form of f is imposed, the vectors γ_k are not identifiable. It is only possible to estimate the space spanned by these vectors, called the effective dimension reduction (e.d.r.) space, which will be denoted by E.

The basic principle of SIR methods is to reverse the role of y and $\mathbf{x}$ and to study the property of the conditional moments of $\mathbf{x}$ given y to recover the e.d.r. space. In this paper, we will focus only on the SIR method which is based on the first conditional moment. To facilitate the estimation of the inverse conditional mean, a slicing on the response variable y is performed. Let us denote by T this transformation of y.

Under this linearity condition :

$$\mathbb{E}(\mathbf{x}'b|\mathbf{x}'\gamma_1,\ldots,\mathbf{x}'\gamma_K) \text{ is linear in } \mathbf{x}'\gamma_1,\ldots,\mathbf{x}'\gamma_K \text{ for any } b,$$

the following geometrical property of SIR has been shown by Li (1991) : the centered inverse regression curve, $\mathbb{E}(\mathbf{x}|T(y))-\mathbb{E}(\mathbf{x})$ as y varies, is contained in the linear subspace of $\mathbb{R}^p$ spanned by the vectors $\Sigma\gamma_1,\ldots,\Sigma\gamma_K$. A straightforward consequence is that the covariance matrix $M = \mathbb{V}(\mathbb{E}(\mathbf{x}|T(y)))$ is degenerated in any direction Σ-orthogonal to the γ_k's. Therefore the eigenvectors associated with the nonnull K eigenvalues of the matrix $\Sigma^{-1}M$ are e.d.r. directions, that is are in the e.d.r. space. Note that the linearity condition is verified when $\mathbf{x}$ follows an elliptically symmetric distribution, for instance when $\mathbf{x}$ is multinormal.

Another way to estimate the e.d.r. space is to consider $\mathbf{z} = \Sigma^{-1/2}(\mathbf{x}-\mu)$, the standardized version of $\mathbf{x}$. Indeed model (3.5) can be written :

$$y = h(\mathbf{z}'\eta_1,\ldots,\mathbf{z}'\eta_K,\varepsilon),$$

where $\eta_k = \Sigma^{1/2}\gamma_k, k = 1\ldots,K$ are standardized e.d.r. directions and the space spanned by these vectors is the standardized e.d.r. space, denoted E_s. Then we compute the covariance matrix of the standardized inverse regression curve :

$$M_s = \mathbb{V}(\mathbb{E}(\mathbf{z}|T(y))).$$

The eigenvectors of M_s associated with the nonnull K eigenvalues of the matrix M_s, denoted $v_1,\ldots,v_K$, are standardized e.d.r. directions. Then the e.d.r. space can be computed from this standardized e.d.r. space, and is spanned by the vectors $\Sigma^{-1/2}v_k, k = 1\ldots,K$.

In this paper we use bagging to improve the estimation of the e.d.r. basis. The idea of the Bagging SIR approach is to generate B bootstrap replications of the observations, by resampling with replacement from the original data set, then estimating in each bootstrap sample the covariance matrix and finally combining them by averaging. We propose four versions of Bagging-SIR which differ in the way the matrices or directions of interest are combined.

In Section 3.3.2 we recall the sample version of SIR and then describe the first version of Bagging-SIR based on the mean of bootstrap covariance matrices M_s. We give the asymptotic distribution of the Bagging-SIR estimator. In Section 3.3.3 we propose three alternative versions of Bagging-SIR. Then a simulation study is carried out in Section 3.3.4 in order to show the numerical performances of the proposed approaches and to compare them with each other and with classical SIR. Finally concluding remarks and extensions are given and discussed in Section 3.3.5.

3.3.2 A first version of Bagging-SIR

We describe a first version of Bagging-SIR, named Bagging-I, and then give the asymptotic distribution of the corresponding estimator. Finally we discuss the choice of dimension K of the model.

3.3.2.1 Sample version of SIR and Bagging-I

Let $S = \{(y_i, \mathbf{x}_i'), i = 1, \ldots, n\}$ be a sample from the reference model (3.5). We first briefly recall how the classical SIR method may be used to estimate a basis $g_1, \ldots, g_K$ that spans the e.d.r. space and then we detail the Bagging-I approach.

Sample version of SIR. Let T denote the slicing of the y_i's into H fixed slices, $s_1, \ldots, s_H$, where $H > K$. With this transformation, the matrix M_s is given by :

$$M_s = \sum_{h=1}^{H} p_h m_{s,h} m_{s,h}',$$

where $p_h = P(y \in s_h)$ and $m_{s,h} = \mathbb{E}(\mathbf{z}|y \in s_h)$. The empirical mean and covariance matrix of the $\mathbf{x}_i$'s are respectively given by $\overline{\mathbf{x}} = n^{-1}\sum_{i=1}^{n}\mathbf{x}_i$ and $\widehat{\Sigma} = n^{-1}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})'$. We standardize the $\mathbf{x}_i$'s, $\mathbf{z}_i = \hat{\Sigma}^{-1/2}(\mathbf{x}_i - \overline{\mathbf{x}})$. The covariance matrix M_s is estimated by :

$$\widehat{M_s} = \sum_{h=1}^{H} \hat{p}_h \hat{m}_{s,h} \hat{m}_{s,h}',$$

where $\hat{p}_h = n^{-1}\sum_{i=1}^{n}\mathbb{I}_{[y_i \in s_h]}$ and $\hat{m}_{s,h} = (n\hat{p}_h)^{-1}\sum_{i=1}^{n}\mathbf{z}_i\mathbb{I}_{[y_i \in s_h]}$, where the notation $\mathbb{I}$ designates the indicator function. The first K eigenvectors $\hat{g}_{s,k}, k = 1 \ldots, K$ of the matrix $\widehat{M_s}$ are estimated standardized e.d.r. directions. Then $\hat{g}_k = \hat{\Sigma}^{-1/2}\hat{g}_{s,k}, k = 1, \ldots, K$ are estimated e.d.r. directions.

Sample version of Bagging-I. Let B denote the number of bootstrap replications and let $S^{*(b)} = \{(y_i^{*(b)}, \mathbf{x}_i^{*(b)'}), i = 1, \ldots, n\}$, for $b = 1, \ldots, B$, be a nonparametric bootstrap sample. The empirical mean and covariance matrix of the $\mathbf{x}_i^{*(b)'}$s are respectively given by $\overline{\mathbf{x}}^{*(b)} = n^{-1}\sum_{i=1}^{n}\mathbf{x}_i^{*(b)}$ and $\widehat{\Sigma}^{*(b)} = n^{-1}\sum_{i=1}^{n}(\mathbf{x}_i^{*(b)} - \overline{\mathbf{x}}^{*(b)})(\mathbf{x}_i^{*(b)} - \overline{\mathbf{x}}^{*(b)})'$. We compute the standardized $\mathbf{x}_i^{*(b)}$'s, $\mathbf{z}_i^{*(b)} = (\widehat{\Sigma}^{*(b)})^{-1/2}(\mathbf{x}_i^{*(b)} - \overline{\mathbf{x}}^{*(b)})$. In each bootstrap sample $S^{*(b)}$, let $T^{(b)}$ be the slicing of the $y_i^{*(b)}$'s into H fixed slices, $s_1^{(b)}, \ldots, s_H^{(b)}$. We compute the matrix $\widehat{M_s}^{*(b)}$ defined by :

$$\widehat{M_s}^{*(b)} = \sum_{h=1}^{H} \hat{p}_h^{*(b)} (\hat{m}_{s,h}^{*(b)})(\hat{m}_{s,h}^{*(b)})',$$

where $\hat{p}_h^{*(b)} = n^{-1}\sum_{i=1}^n \mathbb{I}_{[y_i^{*(b)} \in s_h^{(b)}]}$ and $\hat{m}_{s,h}^{*(b)} = \left(n\hat{p}_h^{*(b)}\right)^{-1}\sum_{i=1}^n \mathbf{z}_i^{*(b)}\mathbb{I}_{[y_i^{*(b)} \in s_h^{(b)}]}$. We compute the mean of these B covariance matrices,

$$\widehat{M}_{s,B}^* = \sum_{b=1}^{B} \widehat{M}_s^{*(b)}/B,$$

which is the Bagging-I estimator of the SIR matrix of interest M_s. The eigenvectors $\hat{g}_{s,k}^*, k = 1\ldots,K$ associated with the K largest eigenvalues of the matrix $\widehat{M}_{s,B}^*$ are the Bagging-I estimators of standardized e.d.r. directions. Finally $\hat{g}_k^* = \widehat{\Sigma}^{-1/2}\hat{g}_{s,k}^*, k = 1\ldots,K$ are the estimated e.d.r. directions.

3.3.2.2 Asymptotic theory

We focus here on asymptotics for the Bagging-I estimator $\widehat{M}_{s,B}^*$ of the SIR matrix of interest M_s. In the sequel, the notation $Z_n \longrightarrow_d Z$ means that Z_n converges in distribution to Z as $n \to \infty$ and the notation $Z^* \longrightarrow_{D*} \widehat{Z}$ denotes bootstrap convergence. Let $D_1 \otimes D_2$ denote the Kronecker product of the matrices D_1 and D_2 (see for instance Harville, 1999 for some useful properties of the Kronecker product). Let $D = [d_1, \ldots, d_q]$ be a $(p \times q)$ matrix, where the d_k's are p-dimensional column vectors. The notation $\mathrm{vec}(D)$ is used to designate the pq-dimensional column vector defined by $\mathrm{vec}(D) = (d_1', \ldots, d_K')'$. Let F be a $(p \times p)$ symmetric matrix. The notation $\mathrm{vech}(F)$ designates the $\frac{p(p+1)}{2}$-dimensional column vector obtained by stacking the columns of F where the "supradiagonal" elements have been previously eliminated.

For $i = 1, \ldots, n$, let $Q_i = Q(y_i, \mathbf{x}_i) = (R(y_i, \mathbf{x}_i)', \mathrm{vech}((\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)')', (\mathbf{x}_i - \mu)')'$, with $R(y_i, \mathbf{x}_i) = ((\mathbf{x}_i - \mu)'\mathbb{I}_{[y_i \in s_1]}, \ldots, (\mathbf{x}_i - \mu)'\mathbb{I}_{[y_i \in s_H]}, \mathbb{I}_{[y_i \in s_1]}, \ldots, \mathbb{I}_{[y_i \in s_H]})'$. Let $\overline{Q}_n = n^{-1}\sum_{i=1}^n Q_i$. By the Central Limit Theorem, we get :

$$\sqrt{n}(\overline{Q}_n - \mathbb{E}(Q)) \to_d U_0 \sim \mathcal{N}(0, \Gamma_{U_0} = \mathbb{V}(Q)). \tag{3.6}$$

Details on $\mathbb{E}(Q)$ and Γ_{U_0} can be found in Saracco (1997) or Barrios and Velilla (2007).

We define the function G_1 from $\mathbb{R}^{pH+H+\frac{p(p+1)}{2}+p}$ to $\mathbb{R}^{p^2}$ by :

$$G_1(u) = \mathrm{vec}((C - cc')^{-1/2}A(C - cc')^{-1/2}),$$

with $u = (\alpha', \beta', \mathrm{vech}(C)', c')'$, where $\alpha = (\alpha_1', \ldots, \alpha_H')'$ is formed by H vectors $\alpha_h \in \mathbb{R}^p, h = 1, \ldots, H$, $\beta = (\beta_1, \ldots, \beta_H)' \in \mathbb{R}^H$, $c \in \mathbb{R}^p$, C is a $p \times p$ positive definite symmetric matrix and $A = \sum_{h=1}^H \beta_h[(\frac{\alpha_h}{\beta_h}) - c][(\frac{\alpha_h}{\beta_h}) - c]'$. Note that $G_1(\overline{Q}_n) = \mathrm{vec}(\widehat{M}_s)$ and $G_1(\mathbb{E}(Q)) = \mathrm{vec}(M_s)$.

We then apply Delta method and get :

$$\sqrt{n}(\mathrm{vec}(\widehat{M}_s) - \mathrm{vec}(M_s)) \to_d U_1 \sim \mathcal{N}(0, \Gamma_{U_1} = D_{G_1}\Gamma_{U_0}D_{G_1}'),$$

where D_{G_1} is the Jacobian matrix of G_1 evaluated at the point $\mathbb{E}(Q)$. Details on the calculation of this matrix can be found in Saracco (1997) or Barrios and Velilla (2007).

For a bootstrap replication $\{(y_i^*, \mathbf{x}_i^{*\prime}), i = 1, \ldots, n\}$ obtained by resampling with replacement from the original data, let $Q_i^* = Q(y_i^*, \mathbf{x}_i^*)$ and $\overline{Q}_n^* = n^{-1} \sum_{i=1}^n Q_i^*$. We have $G_1(\overline{Q}_n^*) = \text{vec}(\widehat{M}_s^*)$. From Barrios and Velilla (2007), we have :

$$\sqrt{n}(\overline{Q}_n^* - \overline{Q}_n) \rightarrow_{D^*} U_0 \quad \text{and} \quad \sqrt{n}(\text{vec}(\widehat{M}_s^*) - \text{vec}(\widehat{M}_s)) \rightarrow_{D^*} U_1. \tag{3.7}$$

Let $1_B = (1, \ldots, 1)'$ be the B-dimensional column vector of ones. Remark that

$$\sqrt{n}(\text{vec}(\widehat{M}_{s,B}^* - \widehat{M}_s)) = \frac{1}{B}(1_B' \otimes I_{p^2}) \begin{pmatrix} \sqrt{n}(\text{vec}(\widehat{M}_s^{*(1)}) - \text{vec}(\widehat{M}_s)) \\ \vdots \\ \sqrt{n}(\text{vec}(\widehat{M}_s^{*(B)}) - \text{vec}(\widehat{M}_s)) \end{pmatrix}.$$

Given the data $\{(y_i, \mathbf{x}_i'), i = 1, \ldots, n\}$, the B bootstrap samples $S^{*(b)}$, $b = 1, \ldots, B$, are independent. Thus we have :

$$\begin{pmatrix} \sqrt{n}(\text{vec}(\widehat{M}_s^{*(1)}) - \text{vec}(\widehat{M}_s)) \\ \vdots \\ \sqrt{n}(\text{vec}(\widehat{M}_s^{*(B)}) - \text{vec}(\widehat{M}_s)) \end{pmatrix} \longrightarrow_{D^*} \mathcal{N}(0, I_B \otimes \Gamma_{U_1}).$$

Then, we get :

$$\sqrt{n}(\text{vec}(\widehat{M}_{s,B}^* - \widehat{M}_s)) \longrightarrow_{D^*} \mathcal{N}(0, \Gamma_V).$$

where $\Gamma_V = \frac{1}{B^2}(1_B' \otimes I_{p^2})(I_B \otimes \Gamma_{U_1})(1_B' \otimes I_{p^2})' = \frac{1}{B^2}(1_B' I_B 1_B) \otimes (I_{p^2} \Gamma_{U_1} I_{p^2}) = \frac{1}{B}\Gamma_{U_1}$.

3.3.2.3 Discussion on the choice of K

Until now we have supposed that the dimension K of the reduction model is known. However, in most applications, this dimension is unknown and must be estimated from the data. Several approaches have been proposed in the literature. The first approaches were hypothesis tests based on the nullity of some eigenvalues, see Li (1991) for a normal distribution of $\mathbf{x}$ and Schott (1994) for an elliptically symmetric distribution. Another approach is that of Ferré (1997, 1998) which is based on the quality of the estimations of the e.d.r. space. Two versions exist for calculating these quantities : a consistent estimator for the "classical" SIR and a cross-validation procedure for the SIR version of Hsing and Carroll (1992). One can also use the bootstrap method of Barrios and Velilla (2007), which combines formal and graphical inference procedures for choosing the dimension of a general regression problem. Another procedure is that proposed by Liquet and Saracco (2008). By a bootstrap estimation of the square trace correlation between the true e.d.r. space and its estimate, it makes it possible to choose in the same step dimension K of the model and parameter α for SIR_α (an extension of the classical SIR method). In practice we advise first choosing dimension $\widehat{K}$ from the original data set via one of the above-mentionned methods. Then we perform the Bagging-SIR approach using dimension $\widehat{K}$.

3.3.3 Alternative versions of Bagging-SIR

In this section, we present three other versions of Bagging-SIR, respectively named Bagging-II, Bagging-III and Bagging-IV, which differ in the strategy of aggregating either the covariance matrices or the (standardized or not) e.d.r. directions.

Bagging-II. This version can be viewed as an unstandardized version of Bagging-I. In each bootstrap sample $S^{*(b)}$, we compute the matrix $\widehat{M}^{*(b)}$ which is defined as $\widehat{M}^{*(b)} = \sum_{h=1}^{H} \hat{p}_h^{*(b)} (\hat{m}_h^{*(b)} - \overline{\mathbf{x}}^{*(b)})(\hat{m}_h^{*(b)} - \overline{\mathbf{x}}^{*(b)})'$ where $\hat{p}_h^{*(b)} = n^{-1} \sum_{i=1}^{n} \mathbb{I}_{[y_i^{*(b)} \in s_h^{(b)}]}$ and $\hat{m}_h^{*(b)} = \left(n\hat{p}_h^{*(b)}\right)^{-1} \sum_{i=1}^{n} \mathbf{x}_i^{*(b)} \mathbb{I}_{[y_i^{*(b)} \in s_h^{(b)}]}$. We calculate the mean of these B covariance matrices $\widehat{M}_B^* = \sum_{b=1}^{B} \widehat{M}^{*(b)}/B$. The eigenvectors $\hat{g}_k^*, k = 1 \ldots, K$ associated with the K largest eigenvalues of the matrix $\hat{\Sigma}^{-1}\widehat{M}_B^*$ are the estimated e.d.r. directions.

For simplicity, the two following versions of Bagging-SIR are described for a single index model ($K = 1$). Note that these versions can be easily extended to multiple-index models (see the simulation study in Section 3.3.4 for details).

Bagging-III. In each bootstrap sample $S^{*(b)}$, we compute the matrix $\widehat{M}_s^{*(b)}$ defined in the first version of Bagging-SIR. The eigenvector $\hat{g}_s^{*(b)}$ associated with the largest eigenvalue of the matrix $\widehat{M}_s^{*(b)}$ is the standardized estimated e.d.r. direction in the bth bootstrap sample. We construct the matrix $\widehat{G}_s^* = [\hat{g}_s^{*(1)}, \ldots, \hat{g}_s^{*(B)}]$. Note that if we consider matrix $G_s = [g_s, \ldots, g_s]$, then the major eigenvector of $G_s G_s'$ is g_s. Thus we consider the major eigenvector $\hat{g}_s^*$ of the matrix $\widehat{G}_s^*(\widehat{G}_s^*)'$ as the estimated standardized e.d.r. direction. Finally we compute the estimated e.d.r. direction as $\hat{g}^* = \hat{\Sigma}^{-1/2}\hat{g}_s^*$.

Bagging-IV. This version can be viewed as an unstandardized version of Bagging-III. In each bootstrap sample $S^{*(b)}$, we compute the matrix $\widehat{M}^{*(b)}$ defined in the second version of Bagging-SIR. The eigenvector $\hat{g}^{*(b)}$ associated with the largest eigenvalue of the matrix $(\widehat{\Sigma}^{*(b)})^{-1}\widehat{M}^{*(b)}$ is the estimated e.d.r. direction in the bth bootstrap sample. We construct the matrix $\widehat{G}^* = [\hat{g}^{*(1)}, \ldots, \hat{g}^{*(B)}]$. Thus we consider the major eigenvector $\hat{g}^*$ of the matrix $\widehat{G}^*(\widehat{G}^*)'$ as the estimated e.d.r. direction in model (3.5).

3.3.4 Numerical comparisons via a simulation study

In this section we perform a simulation study with `R` to illustrate the behaviour of the proposed versions of Bagging-SIR and to compare them with SIR. All the source codes are available from the authors by E-mail. First we introduce the efficiency measure which will be used to compare the performances of the methods. Then, for a single index model ($K = 1$), we study the impact of the sample size, the variance of the error term of the model and the number of bootstrap samples, on the performances of the Bagging-SIR approaches. Finally we illustrate the benefits of the proposed approaches for a multiple-index model where $K = 2$. Note that bagging versions of SIR are not

computationally slow because they are based on multiple uses of SIR, thereby giving a very fast estimation method.

3.3.4.1 Efficiency measure

Let $\check{g}_1, \ldots, \check{g}_K$ be the K estimated e.d.r. directions. We designate $\check{G} = [\check{g}_1, \ldots, \check{g}_K]$ and $\check{E} = \text{Span}(\check{G})$, the linear subspace spanned by the $\check{g}_k$'s. Let $G = [\gamma_1, \ldots, \gamma_K]$ be the matrix of the true directions and let $E = \text{Span}(G)$. Let P_E (resp. $P_{\check{E}}$) be the Σ-orthogonal projector onto E (resp. $\check{E}$) defined as follows : $P_E = G(G'\Sigma G)^{-1}G'\Sigma$ and let $P_{\check{E}} = \check{G}(\check{G}'\Sigma\check{G})^{-1}\check{G}'\Sigma$.
The quality of the estimate $\check{E}$ of E is measured by :

$$m_\Sigma(E, \check{E}) = \text{Trace}(P_E P_{\check{E}})/K.$$

This efficiency measure is that used by Ferré (1998) in order to determine the dimension in SIR and related methods. It belongs to $[0, 1]$ with $m_\Sigma(E, \check{E}) = 0$ if $\check{E}$ and E are Σ-orthogonal and $m_\Sigma(E, \check{E}) = 1$ if $\check{E} = E$. Therefore the closer this value is to one, the better is the estimation. When $K = 1$ (single index model), this measure is the squared cosine of the angle formed by the vectors γ and $\check{g}$.

3.3.4.2 First simulation model : a single index model

We generate simulated data from the following regression model (in which the true dimension K is equal to 1) :

$$y = \exp(\mathbf{x}'\gamma) + \varepsilon, \tag{3.8}$$

where ε is normally distributed with variance δ^2 and $\mathbf{x}$ follows an 8-dimensional standardized normal distribution. The error term ε is independent of $\mathbf{x}$. We consider $\gamma = (1, -1, 0, 0, 0, 0, 0, 0)'$.

Bagging-SIR approaches have been implemented with the number of bootstraps $B = 200$. The impact of the choice of B is studied at the end of this section.

Effect of the sample size. We compare SIR and the four Bagging-SIR approaches on $N = 500$ data replications of size $n = 50, 100, 200$ in model (3.8) where $\delta = 2$. The top of Figure 3.1 shows the boxplots of the squared cosines for $n = 50$ and $\delta = 2$ based on the $N = 500$ data replications. The Bagging approaches seem to perform better than SIR. At the bottom of Figure 3.1, we compare the methods two by two on these 500 simulated samples. The plots on the first line of these graphics give the squared cosines obtained with SIR versus those obtained with the Bagging-SIR approaches. All the Bagging-SIR approaches seem to perform better than the SIR method with a majority of points on the right of the first bisecting line. The plots of the squared cosines of a Bagging-SIR approach versus another Bagging-SIR approach show that these methods give very similar results. The four Bagging-SIR methods provide very similar results of a quality higher than that obtained with the SIR approach, and there is no method that is uniformly better than the others.

Figure 3.2 shows the boxplots of the squared cosines for different sizes of n (50,100, 200) for the SIR approach and Bagging-II. Similar results have been observed with the three other Bagging-SIR approaches. The benefits of Bagging-SIR are greater with small sample sizes. For $n = 50$, the median quality obtained with Bagging-II is 0.83 versus 0.78 with classical SIR. The most interesting feature of Bagging-SIR concerns the range of the boxplots, which is smaller than with classical SIR approaches.

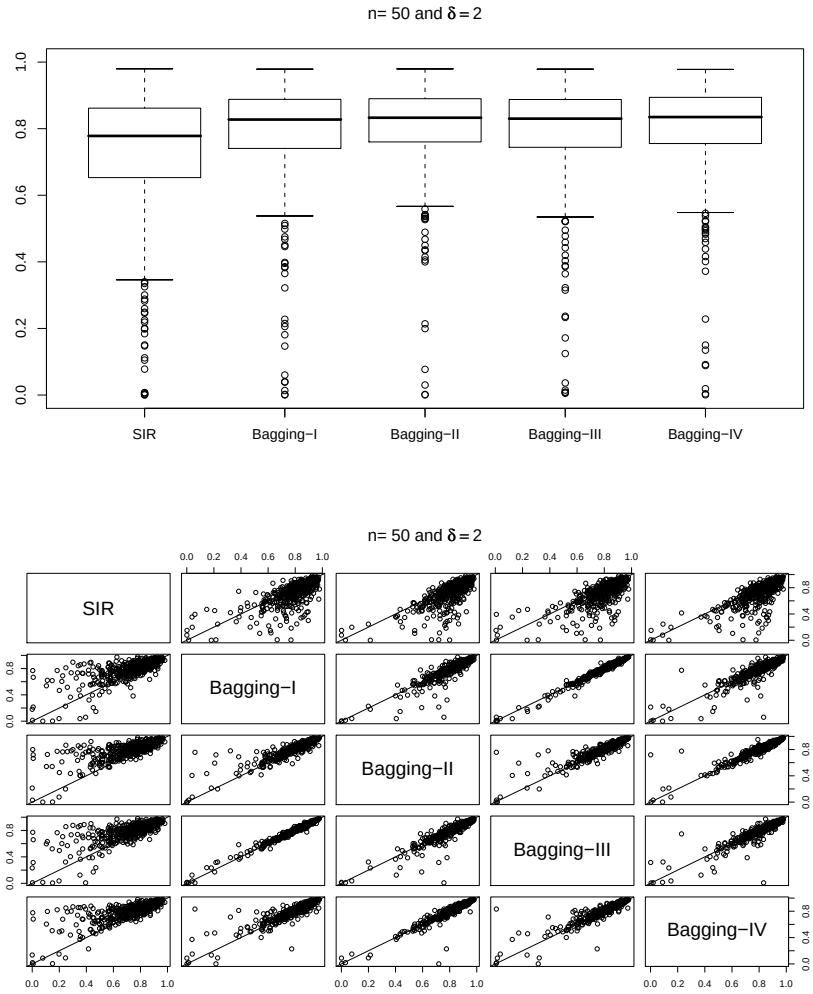

FIG. 3.1 – Comparison of Bagging-SIR and SIR methods for $\delta = 2$, $n = 50$ and $B = 200$ (Boxplots of the squared cosines at the top and scatterplots at the bottom)

Increasing the variance of the error term. We now compare SIR and Bagging-II on $N = 500$ data replications of size $n = 50$ in model (3.8) with δ belonging to the set $\{0.5, 1, 1.5, 2, 2.5, 3\}$. Figure 3.3 shows the boxplots of the squared cosines based on

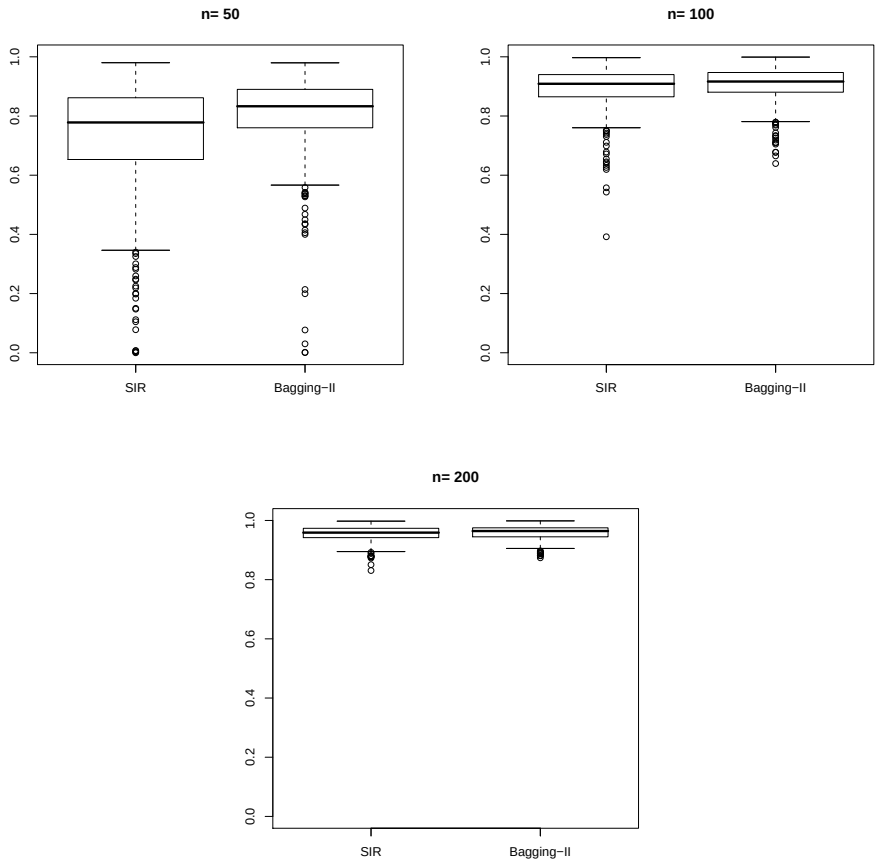

FIG. 3.2 – Boxplots of the squared cosines obtained with SIR and Bagging-II for $\delta = 2$, $B = 200$ and $n = 50, 100, 200$.

the $N = 500$ data replications for the different values of δ. As the variation included in the model increases, the performances of both methods decrease. The benefits of using Bagging-II become more obvious as the variance δ^2 of the error term ε increases. When $\delta = 3$, the median of the squared cosines is about 0.71 with Bagging-II as opposed to 0.63 with SIR, and bagging considerably reduces the range of the boxplot. Thus, Bagging-II seems to be less sensitive to high variations in the model. Similar results may be observed with the other Bagging-SIR approaches.

Impact of the number B of bootstraps. We now study the impact of the number B of bootstrap replications on the performance of Bagging-II. Parameter B will belong to the set $\{10, 20, 30, 50, 70, 100, 200, 500, 1000\}$. We again consider model (3.8) where $\delta = 2$ and $n = 50$. Figure 3.4 shows the boxplots of the quality measure obtained with SIR and Bagging-II for the different values of B. For almost all bootstrap replication numbers (except $B = 10$), Bagging-SIR is more efficient than SIR. The performance

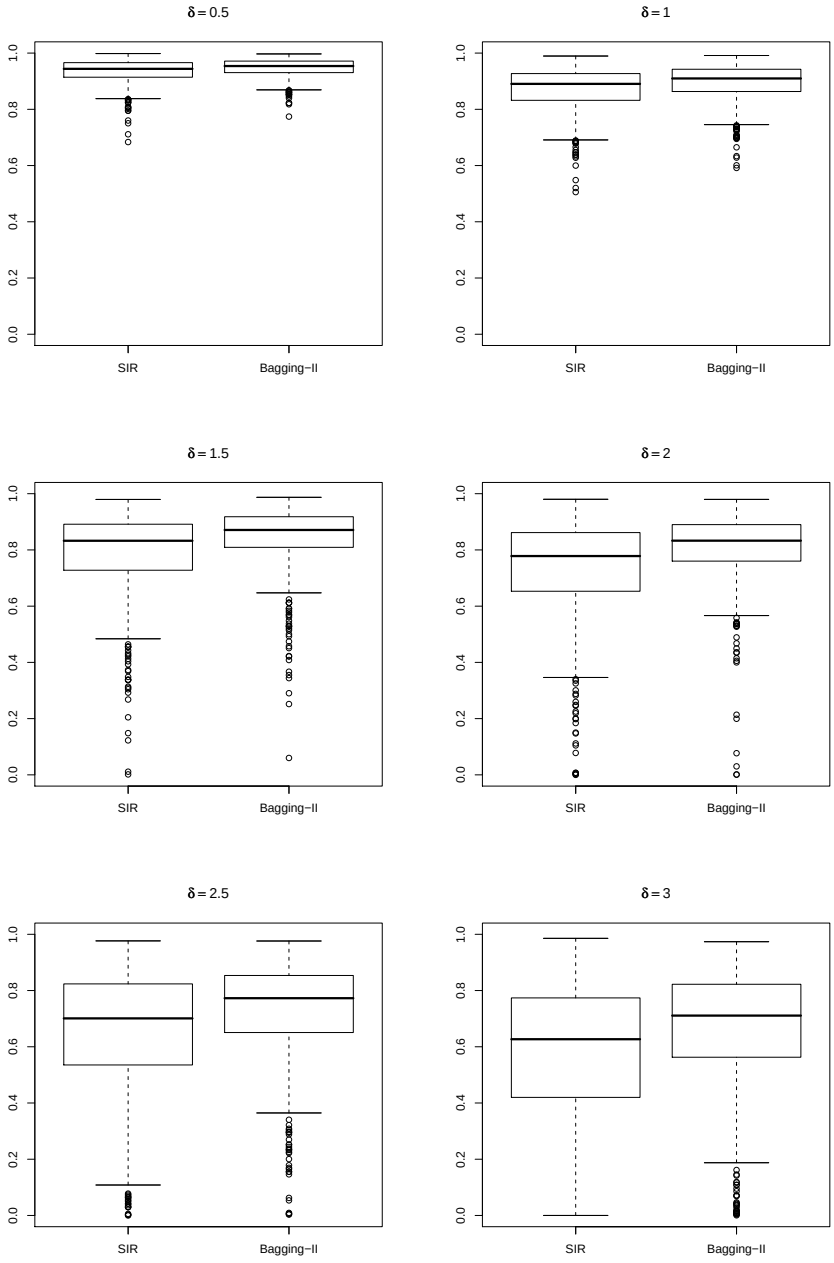

FIG. 3.3 – Boxplots of the squared cosines for $n = 50$, $B = 200$ and $\delta = 0.5, 1, 1.5, 2, 2.5, 3$.

of Bagging-SIR seems to increase until $B = 200$. Moreover, quality does not seem to be improved by a large number of replications ($B = 500$ or 1000). Note that for other simulations, the results may slightly differ but a sufficient number of bootstraps seems

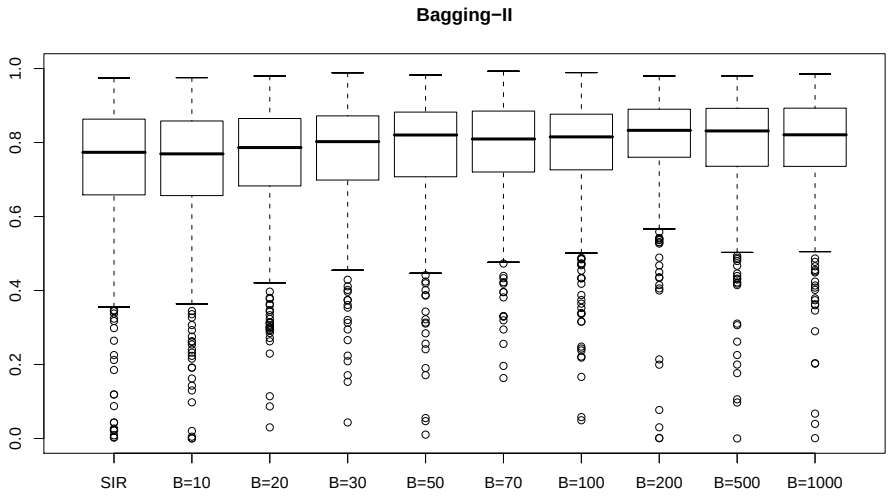

FIG. 3.4 – Boxplots of the squared cosines obtained with SIR and Bagging-II for $n = 50$, $\delta = 2$ and for various values of B.

to be 100 or 200. In the rest of the simulation study, we choose $B = 200$.

3.3.4.3 Second simulation model : multiple-index model

In this section we briefly demonstrate the satisfactory behavior of the Bagging-SIR approaches when the dimension of the reduction model is equal to 2 ($K = 2$). For Bagging-III and Bagging-IV, a slight modification of the approach described in Section 3 is necessary for the two-index model : the matrices $\widehat{G}_s^*$ and $\widehat{G}^*$ are now computed by binding in columns the two estimated e.d.r. directions for the B bootstrap samples. Thus $\widehat{G}_s^*$ and $\widehat{G}^*$ are $(p \times 2B)$ matrices. Then the estimated e.d.r. directions are the two eigenvectors associated with the two largest eigenvalues of the matrices $\widehat{G}_s^*(\widehat{G}_s^*)'$ or $\widehat{G}^*(\widehat{G}^*)'$. We generate simulated data from the following regression model :

$$y = (\mathbf{x}'\gamma_1)\exp(\mathbf{x}'\gamma_2) + \varepsilon, \tag{3.9}$$

where ε is normally distributed with variance δ^2 and $\mathbf{x}$ follows an 8-dimensional standardized normal distribution. The error term ε is independent of $\mathbf{x}$. We take $\gamma_1 = (1,1,0,0,0,0,0,0)'$ and $\gamma_2 = (0,0,0,1,1,0,0,0)'$.

We then compare SIR and the four Bagging-SIR approaches on $N = 500$ data replications of size $n = 100$ in model (3.9) where $\delta = 2.5$. The top of Figure 3.5 shows the boxplots of the quality measure $m_\Sigma(E, \check{E})$ for $n = 100$ and $\delta = 2.5$ based on the $N = 500$ data replications. At the bottom of Figure 3.5, we compare the methods two by two on these 500 simulated samples. The plots on the first line of this graphic represent the quality measures obtained with SIR versus those obtained with the Bagging-SIR

approaches. All the Bagging-SIR approaches seem to perform better than the SIR method, with a majority of points on the right of the first bisecting line. Moreover, all four bagging approaches seem to give similar results which are superior to the classical technique.

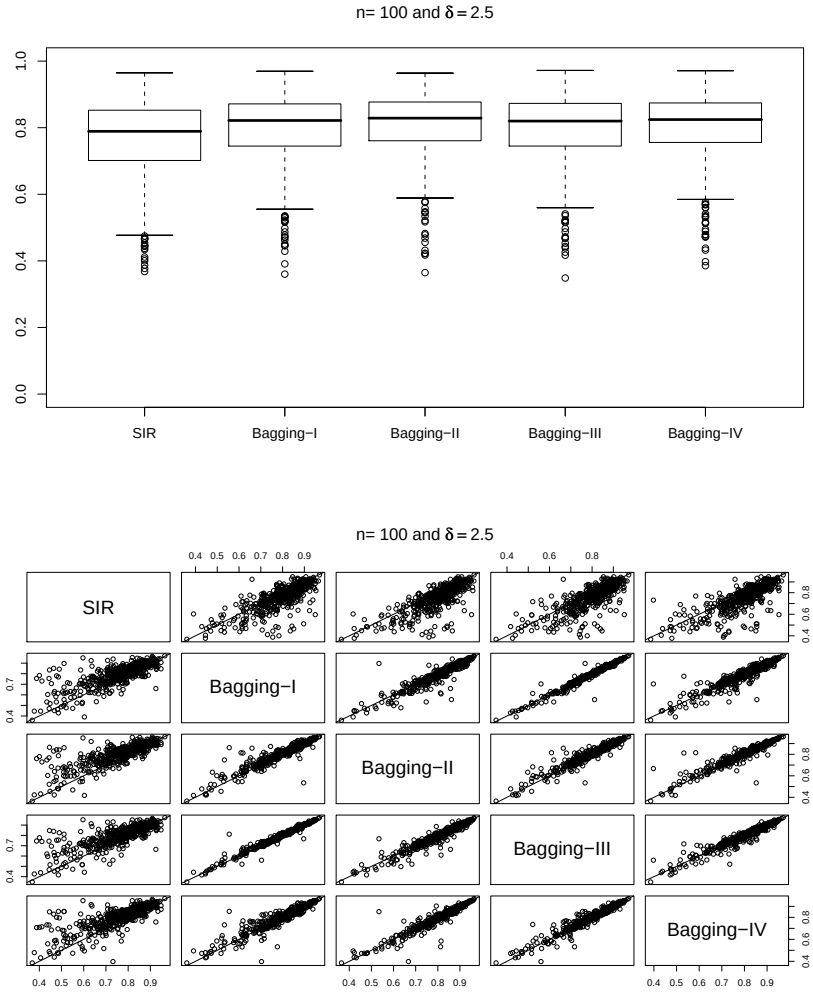

FIG. 3.5 – Comparison of Bagging-SIR and SIR methods where $\delta = 2$, $B = 200$ and $n = 50$ (Boxplots of quality measures at the top and scatterplots at the bottom)

3.3.5 Concluding remarks

In this paper, we propose bagging versions of the classical SIR method. We provide asymptotic results for one of them (Bagging-I), and asymptotics for the other methods could be obtained by using similar arguments. The numerical performance of these new approaches is illustrated in a simulation study. Importantly, Bagging-SIR always pro-

vides equivalent or better estimation than SIR. This is particularly true when the sample size is small (lower than 100) and when the underlying structural relationship between the dependent variable and the indices $\mathbf{x}'\gamma_k$ is relatively noisy. In data analysis, instead of adhering to SIR, we recommend closely comparing the estimation results obtained with SIR and Bagging-SIR and preferring Bagging-SIR if a difference is observed. The bagging versions of SIR have been implemented in `R` and the source codes are available from the authors by E-mail.

It is known that SIR fails to work in the pathological case of symmetric regressions with $y = f(\gamma'\mathbf{x}) + \varepsilon$, where f is a symmetric function of the argument $\gamma'\mathbf{x}$. This is theoretically due to the fact that SIR is only based on the estimation of the conditional mean of $\mathbf{x}$ given y. In order to deal with this problem, several methods have been developed using the estimation of the conditional variance function of the covariates given the response. For instance, there are SAVE (sliced average variance estimation), SIR-II and SIR$_\alpha$ which combines SIR and SIR-II. More details on these methods can be found in Li (1991), Cook and Weisberg (1991), Cook (2000) or Gannoun and Saracco (2003). One of their drawbacks is that the performance of the corresponding slicing estimator is sensitive to the choice of the number of slices (see Li and Zhu, 2007) unlike SIR which is insensitive. It is clearly possible and potentially interesting to extend the proposed bagging approaches of SIR to SAVE, SIR-II or SIR$_\alpha$ methods.

Moreover, multivariate extensions of SIR have been proposed in the literature, see for instance Aragon (1997), Li et al. (2003), Saracco (2005) or Barreda et al. (2007). Bagging versions of the corresponding estimators could be proposed and might become an alternative to the classical slicing approach.

Acknowledgment

The authors are very grateful to the editor and the referee for their valuable comments and constructive suggestions.

References

Aragon, Y. (1997). A Gauss implementation of multivariate sliced inverse regression. *Computational Statistics*, **12**(3), 355-372.

Barreda, L. Gannoun, A., Saracco, J. (2007). Some extensions of multivariate sliced inverse regression. *Journal of Statistical Computation and Simulation*, **77**(1-2), 1-17.

Barrios, M.P., Velilla, S. (2007). A bootstrap method for assessing the dimension of a general regression problem. *Statistics & Probability Letters*, **77**(3), 247-255.

Beran, B.R., Srivastava, M.S. (1985). Bootstrap tests and confidence regions for functions of a covariance matrix. *The Annals of Statistics*, **13**(1), 95-115.

Breiman, L. (1994). Bagging Predictors. *Technical Report No. 421, Department of Statistics University of California.*

Breiman, L. (1996). Bagging Predictors. *Machine Learning*, **24**, 123-140.

Bühlmann, P. (2004). Bagging, boosting and ensemble methods. *Handbook of Computational Statistics*, 877-907, Springer, Berlin.

Cook, R.D., Weisberg, S. (1991). Comment on "Sliced Inverse Regression for Dimension Reduction", K.C. Li, *Journal of the American Statistical Association*, **86**, 328-332.

Cook, R.D. (2000). SAVE : A method for dimension reduction and graphics in regression. *Communications in Statistics - Theory and Methods*, **29**, 2109-2121.

Efron, B., Tibshirani, R. J. (1979). Bootstrap methods : another look at the jackknife. *The Annals of Statistics*, **7**(1), 1-26.

Efron, B., Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*, Chapman & Hall.

Ferré, L. (1997). Dimension choice for sliced inverse regression based on ranks. *Student*, **2**, 95-108.

Ferré, L. (1998). Determining the dimension in Sliced Inverse Regression and related methods. *Journal of the American Statistical Association*, **93**(441), 132-140.

Gannoun, A., Saracco, J. (2003). An asymptotic theory for SIR_α method. *Statistica Sinica*, **13**(2), 297-310.

Harville, D.A. (1997). *Matrix algebra from a statistician's perspective*, Springer-Verlag.

Hsing, T., Carroll, R.J. (1992). An asymptotic theory for Sliced Inverse Regression. *The Annals of Statistics*, **20**(2), 1040-1061.

Li, K.C. (1991). Sliced inverse regression for dimension reduction, with discussion. *Journal of the American Statistical Association*, **86**, 316-342.

Li, K.C., Aragon, Y., Shedden, K., Thomas Agnan, C. (2003). Dimension reduction for multivariate response data. *Journal of the American Statistical Association*, **98**(461), 99-109.

Li, Y., Zhu, L. (2007). Asymptotics for sliced average variance estimation. *The Annals of Statistics*, **35**(1), 41-69.

Liquet, B., Saracco, J. (2008). Application of the bootstrap approach to the choice of dimension and the α parameter in the SIR_α method. *Communications in Statistics - Simulation and Computation*, **37**(6), 1198-1218.

Saracco, J. (1997). An asymptotic theory for Sliced Inverse Regression. *Communications in Statistics - Theory and Methods*, **26**(9), 2141-2171.

Saracco, J. (2005). Asymptotics for pooled marginal slicing estimator based on SIR_α approach. *Journal of Multivariate Analysis*, **96**(1), 117-135.

Schott, J.R. (1994). Determining the dimensionality in Sliced Inverse Regression. *Journal of the American Statistical Association*, **89**(425), 141-148.

Shao, J., Tu, D.S. (1995). *The Jackknife and Bootstrap*, Springer Series in Statistics, Springer-Verlag, New York.

Chapitre 4

Applications et collaborations interdisciplinaires

Dans ce chapitre je décris les thèmes annexes traités lors de ma thèse et qui résultent de collaborations universitaires ou industrielles ainsi que du doctorat-conseil. Tout d'abord, une synthèse de ces travaux est réalisée. Puis je donne certains articles publiés qui ont découlé de ces collaborations.

D'autres documents (articles ou rapports) ont aussi été produits lors de ces collaborations : ils sont très appliqués et leur contenu statistique se limite à des applications ou adaptations de méthodologies existantes. L'interprétation des résultats est la partie fondamentale de ces travaux. C'est la raison pour laquelle j'ai choisi de ne pas les insérer dans mon mémoire de thèse. Cependant je tiens ces documents à disposition pour une éventuelle lecture.

4.1 Synthèse des travaux

Pollution atmosphérique. Ce premier travail est une application autour du problème de l'identification et de la quantification de sources de poussière fine, l'impact néfaste des particules en suspension dans l'air sur la santé ayant été démontré. Cette étude de cas est une réponse à un appel à projet lancé dans le cadre d'un programme de recherche scientifique piloté par le Ministère de l'Ecologie, du Développement et de l'Aménagement Durable ainsi que l'Agence de l'Environnement et de la Maîtrise de l'Energie. Ce projet visant à mettre au point un outil de détermination de la contribution des sources de poussière fine a été réalisé en collaboration avec l'association de surveillance de la qualité de l'air en Aquitaine pour la campagne de prélèvement des filtres de poussière sur le site d'Anglet, le Centre d'Etudes Nucléaires de Bordeaux Gradignan pour la caractérisation des filtres par leur concentration en différents éléments chimiques et l'Institut de Mathématiques de Bordeaux pour la détermination de l'origine et de la contribution des sources de poussière fine. Ce projet s'est échelonné sur 18 mois et a pris fin en Décembre 2006. Afin d'identifier les sources à l'origine de l'empoussièrement des filtres, une Analyse en Composantes Principales (ACP) couplée à

une technique de rotation pour faciliter l'interprétation des résultats a été utilisée. Mon premier article a consisté à décrire le modèle d'Analyse en Facteurs, avec notamment l'utilisation de la rotation, afin de comprendre les justifications théoriques de la possibilité de réaliser une rotation des composantes principales en ACP. Une description des sorties de trois logiciels "référence" en statistique (SAS, SPAD et SPSS) illustre l'intérêt pratique de la rotation en ACP sur un jeu de données "école". Cet article intitulé "Analyse en Facteurs : présentation et comparaison des logiciels SAS, SPAD et SPSS", écrit en collaboration avec Marie Chavent et Jérôme Saracco, a été publié dans la revue *Modulad*. Il est disponible dans la section 4.2. L'approche proposée nous a ainsi permis de mettre en lumière des groupes d'éléments corrélés et ainsi d'identifier la présence d'une source dans le cas où ces éléments étaient par ailleurs les traceurs connus d'une même source.

Dans un second temps, une approche de type récepteur nous a permis de quantifier l'émission des sources dans le phénomène d'empoussièrement global. La méthodologie que nous avons développée pour répondre à la problématique a donné lieu à deux articles écrits en collaboration avec Marie Chavent, Hervé Guégan, Brigitte Patouille et Jérôme Saracco. Le premier intitulé "PCA and PMF based methodology for air pollution sources identification and apportionment" va paraître dans *Environmetrics* et est disponible dans la section 4.3. Le second article intitulé "Apportionment of Air Pollution by Source at a French Urban Site" est publié dans la revue *Case Studies in Business, Industry and Government Statistics*, qui met l'accent sur le développement de nouvelles méthodologies en analyse des données pour traiter des études de cas réelles. Cet article est disponible sur internet à l'adresse suivante : `http://www.bentley.edu/csbigs/documents/chavent.pdf`.

Comptabilité financière. La deuxième application a consisté en une collaboration avec Elisabeth Walliser, Maître de Conférences en Sciences de Gestion à la Faculté de Sciences Economiques de Montpellier, et Corinne Bessieux Ollier, professeure associée au Groupe Sup de Co Montpellier. Ce travail visait à étudier l'impact de l'adoption en France d'une nouvelle loi en comptabilité financière sur les incorporels. Ce travail a duré 6 mois pour s'achever en Novembre 2008. Tout d'abord, je décris l'approche de classification hiérarchique descendante de variables quantitatives que nous avons utilisée et qui a remplacé l'étape classique d'ACP préalable à la classification des observations. Une procédure pour le choix du nombre de classes de variables a également été proposée et est présentée ci-dessous.

Pour la classification de variables quantitatives, la procédure VARCLUS du logiciel SAS est probablement la plus connue et utilisée. Cet algorithme est à la fois divisif et itératif et produit une hiérarchie ou une partition de classes de variables quantitatives. Même si cette méthode fonctionne bien en général, de nombreuses options sont possibles (initialisation, choix de la classe à couper, règles d'arrêt, etc.) conduisant à des types de résultats différents. Ainsi il n'est pas évident de procéder à des choix cohérents et justifiés théoriquement. Dans cet objectif, nous avons développé sous `R` une version simplifiée de cette méthode, avec notamment une procédure différente pour le choix de la classe à diviser.

Soit $\mathbf{X}$ une matrice de données de dimension (n, p) où un ensemble de n objets sont décrits sur un ensemble de p variables quantitatives. Soit $\mathcal{V} = \{\mathbf{x_1}, \ldots, \mathbf{x_p}\}$ l'ensemble des p colonnes de $\mathbf{X}$, appelées pour plus de simplicité variables quantitatives. On désigne par $\mathcal{P}_K = \{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$ une partition de $\mathcal{V}$ en K classes et par $\mathcal{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_K\}$ un ensemble de K variables latentes. Il s'agit de classifier un ensemble de p variables en K classes disjointes telles que les variables au sein d'une même classe soient le plus fortement possible corrélées (sans tenir compte du signe de la liaison). Le critère de partitionnement vise à maximiser l'homogénéité des classes de la partition :

$$H(\mathcal{P}_K) = \sum_{k=1}^{K} S(\mathcal{C}_k), \tag{4.1}$$

où $S(\mathcal{C}_k) = \sum_{\mathbf{x}_j \in \mathcal{C}_k} \mathrm{corr}^2(\mathbf{x}_j, \mathbf{y}_k)$ mesure l'adéquation entre les variables de $\mathcal{C}_k$ et la variable latente $\mathbf{y}_k$. On montre que la variable latente maximisant l'homogénéité d'une classe correspond à la première composante principale issue de l'ACP des variables de la classe. Ainsi on obtient $H(\mathcal{P}_K) = \sum_{k=1}^{K} \lambda_k$ avec λ_k la variance de la première composante principale de $\mathcal{C}_k$. Nous proposons un algorithme de classification hiérarchique descendante. La division d'une classe $\mathcal{C}_l$ en deux sous-classes $\mathcal{A}_l$ et $\bar{\mathcal{A}}_l$ est réalisée à l'aide d'un algorithme de type nuées dynamiques visant à maximiser (4.1). L'initialisation de l'algorithme est inspirée de la procédure VARCLUS. Une ACP est réalisée sur $\mathcal{C}_l$ et les deux premières composantes principales sont retenues pour jouer le rôle de variables latentes de deux classes contenant au départ elles seules. Chaque variable est ensuite affectée à la sous-classe pour laquelle sa corrélation au carré avec la variable latente est la plus forte. Afin de faciliter cette étape et d'obtenir une meilleure partition initiale, une procédure de rotation est utilisée pour obtenir des valeurs proches de 0 ou 1 dans la matrice des corrélations au carré entre les variables et les composantes principales. Pour le choix de la classe à diviser, on choisit de diviser à l'étape K la classe C_l qui fournit la meilleure partition en $K+1$ classes au sens du critère (4.1). Ce critère étant additif, cela revient à choisir de diviser la classe C_l qui maximise la variation du critère suivant $h(\mathcal{C}_l) = S(\bar{A}_l) + S(A_l) - S(\mathcal{C}_l)$. Les classes de la hiérarchie partielle ainsi construites sont indicées par $h(\mathcal{C}_l)$, le gain en homogénéité des classes obtenu grâce à la division de C_l. Cet indice est toujours positif mais pour l'instant, nous ne sommes pas parvenus à démontrer qu'il est bien monotone croissant (c'est-à-dire que si $\mathcal{A} \subset \mathcal{B}$ alors $h(\mathcal{A}) \leq h(\mathcal{B})$). Notons qu'en pratique, sur l'ensemble des jeux de données que nous avons traités, nous n'avons jamais observé d'inversion.

La procédure que nous proposons pour le choix du nombre de classes K est la suivante : pour $K = 1, \ldots, p$,

- Nous construisons à partir de la matrice $\mathbf{X}$ la partition en K classes : $\mathcal{P}_K^{(\mathbf{X})} = (\mathcal{C}_1^{(\mathbf{X})}, \ldots, \mathcal{C}_K^{(\mathbf{X})})$.
- Nous générons B échantillons bootstrap à partir de $\mathbf{X}$: $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(B)}$.
- Pour chaque échantillon $b = 1, \ldots, B$, nous calculons la partition en K classes à l'aide de l'approche divisive proposée : $\mathcal{P}_K^{(b)} = (\mathcal{C}_1^{(b)}, \ldots, \mathcal{C}_K^{(b)})$. Nous comparons les partitions $\mathcal{P}_K^{(\mathbf{X})}$ et $\mathcal{P}_K^{(b)}$ à l'aide de l'indice de Rand corrigé (voir la remarque

ci-dessous sur ce critère).

- Nous calculons ensuite la valeur moyenne de cet indice, ce qui permet d'évaluer en moyenne l'adéquation entre les B partitions, construites à partir des échantillons bootstrap, et la partition calculée sur $\mathbf{X}$.
  (a) Si les partitions sont semblables (critère de Rand moyen proche de 1), cela signifie qu'il existe une structure stable dans les données.
  (b) Si au contraire, la valeur moyenne de l'indice est faible, cela signifie qu'il n'y a pas de structure en K classes dans les données, de légères perturbations entraîne des partitions très différentes.

Remarque. Le critère de Rand mesure le pourcentage d'accords entre deux partitions ayant le même nombre de classes. Nous choisissons d'utiliser la version corrigée de cet indice qui est d'espérance nulle lorsque les accords entre les deux partitions sont dus au hasard et qui permet ainsi de mieux détecter les cas de non adéquation entre les partitions. Deux partitions seront donc d'autant plus proches que l'indice sera proche de 1. Un indice proche de zéro (pouvant également être négatif) indiquera que les partitions sont très différentes. Pour plus de détails sur ce critère, le lecteur pourra se référer à Hubert et Arabie (1985).

L'algorithme de classification hiérarchique descendante de variables quantitatives avec le choix du nombre de classes par bootstrap a remplacé l'étape classique d'ACP préalable à la classification des observations. Cette approche a permis de mettre en lumière une structure en classes des entreprises qui était masquée lors de la classification sur les scores des composantes principales des observations issus d'une ACP. Ainsi différentes stratégies de réaction des entreprises françaises face à la mise en place de cette nouvelle loi ont été détectées. Cette étude a donné lieu à un article écrit en collaboration avec Corinne Bessieux Ollier, Marie Chavent et Elisabeth Walliser, intitulé "The consequences of adopting mandatory IFRS on intangibles : French evidence". Il a été soumis dans une revue internationale de comptabilité. Cette revue étant très appliquée, la méthodologie statistique utilisée pour répondre à la problématique est peu détaillée, au contraire l'interprétation des résultats et la compréhension des phénomènes qui en ressortent sont mis en avant. Pour cette raison, je n'ai pas joint cet article à mon mémoire de thèse. Notons que ce travail a également fait l'objet de plusieurs présentations dans des congrès de comptabilité financière, en particulier lors de l'American Accounting Association à New York, un des plus importants dans la communauté.

Satisfaction des navigants plaisanciers. L'étude suivante a été réalisée en collaboration avec Laurent Morillère de l'entreprise de Marketing Enform en réponse à un appel à projet initié par Voies Navigables de France (VNF). Il s'agissait de définir la méthodologie statistique d'une enquête par sondage des navigants plaisanciers sur le Canal des Deux Mers durant l'été et l'automne 2008. En particulier, différents outils classiques de théorie des sondages, statistique descriptive et inférentielle ainsi que des tests paramétriques et nonparamétriques ont permis de mieux connaître les navigants (origine, attentes, utilisation du canal, etc.). Une classification hiérarchique ascendante

des individus a permis de dégager des profils de navigants afin de comprendre les appréciations divergentes exprimées à propos d'un service ou une offre à tous égards identiques et de fournir à VNF un outil d'aide à la décision en comprenant mieux la cause à l'origine de l'opinion exprimée. Les points forts ainsi que les priorités d'amélioration de l'offre de service proposée sur le canal ont été identifiés. Je tiens à disposition le rapport synthétique qui a été remis à VNF.

Doctorat-conseil chez Danone. Dans le cadre d'un doctorat-conseil chez Danone Research à Paris, j'ai étudié l'application de méthodes Multi-way (voir par exemple Smilde, Bro et Geladi, 2004) aux données cliniques en nutrition-santé. En raison de clauses de confidentialité, la problématique ne peut être davantage précisée. L'objectif premier de cette mission a été de comprendre la théorie sous-jacente à ces méthodes pour justifier de leur utilisation et des résultats statistiques obtenus. Dans un second temps, j'ai réalisé un état de l'art sur les méthodes multi-tableaux afin d'identifier l'approche la mieux adaptée à la problématique traitée. Ainsi l'utilisation de l'Analyse Factorielle Multiple (Escofier et Pages, 2008) s'est révélée pertinente et a permis d'obtenir des résultats justifiés théoriquement avec des interprétations intéressantes.

4.2 Analyse en Facteurs : présentation et comparaison des logiciels SAS, SPAD et SPSS

Analyse en Facteurs : présentation et comparaison des logiciels SAS, SPAD et SPSS

Marie Chavent[1], Vanessa Kuentz[1], Jérôme Saracco[1,2]

[1] Universités Bordeaux 1 et 2,
IMB, UMR CNRS 5251,
351 Cours de la Libération, 33405 Talence Cedex, France
vanessa.kuentz,marie.chavent@math.u-bordeaux1.fr

[2] Université Montesquieu - Bordeaux 4,
GREThA, UMR CNRS 5113,
Avenue Léon Duguit, 33608 Pessac Cedex, France
jerome.saracco@u-bordeaux4.fr

Abstract In data analysis, factorial methods are essential. These techniques can be used as an end in themselves, seeking to highlight underlying common factors in a group of variables. They can also be used as input to another analysis. Then, they consist in data dimension reduction and operate by replacing the original variables, sometimes highly correlated, by a smaller number of linearly independent variables. Factor Analysis (F.A.) is one possible method for quantitative data. This article aims at presenting in a synthetic way the F.A. model, rarely described in French books, but frequent in the Anglo-Saxon literature, and often available in softwares. The presentation of the estimation techniques for the F.A. model enables to establish the existing connection between Principal Component Analysis (P.C.A.) and F.A. The usefulness of rotation techniques, which can facilitate the interpretation of the results, will also be shown. An application on crime data of American cities will be carried out and will allow to describe the results provided by three of the most used statistical softwares : SAS, SPAD and SPSS. Then it will help to clarify the vocabulary, sometimes confused for the user.

Keywords : Factor Analysis, Principal Component Analysis, Singular Value Decomposition, Rotation.

Résumé En analyse des données, les méthodes factorielles sont fondamentales. Ces techniques peuvent être utilisées comme but en soi, il s'agit alors de faire ressortir des facteurs sous-jacents communs à un groupe de variables. Elles peuvent également constituer une étape préalable à d'autres études. Elles consistent alors à réduire la dimension des données en remplaçant les variables d'origine, qui peuvent être corrélées, par un plus petit nombre de variables linéairement indépendantes. Lorsque les données sont quantitatives, l'Analyse en Facteurs (A.F.) est une des méthodes possibles. L'objectif de cet article est de dresser une présentation synthétique du modèle d'A.F., peu développé dans les manuels francophones, mais fréquent dans la littérature anglo-saxonne, et souvent présent dans les logiciels statistiques. La

présentation des techniques d'estimation du modèle d'A.F. permet d'établir le lien existant entre l'Analyse en Composantes Principales (A.C.P.) et l'A.F. Il s'agit également de montrer l'utilité des techniques de rotation, qui peuvent faciliter l'interprétation des résultats. Un exemple d'application sur des données de criminalité de villes américaines permet enfin de décrire les résultats fournis par trois des logiciels statistiques les plus utilisés : SAS, SPAD et SPSS, et ainsi de clarifier le vocabulaire, parfois confus pour l'utilisateur.

Mots-clés : Analyse en Facteurs, Analyse en Composantes Principales, Décomposition en Valeurs Singulières, Rotation.

1 Introduction

L'A.F. trouve son origine en psychométrie lorsqu'en 1904, Spearman développe une théorie psychologique selon laquelle l'esprit humain s'explique par un facteur commun à tous les individus et par plusieurs facteurs spécifiques à chacun. Ce modèle est généralisé pour plusieurs facteurs communs par Garnett en 1919. De nombreuses applications sont alors réalisées pour déterminer un nombre relativement faible de tests qui permettraient de décrire l'esprit humain de façon aussi complète que possible.

Ainsi, l'A.F. vise à écrire chaque variable aléatoire du problème en fonction de facteurs sous-jacents communs à toutes les variables, et d'un facteur spécifique ou unique à la variable aléatoire considérée. Il repose sur différentes hypothèses dont principalement la non corrélation des facteurs communs. Différentes méthodes d'estimation existent, les plus courantes sont l'estimation via les composantes principales, la méthode du facteur principal et le maximum de vraisemblance. L'estimation du modèle d'A.F. via l'A.C.P. ne garantit pas que les hypothèses du modèle soient vérifiées. Cependant cette technique est la plus utilisée car elle fournit souvent une approximation convenable.

Cet article met également en lumière un point essentiel de l'A.F. : le choix du nombre q de facteurs communs. Différents critères empiriques et théoriques existent pour le choisir. Nous insisterons sur le fait que ces règles sont une aide partielle qui ne doit pas se substituer à une interprétation rigoureuse des résultats. Notons que l'enjeu de ce choix est majeur car la qualité des résultats en dépend.

Suite à l'estimation du modèle d'A.F., la lecture des résultats peut s'avérer délicate. Les facteurs obtenus peuvent être difficiles à interpréter, sembler ne pas avoir d'intérêt pour l'étude, ou ne pas exliquer le phénomène considéré, etc. Des résultats sont pourtant parfois présents, mais leur lecture n'est pas directe et intuitive. L'utilisateur peut alors passer à côté de résultats importants. Une rotation orthogonale des facteurs peut aider dans cette phase. La justification de la possibilité d'effectuer une rotation provient de la non-unicité de la solution du modèle d'A.F. Nous verrons de plus que la rotation est possible en A.C.P. à condition d'effectuer convenablement la transformation. Bien que les techniques de rotation peuvent faciliter de façon significative la lecture des résultats, elles sont peu présentées dans les ouvrages francophones, contrairement à leurs voisins anglo-saxons. L'utilité de la rotation des facteurs sera mise en exergue sur une application concernant la criminalité de seize villes américaines (données issues de *U.S. Statistical Abstract*, 1970).

Enfin, l'estimation du modèle peut se faire à l'aide de logiciels statistiques, comme SAS, SPAD et SPSS. Le vocabulaire employé diffère d'un logiciel à l'autre et peut rapidement devenir source de confusion. L'exemple d'application précise ce vocabulaire et pourra ainsi aider les utilisateurs dans la lecture des sorties numériques des logiciels.

Le présent article s'articule autour de cinq parties. Le modèle d'A.F. est présenté à la section 2. L'estimation des paramètres du modèle est ensuite décrite à la section 3. Les techniques de rotation des facteurs, facilitant la détection de groupes de variables corrélées, sont présentées à la section 4. Enfin, à la section 5, une application de ce modèle d'A.F. est réalisée sur des données de criminalité dans différentes villes des Etats-Unis et permet de comparer les résultats fournis par les trois logiciels statistiques SAS, SPAD et SPSS.

2 Le modèle d'A.F.

Soit $\mathbf{x} = (x^1, x^2, \ldots, x^p)'$ un vecteur aléatoire de $\mathbb{R}^p$ d'espérance $\mu \in \mathbb{R}^p$. On note $\tilde{\mathbf{x}} = \mathbf{x} - \mu$ la version centrée de $\mathbf{x}$.

Le modèle d'A.F. s'écrit :

$$\underset{(p\times 1)}{\tilde{\mathbf{x}}} = \underset{(p\times q)(q\times 1)}{A_q\mathbf{f}} + \underset{(p\times 1)}{\mathbf{e}} \tag{1}$$

où :

- A_q est une matrice $(p\times q)$ de coefficients a_j^α, $j = 1, \ldots, p$, $\alpha = 1, \ldots, q$ ("loadings" en anglais). Elle est appelée matrice de saturation ("factor loadings matrix" ou "factor pattern matrix").
- $\mathbf{f} = (f^1, \ldots, f^q)'$ est un vecteur aléatoire de $\mathbb{R}^q$, composé des q facteurs communs ("common factors") aux p variables aléatoires $\tilde{x}^1, \tilde{x}^2, \ldots, \tilde{x}^p$.
- $\mathbf{e} = (e^1, \ldots, e^p)'$ est un vecteur aléatoire centré de $\mathbb{R}^p$, composé des p facteurs spécifiques (ou uniques) ("unique factors") à chaque variable $\tilde{x}^j, j = 1, \ldots, p$.

Il découle de (1) et de $\mathbb{E}(\mathbf{e}) = 0$ la propriété suivante :

$$\mathbb{E}(\mathbf{f}) = 0. \tag{2}$$

Pour tout $j = 1, \ldots, p$, on a :

$$\tilde{x}^j = \sum_{\alpha=1}^{q} a_j^\alpha f^\alpha + e^j. \tag{3}$$

Chaque variable $\tilde{x}^j$ s'écrit comme la somme d'une combinaison linéaire de facteurs $f^1, \ldots, f^q$ communs à toutes les variables $\tilde{x}^1, \ldots, \tilde{x}^p$ et d'un facteur e^j spécifique à la variable considérée $\tilde{x}^j$.

On insiste sur le fait que les facteurs communs $f^1, \ldots, f^q$ sont aléatoires. Ainsi, le modèle d'A.F. est souvent désigné comme un modèle à effets aléatoires ou modèle structurel (Baccini et Besse, 2005).

Le modèle (1) repose sur plusieurs hypothèses.

(H_1) : $\mathbb{E}(\mathbf{f}\mathbf{f}') = I_q$, où I_q est la matrice identité $(q \times q)$.

(H_2) : $\mathbb{E}(\mathbf{e}\mathbf{e}') = \Xi$, où $\Xi = diag(\xi^j, j = 1, \ldots, p)$.
(H_3) : $\mathbb{E}(\mathbf{e}\mathbf{f}') = 0$.

L'hypothèse (H_1) signifie que les facteurs communs $f^\alpha, \alpha = 1, \ldots, q$, sont non corrélés et de variance 1. Cette hypothèse de non corrélation des facteurs s'explique par le fait que l'on souhaite exprimer les variables aléatoires $\tilde{x}^j$ en fonction du plus petit nombre de facteurs possible, et donc éviter des redondances.

L'hypothèse (H_2) signifie que les facteurs uniques $e^j, j = 1, \ldots, p$, ne sont pas corrélés. Ils expriment pour chaque variable la part non expliquée par les facteurs communs. Ils ont chacun une variance spécifique ξ^j.

L'hypothèse (H_3) traduit le fait que chaque variable $e^j, j = 1, \ldots, p$, traduit la part spécifique à la variable $\tilde{x}^j$ qui n'a pu être exprimée par les facteurs communs $f^\alpha, \alpha = 1, \ldots, q$, donc les variables e^j et f^α, ne sont pas corrélées.

On note Σ la matrice de variance covariance de $\mathbf{x}$. On déduit du modèle (1) que :

$$\begin{aligned} \mathbb{E}(\tilde{\mathbf{x}}\tilde{\mathbf{x}}') &= A_q \mathbb{E}(\mathbf{f}\mathbf{f}') A_q' + \mathbb{E}(\mathbf{e}\mathbf{e}') \text{ et donc} \\ \Sigma &= A_q A_q' + \Xi. \end{aligned} \tag{4}$$

L'équation (4) est appelée modèle de structure de covariance.
D'après (1) ou (3), on peut écrire pour tout $j = 1, \ldots, p$:

$$\begin{aligned} \mathbb{V}(x^j) &= (a_j^1)^2 + (a_j^2)^2 + \ldots + (a_j^q)^2 + \xi^j \\ &= \sum_{\alpha=1}^{q} (a_j^\alpha)^2 + \xi^j \\ &= h_j^2 + \xi^j. \end{aligned} \tag{5}$$

De même, pour $j \neq k$:

$$cov(x^j, x^k) = \sum_{\alpha=1}^{q} a_j^\alpha a_k^\alpha + 0. \tag{6}$$

On voit ainsi que les covariances des variables aléatoires $x^j, j = 1, \ldots, p$, sont complètement reconstituées par la matrice de saturation A_q tandis que les variances se décomposent en une part due aux facteurs communs, appelée communalité ou variance commune, et une part due aux facteurs spécifiques, appelée variance spécifique ou résiduelle.

On remarque également que A_q est la matrice des covariances entre les variables aléatoires $x^j, j = 1, \ldots, p$, et les facteurs communs $f^\alpha, \alpha = 1, \ldots, q$. En effet :

$$\begin{aligned} cov(\mathbf{x}, \mathbf{f}) = \mathbb{E}(\mathbf{x}\mathbf{f}') &= \mathbb{E}((A_q\mathbf{f} + \mathbf{e} + \mu)\mathbf{f}') \\ &= A_q \mathbb{E}(\mathbf{f}\mathbf{f}') + \mathbb{E}(\mathbf{e}\mathbf{f}') + \mu \mathbb{E}(\mathbf{f}') \\ &= A_q. \end{aligned} \tag{7}$$

On travaille maintenant sur les variables standardisées, c'est-à-dire que $\tilde{\mathbf{x}}$ correspond au vecteur $\mathbf{x}$ centré réduit : $\tilde{\mathbf{x}} = \Sigma^{-1/2}(\mathbf{x} - \mu)$.

Dans ce cas, la matrice A_q devient la matrice des corrélations linéaires entre les variables x^j et les facteurs f^α, et l'équation (4) s'écrit :

$$\Upsilon = A_q A_q' + \Xi \tag{8}$$

où Υ est la matrice de corrélation linéaire de $\mathbf{x}$.

De façon analogue à (5), on a :

$$1 = h_j^2 + \xi^j. \tag{9}$$

Dans la suite de cet article, nous considérons que le vecteur $\tilde{\mathbf{x}}$ correspond au vecteur $\mathbf{x}$ centré réduit.

3 Estimation du modèle

On veut estimer A_q et $\mathbf{f}$ dans le modèle (1). Rigoureusement, on ne devrait pas parler d'estimation pour $\mathbf{f}$ car il s'agit d'un vecteur aléatoire, on va donc obtenir une réalisation et non une estimation de $\mathbf{f}$. Nous nous conformerons cependant à cet abus de langage, fréquent dans la littérature.

Pour cela, on dispose d'un échantillon $\{\mathbf{x_1}, \ldots, \mathbf{x_n}\}$ de n réalisations indépendantes et identiquement distribuées du vecteur aléatoire $\mathbf{x}$ de $\mathbb{R}^p$.

D'après (1), on peut écrire pour tout $i = 1, \ldots, n$:

$$\tilde{\mathbf{x}}_\mathbf{i} = A_q \mathbf{f_i} + \mathbf{e_i}. \tag{10}$$

On note :
- $\tilde{X}$ la matrice $(n \times p)$ des données centrées réduites.
- F_q la matrice $(n \times q)$ correspondant aux n réalisations des q facteurs communs. Elle est appelée matrice des scores des facteurs communs ("factor scores matrix").
- E_q la matrice $(n \times p)$ des erreurs spécifiques.

Le modèle d'A.F. sur échantillon s'écrit alors :

$$\underset{(n \times p)}{\tilde{X}} = \underset{(n \times q)(q \times p)}{F_q A_q'} + \underset{(n \times p)}{E_q.} \tag{11}$$

Nous présentons ici trois méthodes d'estimation de A_q et F_q. Pour toute méthode d'estimation, il faut ensuite choisir le nombre q (avec $q \leq p$) de facteurs communs que l'on retient. Quelques critères pour le choisir sont discutés dans la section 3.4.

3.1 Estimation du modèle via les composantes principales

Cette technique utilise l'A.C.P. comme méthode d'estimation du modèle d'A.F. Nous rappelons donc dans un premier temps le principe de l'A.C.P., puis nous expliquons comment cette méthode est utilisée pour estimer le modèle d'A.F.

3.1.1 Présentation de l'A.C.P.

L'A.C.P est proposée pour la première fois par Pearson en 1901, elle est ensuite intégrée à la statistique mathématique par Hotelling en 1933. L'A.C.P. peut être considérée selon différents points de vue. La présentation la plus fréquente dans la littérature francophone est géométrique. L'A.C.P est alors vue comme une technique visant à représenter de façon optimale des données, selon certains critères géométriques et algébriques. Le lecteur pourra se reporter à l'ouvrage de Lebart et al. (1997). L'A.C.P peut être considérée sur un plan probabiliste, elle est alors un cas particulier du modèle d'A.F. où les variances spécifiques sont nulles ou égales, voir par exemple Tipping et Bishop (1999). Dans cet article, nous adopterons une présentation de l'A.C.P. qui nous permettra de faire le lien avec l'A.F.

L'A.C.P présente deux variantes, elle peut être réalisée à partir des données centrées ou des données centrées réduites. Dans le premier cas, on parle d'A.C.P. non normée ou A.C.P sur matrice des covariances. Dans le second cas, on parle d'A.C.P normée ou A.C.P. sur matrice des corrélations. Nous présentons dans cet article l'A.C.P. normée, variante la plus utilisée.

L'objectif de l'analyse du nuage des individus $\{x_1, \ldots, x_n\}$ en A.C.P. est de déterminer q nouvelles variables $\psi^1, \ldots, \psi^q$ avec $q \leq p$, permettant de résumer "au mieux" les p variables $\tilde{x}^1, \ldots, \tilde{x}^p$. Ces q nouvelles variables sont appelées les composantes principales des individus. Elles sont définies comme des combinaisons linéaires des p variables $\tilde{x}^1, \ldots, \tilde{x}^p$. On a donc, pour $\alpha = 1, \ldots, q$:

$$\psi^\alpha = v_1^\alpha \tilde{x}^1 + \ldots + v_p^\alpha \tilde{x}^p = \tilde{X} v^\alpha. \tag{12}$$

On suppose que l'espace $\mathbb{R}^n$ est muni de la métrique M, matrice de dimension $(n \times n)$, avec $M = diag(1/\sqrt{m}, \ldots, 1/\sqrt{m})$, où $m = n$ ou $m = n - 1$ selon l'estimateur de la variance choisi.

On veut que ces composantes soient de variance maximale et deux à deux orthogonales. Par construction, les colonnes de $\tilde{X}$ sont centrées et donc les composantes principales le sont aussi. On a donc :

$$\mathbb{V}(\psi^\alpha) = (\psi^\alpha)' M \psi^\alpha = (v^\alpha)' R v^\alpha \tag{13}$$

où $R = \tilde{X}' M \tilde{X}$ est la matrice des corrélations empiriques entre les variables initiales $x^1, \ldots, x^p$.

En ajoutant la contrainte $(v^\alpha)'(v^\alpha) = 1$, on démontre, voir par exemple Lebart et al. (1997), que pour $\alpha = 1, \ldots, q$, v^α est le vecteur propre associé à la $\alpha^{\text{ème}}$ plus grande valeur propre de la matrice des corrélations R.

On construit ainsi la matrice Ψ_q dont les colonnes sont les composantes principales des individus $\psi^\alpha, \alpha = 1, \ldots, q$:

$$\Psi_q = \tilde{X} V_q \tag{14}$$

où V_q est la matrice $(p \times q)$ dont les colonnes sont les vecteurs propres $v^\alpha, \alpha = 1, \ldots, q$, associés aux q plus grandes valeurs propres de la matrice R.

3.1.2 Estimation du modèle d'A.F.

L'A.C.P. peut être utilisée comme méthode d'estimation du modèle d'A.F. Le lien entre l'A.C.P. et l'A.F. s'obtient facilement à partir de la décomposition en valeurs singulières (D.V.S.) de la matrice $Z = M\tilde{X}$.

On note r (avec $r \leq p < n$) le rang de la matrice Z et on écrit sa D.V.S. :

$$Z = U\Lambda V' \tag{15}$$

où :
- $\Lambda = diag(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_r})$ des valeurs singulières des matrices ZZ' et $Z'Z$ rangées par ordre décroissant ($\sqrt{\lambda_1} \geq \sqrt{\lambda_2} \geq \ldots \geq \sqrt{\lambda_r} > 0$).
- U est la matrice orthonormée $(n \times r)$ dont les colonnes sont les vecteurs propres de ZZ' associés aux r valeurs propres.
- V est la matrice orthonormée $(p \times r)$ dont les colonnes sont les vecteurs propres de $Z'Z$ associés aux r valeurs propres.

On a donc :

$$\tilde{X} = M^{-1}Z = M^{-1}U\Lambda V'. \tag{16}$$

On note U_q, Λ_q et V_q les matrices contenant respectivement les q premières colonnes de U, Λ et V.

• **Avec q = r.**
Pour se ramener au modèle d'A.F. (11), on pose :

$$\hat{F}_q = M^{-1}U_q \tag{17}$$
$$\hat{A}_q = V_q\Lambda_q. \tag{18}$$

On note que dans ce cas $\hat{E}_q = 0$.

Comme $U_q'U_q = I_r$, on montre que :

$$\hat{A}_q = Z'U_q = \tilde{X}'M^{-1}U_q. \tag{19}$$

Cette écriture est utilisée pour démontrer que les éléments de la matrice $\hat{A}_q$, notés $\hat{a}_j^\alpha$, sont les corrélations empiriques entre les variables x^j et les facteurs f^α (détails en annexe 7.1).

Comme $V_qV_q' = I_p$, on montre également que :

$$\hat{F}_q = \tilde{X}\underbrace{V_q\Lambda_q^{-1}}_{V_q^*}. \tag{20}$$

Cette écriture de $\hat{F}_q$ en fonction de $\tilde{X}$ fait ainsi apparaitre la matrice V_q^* des coefficients des scores des facteurs communs, calculée par certains logiciels statistiques.

• **Avec q < r.**
En ne retenant que les vecteurs propres associés aux q plus grandes valeurs propres, on a l'approximation de $\tilde{X}$ suivante :

$$\tilde{X} = \hat{F}_q\hat{A}_q + \hat{E}_q \tag{21}$$

où :

– $\hat{F}_q$ contient les q premières colonnes de $\hat{F}_q$ définie dans (17).
– $\hat{A}_q$ contient les q premières colonnes de $\hat{A}_q$ définie dans (18).
– $\hat{E}_q$ est la matrice des erreurs associée à cette approximation.

Avec cette méthode d'estimation, on montre facilement (voir les détails en annexe 7.1) que les facteurs communs estimés possèdent les bonnes propriétés mais que les hypothèses du modèle ne sont pas nécessairement toutes vérifiées.

3.1.3 Lien avec l'A.C.P.

Les facteurs communs estimés par cette méthode correspondent aux composantes principales des individus (trouvées en A.C.P.) standardisées. En effet, d'après les égalités (14) et (20), on voit que :

$$\hat{F}_q = \Psi_q \Lambda_q^{-1} \tag{22}$$

De plus, la matrice de saturations $\hat{A}_q$ est égale à la matrice des composantes principales des variables. En effet, si on présente l'A.C.P. d'un point de vue géométrique, on réalise généralement non seulement l'analyse des points-individus, comme présenté ici, mais également celle des points-variables. On montre que ces composantes correspondent aux corrélations entre les variables x^j et les facteurs f^α, et donc aux saturations (voir l'ouvrage de Lebart et al., 1997).

3.1.4 Quelques éléments de vocabulaire

On peut introduire, à partir de ces premiers résultats, le vocabulaire utilisé en A.C.P. et en A.F. (tableau 1).

TAB. 1 – Quelques éléments de vocabulaire en A.F. et A.C.P.

	Matrices	**Français**	**Anglais**
ACP	Ψ_q	Composantes principales	Principal component scores
	V_q	Coefficient des composantes principales	Principal component scoring coefficients
AF	F_q	Facteurs communs	Factor scores ou Standardized principal component scores
	V_q^*	Coefficients des facteurs communs	Factor scoring coefficients ou Standardized principal component scoring coefficients

3.2 Méthode du facteur principal

A partir de l'échantillon $\{\mathbf{x_1}, \ldots, \mathbf{x_n}\}$, on calcule la matrice des corrélations empiriques définie par :

$$R = \tilde{X}' M \tilde{X} \tag{23}$$

L'équation (8) du modèle de structure de covariance sur échantillon s'écrit alors :

$$R = \hat{A}_q \hat{A}_q' + \hat{\Xi} \tag{24}$$

Il faut donc déterminer $\hat{A}_q$ et $\hat{\Xi}$.

Pour cela, la méthode du facteur principal estime Ξ (en fait $\Upsilon - \Xi$) et factorise $R - \hat{\Xi}$ pour obtenir $\hat{A}_q \hat{A}'_q$ en utilisant les valeurs propres et vecteurs propres de $R - \hat{\Xi}$.

• **Estimation de Ξ.**
D'après l'équation (9), un estimateur de $\Upsilon - \Xi$ est donné par :

$$R - \hat{\Xi} = \begin{pmatrix} \hat{h}_1^2 & r_{12} & \dots & r_{1p} \\ r_{21} & \hat{h}_2^2 & \dots & r_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ r_{p1} & r_{p2} & \dots & \hat{h}_p^2 \end{pmatrix}$$

où $\hat{h}_j^2$ est l'estimation de la $j^{\text{ème}}$ communalité définie par : $\hat{h}_j^2 = 1 - \hat{\xi}^j$.

D'après (5), la communalité h_j^2 traduit la part commune entre x^j et les $p-1$ variables restantes. Ainsi, une estimation courante pour la communalité h_j^2 est R_j^2, le coefficient de corrélation multiple entre x^j et les $p-1$ variables restantes.

Ainsi,

$$\hat{h}_j^2 = R_j^2 = 1 - \frac{1}{r^{jj}} \tag{25}$$

où r^{jj} est le $j^{\text{ème}}$ élément diagonal de R^{-1}.

Pour effectuer ces estimations, R doit être régulière. Si R est singulière, on utilise pour $\hat{h}_j^2$ la valeur absolue (ou le carré) de la plus grande corrélation de x^j avec les $p-1$ autres variables. Un autre moyen de remédier à cette singularité est de remarquer que, puisque R est singulière, cela signifie qu'il existe des combinaisons linéaires des variables. On peut donc supprimer les redondances et rendre ainsi R inversible.

Notons t le rang de la matrice $R - \hat{\Xi}$.

• **Avec $\mathbf{q = t}$.**

- Estimation de $\mathbf{A_q}$.
On écrit la décomposition spectrale de la matrice $R - \hat{\Xi}$:

$$R - \hat{\Xi} = CDC' = (CD^{1/2})(CD^{1/2})' = \hat{A}_q \hat{A}'_q \tag{26}$$

où :
- $D = diag(\theta_1, \theta_2, \dots, \theta_t)$ des valeurs propres non nulles de $R - \hat{\Xi}$.
- C est la matrice orthonormale dont les colonnes sont les vecteurs propres normés de $R - \hat{\Xi}$ associés aux t valeurs propres non nulles.

- Estimation de $\mathbf{F_q}$.
Après avoir estimé A_q, il faut "estimer" F_q. Une méthode possible est de choisir l'estimateur linéaire $\hat{\mathbf{f}} = L\tilde{\mathbf{x}}$ qui minimise l'erreur quadratique moyenne :

$$E[\|\hat{\mathbf{f}} - \mathbf{f}\|^2] = E[\|L\tilde{\mathbf{x}} - \mathbf{f}\|^2] = E[\|LA_q\mathbf{f} + L\mathbf{e} - \mathbf{f}\|^2]. \tag{27}$$

Seber (1984) montre que (27) est égale à :

$$\text{trace}(L'L\Sigma) - 2\text{trace}(LA_q) + q. \tag{28}$$

En différenciant par rapport à A_q, on obtient :

$$\begin{aligned} 2L\Sigma - 2A'_q &= 0 \\ L &= A'_q\Sigma^{-1} \end{aligned} \tag{29}$$

et donc :

$$\hat{\mathbf{f}} = A'_q\Sigma^{-1}\tilde{\mathbf{x}}. \tag{30}$$

En remplaçant A_q par son estimateur (26), et Σ par la matrice de variance covariance empirique S, on obtient :

$$\begin{aligned} \hat{F}_q &= \tilde{X}S^{-1}\hat{A}_q \\ &= \tilde{X}R^{-1}\hat{A}_q \text{ car on travaille avec la matrice } \tilde{X} \text{ centrée réduite.} \end{aligned} \tag{31}$$

Remarque. En développant le calcul de $\Sigma^{-1} = (A_qA'_q + \Xi)^{-1}$ dans (30), on montre (voir l'ouvrage de Seber, 1984) que $\hat{\mathbf{f}}$ est l'estimateur "ridge" de $\mathbf{f}$:

$$\hat{\mathbf{f}} = (I_q + A'_q\Xi^{-1}A_q)^{-1}A'_q\Xi^{-1}\tilde{\mathbf{x}}. \tag{32}$$

• Avec q < t.
Afin de retenir seulement q facteurs communs dans le modèle d'A.F., on ne conserve que les q, premières colonnes des matrices $\hat{A}_q$ et $\hat{F}_q$ définies respectivement dans (26) et (31).

• Itération de la méthode.
Cette méthode du facteur principal peut facilement être itérée afin d'améliorer l'estimation de A_q. Après avoir estimé A_q à partir de (26), nous pouvons obtenir une nouvelle estimation de la communalité en utilisant (5) :

$$\hat{h}_j^2 = \sum_{\alpha=1}^{q} (\hat{a}_j^\alpha)^2.$$

Ces valeurs sont alors insérées dans la diagonale de $R - \hat{\Xi}$, ce qui nous permet d'obtenir une nouvelle estimation de A_q à partir de la décomposition spectrale de la nouvelle matrice $R - \hat{\Xi}$, comme dans l'équation (26). Ce processus est alors itéré jusqu'à ce que les estimations de la communalité se stabilisent.

Cependant, un défaut majeur de la méthode itérée est qu'elle ne converge pas toujours. De plus, lors de ces itérations, $\hat{h}_j^2$ peut devenir supérieur à 1, ce qui implique $\hat{\xi}^j < 0$. Or, ceci est impossible car on ne peut pas avoir une variance spécifique estimée négative. Ce problème est connu sous le nom de *Heywood case* (Heywood, 1931).

3.3 Maximum de vraisemblance

On suppose que l'échantillon $\{\mathbf{x_1}, \ldots, \mathbf{x_n}\}$ est issu d'une loi multi-normale $N_p(\mu, \Sigma)$. Alors $(n-1)S$ suit la loi de Wishart $W_p(n-1, \Sigma)$ et la log-vraisemblance de l'échantillon est donnée par :

$$\log L(A_q, \Xi) = c - \frac{1}{2}(n-1)(\ln|\Sigma| + \text{trace}(\Sigma^{-1}S)) \tag{33}$$

où c est une constante et $|M|$ désigne le déterminant de M.

Les paramètres A_q et Ξ vont alors pouvoir être estimés en maximisant $\log L(A_q, \Xi)$, sous la contrainte $\Sigma = A_q A_q' + \Xi$ avec Ξ matrice diagonale. La condition suivante : $A_q' \Xi^{-1} A_q$ diagonale, est souvent rajoutée afin d'avoir une solution unique.

L'équation du maximum de vraisemblance n'a pas de solution analytique, la résolution se fait donc numériquement par itérations successives. Cependant, cette méthode ne converge pas toujours. De plus, des cas de variances spécifiques estimées négatives peuvent là encore se produire (*Heywood case*).

Notons qu'avec cette méthode, les facteurs communs obtenus $\hat{f}^\alpha$ ne sont pas forcément ordonnés par variance expliquée décroissante comme avec la méthode des composantes principales et la méthode du facteur principal.

Après avoir estimé A_q et Ξ, il faut estimer la matrice F_q des scores des facteurs communs. Pour cela, on utilise souvent la méthode des moindres carrés généralisés :

$$\hat{F}_q = (\hat{A}_q' \hat{\Xi}^{-1} \hat{A}_q)^{-1} \hat{A}_q' \hat{\Xi}^{-1} \tilde{X}. \tag{34}$$

On ne retient ensuite que les q, avec $q \leq p$, premières colonnes des matrices $\hat{F}_q$ et $\hat{A}_q$.
Pour de plus amples détails sur cette méthode d'estimation, le lecteur pourra se référer à l'ouvrage de Seber (1984).

3.4 Choix du nombre de facteurs

Comme dans toute méthode factorielle, une étape importante de l'A.F. est le choix du nombre q de facteurs communs. La qualité des estimations du modèle dépend de q. En effet, si q est trop grand, certains facteurs spécifiques vont être mélangés aux facteurs communs. A l'inverse si q est trop petit, des facteurs communs importants risquent d'être oubliés. Différents critères théoriques et empiriques peuvent être utilisés pour choisir q.

Voici deux critères théoriques reposant sur la normalité de l'échantillon :

• Critère 1.
Ce critère consiste à déterminer si les $(p-k)$ dernières valeurs propres de la matrice de covariance Σ sont significativement différentes entre elles. On fait pour cela l'hypothèse que les n observations sont les réalisations d'un vecteur aléatoire gaussien dont les $(p-k)$ dernières valeurs propres $\lambda_{k+1}, \ldots, \lambda_p$ de la matrice Σ sont égales. Sous cette hypothèse,

la moyenne arithmétique m_a des $(p-k)$ dernières valeurs propres doit être peu différente de la moyenne géométrique m_g. On définit :

$$T_1 = \left(n - \frac{2p+11}{6}\right)(p-k)\log\left(\frac{m_a}{m_g}\right). \tag{35}$$

Sous H_0, on peut montrer que T_1 suit une loi du Khi-deux à $v_1 = \frac{(p-k+2)(p-k-1)}{2}$ degrés de liberté.

On rejette donc H_0 au seuil de signification α si l'inégalité suivante est vérifiée :

$$T_1 > \chi^2_{v_1,1-\alpha} \tag{36}$$

où $\chi^2_{v_1,1-\alpha}$ est le fractile d'ordre $(1-\alpha)$ de la loi du Khi-deux à v_1 degrés de liberté.

Certains auteurs (voir par exemple Bouveyron, 2006) soulignent le fait que ce critère surestime très souvent le nombre de facteurs à retenir.

• Critère 2.
Ce critère est utilisé lorsque la méthode d'estimation du modèle d'A.F. est le maximum de vraisemblance.
On désire tester l'hypothèse que q_0 est le bon nombre de facteurs communs. Les hypothèses sont donc : $H_0 : \Sigma = A_{q_0}(A_{q_0})' + \Xi$, où A_{q_0} est de dimension $(p \times q_0)$ contre $H_1 : \Sigma = A_q A_q' + \Xi$, où A_q est de dimension $(p \times q)$ avec $q > q_0$.
La statistique de test est :

$$T_2 = \left(n - \frac{2p+4q_0+11}{6}\right)\log\left(\frac{|\hat{A}_{q_0}\hat{A}'_{q_0} + \hat{\Xi}|}{|S|}\right) \tag{37}$$

où $\hat{A}_{q_0}$ et $\hat{\Xi}$ sont les estimateurs du maximum de vraisemblance de A_q et Ξ, obtenus avec q_0 facteurs communs.

Sous H_0, on peut montrer que T_2 suit une loi du Khi-deux à $v_2 = \frac{1}{2}[(p-q_0)^2 - p - q_0]$ degrés de liberté.

On rejette donc H_0 au seuil de signification α si la condition (36) est vérifiée (où l'on aura préalablement substitué v_2 à v_1). Si H_0 est rejetée, cela signifie que le nombre q_0 de facteurs communs est trop petit.

En pratique, on commence souvent avec $q_0 = 1$, et on ajoute des facteurs jusqu'à trouver la valeur q pour laquelle H_0 est vérifiée. Ainsi, le risque associé à la procédure pour trouver le bon nombre de facteurs q_0 est supérieur à α, du fait de la multiplicité des tests.

Il faut noter que cette technique est souvent utilisée pour fixer la borne supérieure du nombre de facteurs. En effet, on peut trouver dans la littérature, voir par exemple Rencher (2002), que lorsque le nombre d'observations est grand, cette méthode a tendance à surestimer le nombre de facteurs communs.

Différentes règles empiriques peuvent également être utilisées pour choisir q. Dans la définition des critères ci-dessous, $\lambda_i, i = 1, \ldots, p$, fait référence aux valeurs propres de R

(ou S) ou bien $R-\hat{\Xi}$ (ou $S-\hat{\Xi}$), selon que la méthode d'estimation utilisée est la méthode des composantes principales ou du facteur principal. Voici quelques exemples de critères empiriques.

• Pourcentage de variance expliquée.
On choisit q tel que le pourcentage de variance expliquée par les q facteurs soit supérieur ou égal à un seuil fixé par l'utilisateur. L'appréciation de ce pourcentage doit tenir compte du nombre de variables p et du nombre d'observations n. En effet, un pourcentage de 10% peut être considéré comme élevé pour $p = 100$ variables et au contraire, faible pour $p = 10$. Notre expérience nous a montré que ce critère a souvent tendance à surestimer le nombre de facteurs q.

• Le test du coude.
On utilise le test du coude de Cattell, ou *scree-test*, basé sur l'analyse des différences consécutives entre les valeurs propres. On calcule les différences premières :

$$\epsilon_i = \lambda_i - \lambda_{i+1}$$

puis les différences secondes :

$$\delta_i = \epsilon_i - \epsilon_{i+1}.$$

On retient alors les valeurs propres $\lambda_1, \lambda_2, \ldots, \lambda_k, \lambda_{k+1}$ tels que $\delta_1, \ldots, \delta_k$ soient tous positifs.
Visuellement, ce critère revient à détecter un coude dans le graphe de l'éboulis des valeurs propres. En pratique, la détection graphique de ce coude peut se révéler difficile.

• La règle de Kaiser.
Sachant que la variance expliquée par un facteur $\hat{f}^\alpha$ est λ_α, il s'agit de retenir les facteurs dont la variance expliquée dépasse la moyenne de la variance totale expliquée : $\overline{\lambda} = \frac{\sum_{i=1}^{p} \lambda_i}{p}$. Pour la matrice de corrélation R, on a : $\overline{\lambda} = 1$. Cette valeur 1 peut aussi être vue comme la variance de chaque variable x^j et on retient donc un facteur s'il explique au moins autant de variance qu'une variable toute seule.

On insiste sur le fait que ces critères ne doivent pas se substituer à une analyse approfondie de l'interprétation des facteurs. Il est indispensable d'examiner l'information apportée par un facteur, et ainsi juger de sa pertinence et de son intérêt quant aux objectifs de l'étude. L'utilisateur retiendra, par exemple, un facteur dont la part de variance expliquée est faible, mais dont l'intérêt est significatif pour la problématique traitée. Au contraire, il pourra rejeter un facteur qui possède une part de variance expliquée élevée, mais qui n'aide pas à comprendre le phénomène étudié. Ainsi, on peut utiliser ces critères comme valeur initiale du nombre q_0 de facteurs, puis au vu de l'interprétation des résultats et des objectifs de l'étude, on peut augmenter ou diminuer ce nombre q_0 afin de trouver une interprétation des résultats satisfaisante.

3.5 Choix de la méthode d'estimation

On trouve dans la littérature (voir par exemple Rencher, 2002) que les solutions obtenues avec la méthode des composantes principales et la méthode du facteur principal (itérée ou non) sont très proches lorsque l'une des deux conditions suivantes est vérifiée :

– Les corrélations entre les variables $x^j, j = 1, \ldots, p$, sont élevées.
– Le nombre de variables p est grand.

Cependant, il est important de noter que la méthode d'estimation la plus utilisée est celle des composantes principales. C'est une technique qui fournit une approximation convenable de la solution et qui est facile à mettre en oeuvre. Ainsi, c'est la méthode utilisée par défaut lorsqu'on estime un modèle d'A.F. sous les logiciels SAS et SPSS. Enfin, contrairement aux deux autres techniques, elle ne présente pas le problème de *Heywood case*.

4 La rotation des facteurs

Dans cette section, nous allons présenter les motivations de la rotation orthogonale des facteurs estimés, puis nous montrerons que cette rotation est justifiée car elle conserve les propriétés des facteurs. Quelques critères permettant la mise en place d'une rotation optimale seront ensuite discutés. Enfin, nous montrerons brièvement que la rotation est possible en A.C.P., comme en A.F., à condition d'effectuer convenablement la transformation.

4.1 Motivations

Après avoir estimé le modèle d'A.F., on peut vouloir interpréter les facteurs communs obtenus en détectant des groupes de variables corrélées aux différents facteurs. La matrice de saturation A_q, représentant les corrélations entre les variables x^j et les facteurs communs f^α, il s'agit de faire apparaître des variables fortement corrélées à un même facteur.

Il est donc souhaitable que pour chaque colonne de A_q les valeurs soient proches soit de 0, soit de 1 et qu'il n'y ait ainsi pas de valeur intermédiaire. Cela permet alors d'associer clairement des variables à un facteur. Il faut également s'assurer que sur chaque ligne de A_q, il n'y aura qu'une seule valeur proche de 1. En effet, si la corrélation entre une variable x^j et un facteur f^α est proche de 1, les corrélations de cette variable avec les facteurs restants doivent être proches de 0, car les facteurs sont orthogonaux entre eux. Cela se traduit par la condition d'orthonormalité de la matrice de transformation T. Ainsi, chaque variable x^j ne pourra être parfaitement associée qu'à un seul facteur f^α.

Cependant, l'estimation $\hat{A}_q$ trouvée ne présente pas toujours une telle structure. Afin de se rapprocher de cette situation, il est possible de réaliser une rotation des facteurs. La justification de la possibilité de faire une rotation provient de la non-unicité de la solution du modèle d'A.F. Ainsi, il s'agit de choisir la solution optimale du point de vue de l'interprétation des résultats.

Remarque. Il existe des rotations qui ne conservent pas la propriété d'orthogonalité des facteurs communs. Ce type de rotation dites obliques ne sera pas abordé dans cet article. Pour plus de détails, le lecteur pourra se reporter à l'ouvrage de Rencher (2002).

4.2 Justifications de la rotation

La solution du modèle d'A.F. n'est pas unique. En effet, soit T une matrice orthonormale de dimension $(q \times q)$. On peut écrire :

$$\begin{aligned}\tilde{X} &= \hat{F}_q \hat{A}'_q + \hat{E}_q \\ &= \hat{F}_q T T' \hat{A}'_q + \hat{E}_q \\ &= \hat{G}_q \hat{B}'_q + \hat{E}_q \text{ avec } \hat{G}_q = \hat{F}_q T \text{ et } \hat{B}_q = \hat{A}_q T. \end{aligned} \tag{38}$$

Ainsi $\hat{G}_q$ est l'estimation de la matrice des facteurs après la rotation et $\hat{B}_q$ est l'estimation de la matrice de saturation après la rotation.
La transformation orthogonale entraîne une rotation "rigide" des q axes définis par les facteurs communs, c'est-à-dire que les q nouveaux axes restent perpendiculaires après la rotation.

Propriété :
Les facteurs communs g^α et les saturations $\hat{b}_j^\alpha$ vérifient toujours les propriétés et hypothèses du modèle d'A.F. après la rotation.

La démonstration est disponible en annexe 7.2.

On montre en particulier que les saturations après rotation, $\hat{b}_j^\alpha$, sont toujours les corrélations des variables d'origine x^j aux facteurs après rotation, g^α.

Cependant, même si suite à la rotation, les communalités estimées sont inchangées et que l'on a :

$$\hat{h}_j^2 = \sum_{\alpha=1}^{q} (\hat{b}_j^\alpha)^2 = \sum_{\alpha=1}^{q} (\hat{a}_j^\alpha)^2, \tag{39}$$

la variance expliquée par chaque facteur f^α change lors de la rotation. En effet :

$$\begin{aligned}\sum_{j=1}^{p} (\hat{b}_j^\alpha)^2 &= \sum_{j=1}^{p} (\hat{a}_j t^\alpha)^2 \text{ où } t^\alpha \text{ est la } \alpha^{\text{ème}} \text{ colonne de } T \\ &= \sum_{j=1}^{p} (\sum_{k=1}^{q} \hat{a}_j^k t_k^\alpha)^2 \\ &\neq \sum_{j=1}^{p} (\hat{a}_j^\alpha)^2. \end{aligned} \tag{40}$$

Après la rotation, les facteurs ne sont donc plus forcément rangés par ordre de variance expliquée décroissante.

4.3 Comment faire la rotation ?

Afin d'effectuer la rotation, il faut déterminer la matrice T qui fournit la "meilleure" interprétation des résultats, c'est-à-dire telle que les éléments $\hat{b}_j^\alpha$ de la matrice $\hat{B}_q = \hat{A}_q T$ soient proches de 0 ou de 1, on parle de "structure simple" de $\hat{B}_q$. Différents critères existent, le plus utilisé est Varimax.

4.3.1 Varimax

Comme toutes les variables n'ont pas la même commmunalité, le critère Varimax est souvent appliqué sur les valeurs standardisées de $\hat{B}_q$, obtenues en divisant chaque ligne de la matrice $\hat{B}_q$ par $\hat{h}_j$. On travaille avec le carré de ces éléments, $(\hat{b}_j^\alpha/\hat{h}_j)^2$, afin de se ramener à des valeurs comprises entre 0 et 1. On note $\hat{B}_q^*$ cette matrice. On veut que ses éléments soient aussi proches que possible de 0 ou de 1. Pour cela, il faut maximiser la variance empirique de chaque colonne de $\hat{B}_q^*$ afin de donner plus de poids aux valeurs extrêmes 0 et 1.

Ceci équivaut à maximiser la somme sur l'ensemble des q facteurs des variances empiriques de chaque colonne de $\hat{B}_q^*$, c'est-à-dire la quantité :

$$\sum_{\alpha=1}^{q}\left\{\frac{\sum_{j=1}^{p}((\hat{b}_j^\alpha)^2/\hat{h}_j^2)^2}{p}-\left(\frac{\sum_{j=1}^{p}((\hat{b}_j^\alpha)^2/\hat{h}_j^2)}{p}\right)^2\right\} \tag{41}$$

avec $\hat{b}_j^\alpha = \hat{a}_j t^\alpha$, $\hat{a}_j$ est la $j^{\text{ème}}$ ligne de la matrice $\hat{A}_q$, et t^α est la $\alpha^{\text{ème}}$ colonne de la matrice T.

La maximisation de la quantité (41) se fait donc de façon itérative, par rapport à $t^\alpha, \alpha = 1,\ldots,q$, sous la contrainte $t^\alpha(t^\alpha)' = 1$ et $t^k(t^l)' = 0$ pour $k \neq l$.

4.3.2 Quartimax

Le critère Quartimax maximise la somme des variances des éléments $(\hat{b}_j^\alpha)^2$ sur toute la matrice $\hat{B}_q$, c'est-à-dire la quantité :

$$\frac{\sum_{\alpha=1}^{q}\sum_{j=1}^{p}(\hat{b}_j^\alpha)^4}{pq}-\left(\frac{\sum_{\alpha=1}^{q}\sum_{j=1}^{p}(\hat{b}_j^\alpha)^2}{pq}\right)^2. \tag{42}$$

On trouve dans la littérature (Jobson, 1992) que ceci équivaut à maximiser la quantité $\sum_{\alpha=1}^{q}\sum_{j=1}^{p}(\hat{b}_j^\alpha)^4$. Il est de plus mentionné que cette méthode a tendance à produire un facteur commun général car elle maximise la variance des $(\hat{b}_j^\alpha)^2$ sur la totalité de la matrice de saturation $\hat{B}_q^*$ et non sur chaque colonne, comme le critère Varimax.

4.3.3 Orthomax

Le critère Orthomax est une généralisation des critères de rotation orthogonale. Il s'agit de maximiser la quantité :

$$\sum_{\alpha=1}^{q}\left\{\sum_{j=1}^{p}(\hat{b}_j^\alpha)^4-\frac{\delta}{p}\left(\sum_{j=1}^{p}(\hat{b}_j^\alpha)^2\right)^2\right\} \tag{43}$$

Pour $\delta = 0$ et $\delta = 1$, on retrouve respectivement le critère Quartimax et la version non standardisée de Varimax. Pour $\delta = 0.5$, le critère s'appelle Biquartimax et pour $\delta = \frac{q}{2}$, ce critère est connu sous le nom de Equamax.

4.4 Remarques sur la rotation et l'A.C.P.

La rotation en A.C.P est possible, comme en A.F., mais il faut être prudent et appliquer la transformation T aux bonnes matrices.

Si on applique la rotation directement sur les composantes principales Ψ_q, les nouvelles composantes après rotation $\Psi_q T$ ne sont pas nécessairement non corrélées. En effet :

$$(\Psi_q T)' M \Psi_q T = T' \Psi_q' M \Psi_q T = T' \Lambda^2 T. \tag{44}$$

On en déduit que $(\Psi_q T)' M \Psi_q T$ n'est pas forcément la matrice identité.

Afin d'effectuer convenablement une rotation en A.C.P. il faut donc introduire la matrice T au bon endroit dans l'écriture de $\tilde{X}$.

Il ne faut pas écrire :

$$\tilde{X} = \overbrace{M^{-1} U \Lambda}^{\Psi_q} \mathbf{TT'} V' \tag{45}$$

mais :

$$\tilde{X} = \overbrace{M^{-1} U}^{\hat{F}_q} \mathbf{TT'} \Lambda V'. \tag{46}$$

Ainsi les composantes obtenues après rotation correspondent en fait aux facteurs communs obtenus après rotation $\hat{G}_q = M^{-1} U T$. Nous avons déjà vérifié en annexe 7.2 que ces facteurs ne sont pas corrélés.

On comprend ainsi la raison pour laquelle les logiciels qui calculent les composantes principales ne proposent pas de rotation des composantes, celles-ci deviendraient en effet corrélées. A l'inverse, les logiciels qui construisent les facteurs communs proposent une rotation.

5 Un exemple d'application sur des données de criminalité

5.1 Problématique

• Données.
Dans cette exemple, nous étudions la criminalité de seize villes américaines, données étudiées par de nombreux auteurs dont Rencher (2002) (extraites de *U.S. Statistical Abstract*, 1970). Pour cela, sept types d'effractions sont relevés et un taux pour 100 000 habitants est calculé (tableau 2). L'objectif est de résumer la criminalité de ces villes à l'aide de facteurs communs. Nous allons étudier les résultats fournis par les logiciels

SAS, SPAD et SPSS. Nous choisissons d'estimer le modèle d'A.F. via les composantes principales.

Il faut noter que le logiciel SPAD utilise la définition de l'estimateur biaisé de l'écart-type ($m = n$), les logiciels SAS et SPSS utilise au contraire la définition de l'estimateur non biaisé de l'écart-type ($m = n - 1$). Il est cependant possible de préciser au logiciel SAS d'utiliser l'estimation biaisée de l'écart-type avec l'option "*vardef=n*".

TAB. 2 – Criminalité

Ville	Meurtre	Viol	Vol	Agression	Cambriolage	Vol avec effraction	Vol de voiture
Atlanta	16.5	24.8	106	147	1112	905	494
Boston	4.2	13.3	122	90	982	669	954
Chicago	11.6	24.7	340	242	808	609	645
Dallas	18.1	34.2	184	293	1668	901	602
Denver	6.9	41.5	173	191	1534	1368	780
Detroit	13	35.7	477	220	1566	1183	788
Hartford	2.5	8.8	68	103	1017	724	468
Honolulu	3.6	12.7	42	28	1457	1102	637
Houston	16.8	26.6	289	186	1509	787	697
Kansas City	10.8	43.2	255	226	1494	955	765
Los Angeles	9.7	51.8	286	355	1902	1386	862
New Orleans	10.3	39.7	266	283	1056	1036	776
New York	9.4	19.4	522	267	1674	1392	848
Portland	5	23	157	144	1530	1281	488
Tucson	5.1	22.9	85	148	1206	756	483
Washington	12.5	27.6	524	217	1496	1003	793

On note $X = (x_i^j), i = 1, \ldots, n, j = 1, \ldots, p$, la matrice des données présentées dans le tableau 2 avec $n = 16$ observations et $p = 7$ variables.

• **Choix de q.**
Les différents auteurs, qui ont étudiés ces données, ont conservé $q = 3$ facteurs communs (voir par exemple Rencher, 2002). Nous vérifions, par une étude approfondie, que la valeur 3 permet une bonne interprétation des résultats. Pour cela, nous comparons les valeurs de q proposées par les critères empiriques discutés dans la section 3.4 et examinons attentivement les valeurs propres de la matrice des corrélations R (figure 1). Dans cette étape, nous privilégions l'interprétation des facteurs et leur intérêt pour la problématique étudiée. Nous vérifions ainsi que la valeur 3 permet des interprétations intéressantes pour l'étude menée. Dans la suite de cet article, la valeur q est donc choisi égale à 3, pour l'A.C.P. comme pour l'A.F.

FIG. 1 – Valeurs propres de la matrice R

Eigenvalues of the Correlation Matrix

	Valeur propre	Différence	Proportion	Cumulée
1	3.46216658	2.13250001	0.4946	0.4946
2	1.32966657	0.37144677	0.1900	0.6845
3	0.95821980	0.34730456	0.1369	0.8214
4	0.61091525	0.25013602	0.0873	0.9087
5	0.36077923	0.19073695	0.0515	0.9602
6	0.17004228	0.06183198	0.0243	0.9845
7	0.10821030		0.0155	1.0000

5.2 Logiciel SAS

La première partie présente la procédure PRINCOMP qui réalise une A.C.P. sur les données. Dans un second temps, nous décrivons les résultats de la procédure FACTOR avec pour méthode d'estimation l'A.C.P. et ainsi nous comparons les résultats des deux procédures.

5.2.1 Procédure PRINCOMP

Le code SAS de la procédure PRINCOMP est présenté dans la figure 2. L'option "$n = 3$" permet de réduire l'affichage des résultats à 3 composantes principales.

FIG. 2 – Code SAS de la procédure PRINCOMP

```
proc PRINCOMP data=doncrim.crimrates n=3 outstat=loadACP out=comp;
var Meurtre Viol Vol Agression Cambriolage Vol_avec_effraction Vol_de_voiture;
run;
```

La procédure propose comme résultats la matrice des corrélations empiriques $R = \tilde{X}'M\tilde{X}$, ses valeurs propres λ_α et les vecteurs propres associés, c'est-à-dire la matrice V_q (figure 3). Cette matrice peut également être obtenue dans une table avec l'option "*outstat=loadACP*". Il s'agit en fait de la matrice des coefficients des composantes principales.

FIG. 3 – Sorties numériques de la procédure PRINCOMP

Correlation Matrix

	Meurtre	Viol	Vol	Agression	Cambriolage	Vol_avec_effraction	Vol_de_voiture
Meurtre	1.0000	0.4349	0.4374	0.5558	0.2318	-.0681	0.0630
Viol	0.4349	1.0000	0.3150	0.7722	0.4973	0.4574	0.3725
Vol	0.4374	0.3150	1.0000	0.6065	0.3402	0.3201	0.5433
Agression	0.5558	0.7722	0.6065	1.0000	0.4200	0.3425	0.3790
Cambriolage	0.2318	0.4973	0.3402	0.4200	1.0000	0.7592	0.2751
Vol_avec_effraction	-.0681	0.4574	0.3201	0.3425	0.7592	1.0000	0.3088
Vol_de_voiture	0.0630	0.3725	0.5433	0.3790	0.2751	0.3088	1.0000

Eigenvalues of the Correlation Matrix

	Valeur propre	Différence	Proportion	Cumulée
1	3.46216658	2.13250001	0.4946	0.4946
2	1.32966657	0.37144677	0.1900	0.6845
3	0.95821980		0.1369	0.8214

Eigenvectors

	Prin1	Prin2	Prin3
Meurtre	0.284077	0.601420	-.290666
Viol	0.434117	0.057951	-.280953
Vol	0.387340	0.193196	0.454238
Agression	0.458700	0.265038	-.090328
Cambriolage	0.388251	-.403606	-.293904
Vol_avec_effraction	0.346301	-.596641	-.125837
Vol_de_voiture	0.315820	-.092125	0.721022

FIG. 4 – Matrice V_q des coefficients des composantes principales

SCORE	Prin1	0.2840770255	0.4341165199	0.3873395476	0.4587002129	0.3882506994	0.346300896	0.3158201451
SCORE	Prin2	0.6014200757	0.0579510469	0.1931960648	0.2650381889	-0.403606485	-0.596640949	-0.092124862
SCORE	Prin3	-0.290666091	-0.280952874	0.4542381853	-0.090328373	-0.293904228	-0.125837414	0.7210217837

L'option "*out=comp*" permet d'obtenir dans une table appelée "*comp*", la matrice Ψ_q des composantes principales (figure 5).

FIG. 5 – Matrice Ψ_q des composantes principales

Prin1	Prin2	Prin3
-1.211983352	1.1808854874	-1.303102471
-2.133959083	-0.107709329	2.2119885684
-0.872935491	2.1520641374	0.7123636412
1.130005234	1.184421741	-1.571231827
0.9796205812	-1.48142512	-0.255612807
1.8426143897	0.077312182	0.4616195197
-3.344991804	-0.241562028	-0.108412417
-2.202267721	-1.901723853	-0.066729639
0.3019595472	1.1949108748	-0.242587142
1.0299148373	0.2190691351	-0.160284845
3.3434931322	-1.012810751	-0.486160635
0.7842181022	0.7215288423	0.356669141
1.9519465049	-0.895021258	1.213102309
-0.815483767	-1.590279827	-1.04171094
-2.086689049	-0.025286754	-0.716503761
1.3045379405	0.5256265198	0.996593304

5.2.2 Procédure FACTOR

La procédure SAS permettant d'estimer un modèle d'A.F. est FACTOR. Cette procédure nous propose différentes méthodes d'estimation dont la méthode des composantes principales que nous spécifions avec l'option "*method=prin*" (figure 6). L'option "*nfactors=3*" permet de fixer q.

FIG. 6 – Code SAS de la procédure FACTOR

```
proc FACTOR data=doncrim.crimrates method=prin nfactors=3 outstat=loadAF out=fact;
var Meurtre Viol Vol Agression Cambriolage Vol_avec_effraction Vol_de_voiture;
run;
```

La procédure FACTOR fournit comme la procédure PRINCOMP les valeurs propres λ_α de la matrice des corrélations empiriques $R = \tilde{X}'M\tilde{X}$.

En précisant l'option "*out=fact*", le logiciel calcule la réalisation des facteurs communs $f^\alpha, \alpha = 1, \ldots, q$, sur les n observations, c'est-à-dire la matrice $\hat{F}_q$ (figure 7).

FIG. 7 – Matrice $\hat{F}_q$ des facteurs communs

Factor1	Factor2	Factor3
-0.651362384	1.024085956	-1.331208242
-1.146864496	-0.093407542	2.2596975134
-0.469146166	1.8663102248	0.72772815
0.6073044668	1.0271526611	-1.605120706
0.5264824771	-1.28471954	-0.261125953
0.9902856339	0.0670465686	0.4715758913
-1.797715978	-0.209487103	-0.110750694
-1.183575955	-1.649210454	-0.068168887
0.162283657	1.0362490339	-0.247819347
0.5535123752	0.1899808466	-0.163741925
1.7969105388	-0.878328404	-0.496646318
0.4214663279	0.6257232916	0.3643619061
1.0490445493	-0.776179155	1.2392669249
-0.438269594	-1.379120376	-1.064178926
-1.121459921	-0.021929146	-0.731957565
0.701104468	0.4558331377	1.0180881778

Le logiciel calcule la matrice V_q^* des coefficients des facteurs communs. Dans les sorties du logiciel, elle est appelée "*standardized scoring coefficents*" (figure 8). Avec l'option "*outstat=loadAF*", on peut également obtenir cette matrice dans la table "*loadAF*".

FIG. 8 – Matrice V_q^* des coefficients des scores des facteurs communs

Standardized Scoring Coefficients

	Factor1	Factor2	Factor3
Meurtre	0.15267	0.52156	-0.29694
Viol	0.23331	0.05026	-0.28701
Vol	0.20817	0.16754	0.46404
Agression	0.24652	0.22985	-0.09228
Cambriolage	0.20866	-0.35002	-0.30024
Vol_avec_effraction	0.18611	-0.51742	-0.12855
Vol_de_voiture	0.16973	-0.07989	0.73657

SCORE	Factor1	0.1526729623	0.2333094518	0.2081698654	0.2465215911	0.2086595503	0.1861142539	0.1697328287
SCORE	Factor2	0.5215627254	0.0502562305	0.1675432367	0.2298460689	-0.350015084	-0.517418178	-0.079892402
SCORE	Factor3	-0.296935279	-0.287012564	0.4640353537	-0.092276607	-0.300243257	-0.12855152	0.7365730343

La matrice de saturation estimée $\hat{A}_q$ est présentée sous le nom de "factor pattern" (figure 9). Ses coefficients sont les corrélations des variables d'origine x^j aux facteurs communs f^α.

FIG. 9 – Matrice $\hat{A}_q$ de saturation

Factor Pattern

	Factor1	Factor2	Factor3
Meurtre	0.52858	0.69350	-0.28453
Viol	0.80776	0.06682	-0.27502
Vol	0.72072	0.22278	0.44465
Agression	0.85350	0.30562	-0.08842
Cambriolage	0.72241	-0.46540	-0.28770
Vol_avec_effraction	0.64436	-0.68799	-0.12318
Vol_de_voiture	0.58764	-0.10623	0.70580

En rajoutant l'option "*rotate=varimax*" dans le code présenté dans la figure 6, nous demandons au logiciel d'effectuer une rotation orthogonale avec le critère Varimax. La matrice T de transformation orthogonale, estimée selon ce critère, est présentée dans la figure 10.

FIG. 10 – Matrice T de transformation orthogonale

Orthogonal Transformation Matrix

	1	2	3
1	0.61059	0.60511	0.51090
2	0.69393	-0.71967	0.02304
3	-0.38163	-0.34046	0.85933

En rajoutant l'option "*out=factrotation*", on obtient la matrice $\hat{G}_q = \hat{F}_q T$ des facteurs communs après la rotation (figure 11).

La matrice de saturation après rotation, $\hat{B}_q = \hat{A}_q T$, est calculée et nommée "rotated factor pattern" (figure 12). Les valeurs de cette matrice sont les corrélations des variables d'origine x^j aux facteurs communs g^α après la rotation, donnés en figure 11.

On voit très clairement sur cet exemple que la rotation des "loadings" facilite la lecture des résultats. Les valeurs de $\hat{A}_q$ (figure 9) sont très dispersées et rendent difficile la détection de groupes de variables corrélées à un même facteur. Au contraire, la matrice $\hat{B}_q$ (figure 12) possède beaucoup plus de valeurs soit proches de 1, soit proches de 0. On peut ainsi associer clairement chaque variable à un facteur.

FIG. 11 – Matrice $\hat{G}_q$ des facteurs après la rotation

Factor1	Factor2	Factor3
0.8209573792	-0.677921774	-1.453131596
-1.627438853	-1.396103804	1.3537329652
0.7309213041	-1.874778112	0.4286739055
1.6961441603	0.1747618885	-1.045383689
-0.470394333	1.3320583558	0.0149855921
0.4712167452	0.390426783	0.9127247156
-1.200768377	-0.899349739	-1.018458959
-1.84110518	0.4939011466	-0.701275648
0.9127510435	-0.563183957	-0.10616891
0.5322902033	0.2539612555	0.146461291
0.6772028014	1.8885272871	0.4710274804
0.5525029634	-0.31933239	0.5428539416
-0.371019377	0.7714553414	1.5830136881
-0.818502468	1.0896243119	-1.17017204
-0.420633311	-0.413620834	-1.202456046
0.3558752995	-0.25042576	1.2435733079

FIG. 12 – Matrice $\hat{B}_q$ de saturation après la rotation

Rotated Factor Pattern

	Factor1	Factor2	Factor3
Meurtre	0.91257	-0.08237	0.04153
Viol	0.64453	0.53433	0.17789
Vol	0.42497	0.12440	0.75545
Agression	0.76696	0.32662	0.36712
Cambriolage	0.22793	0.87003	0.11113
Vol_avec_effraction	-0.03698	0.92698	0.20750
Vol_de_voiture	0.01574	0.19174	0.90429

On peut alors décrire les trois facteurs communs de la criminalité dans ces seize villes américaines. Le premier facteur fait référence aux crimes violents contre une personne : meurtre, viol et agression. Le second facteur se rapporte aux crimes en rapport avec la maison : cambriolage et vol avec effraction. Enfin le troisième facteur fait référence aux vols commis à l'extérieur : vol et vol de voiture.

5.3 Logiciel SPSS

Avec le logiciel SPSS, dans le menu "*Analyse* → *Factorisation* → *Analyse factorielle*", on peut choisir différentes méthodes d'extraction des facteurs en cliquant sur le bouton "*Extraction*" : Maximum de vraisemblance, Composantes principales, Factorisation en axes principaux, etc (figure 13).

En cliquant sur le bouton "*Facteur*"(figure 13), on peut demander au logiciel d'afficher l'estimation de la matrice $\hat{F}_q$ des scores des facteurs communs, et l'estimation de la matrice des coefficients des scores des facteurs communs V_q^*.

L'estimation de ces deux matrices $\hat{F}_q$ et V_q^* est présentée à la figure 14. Nous retrouvons les résultats de la procédure FACTOR de SAS, présentés aux figures 7 et 8.

Le logiciel propose comme résultats la matrice de saturation estimée, $\hat{A}_q$, appelée "matrice des composantes" (figure 15). On voit là l'erreur commise par le logiciel car le terme "matrice des composantes" est réservé à Ψ_q. Ce problème de vocabulaire provient peut-être d'une mauvaise traduction française de ce logiciel anglais.

Le logiciel nous propose également différentes rotations. En choisissant le critère Varimax,

FIG. 13 – Estimation du modèle d'A.F. avec SPSS

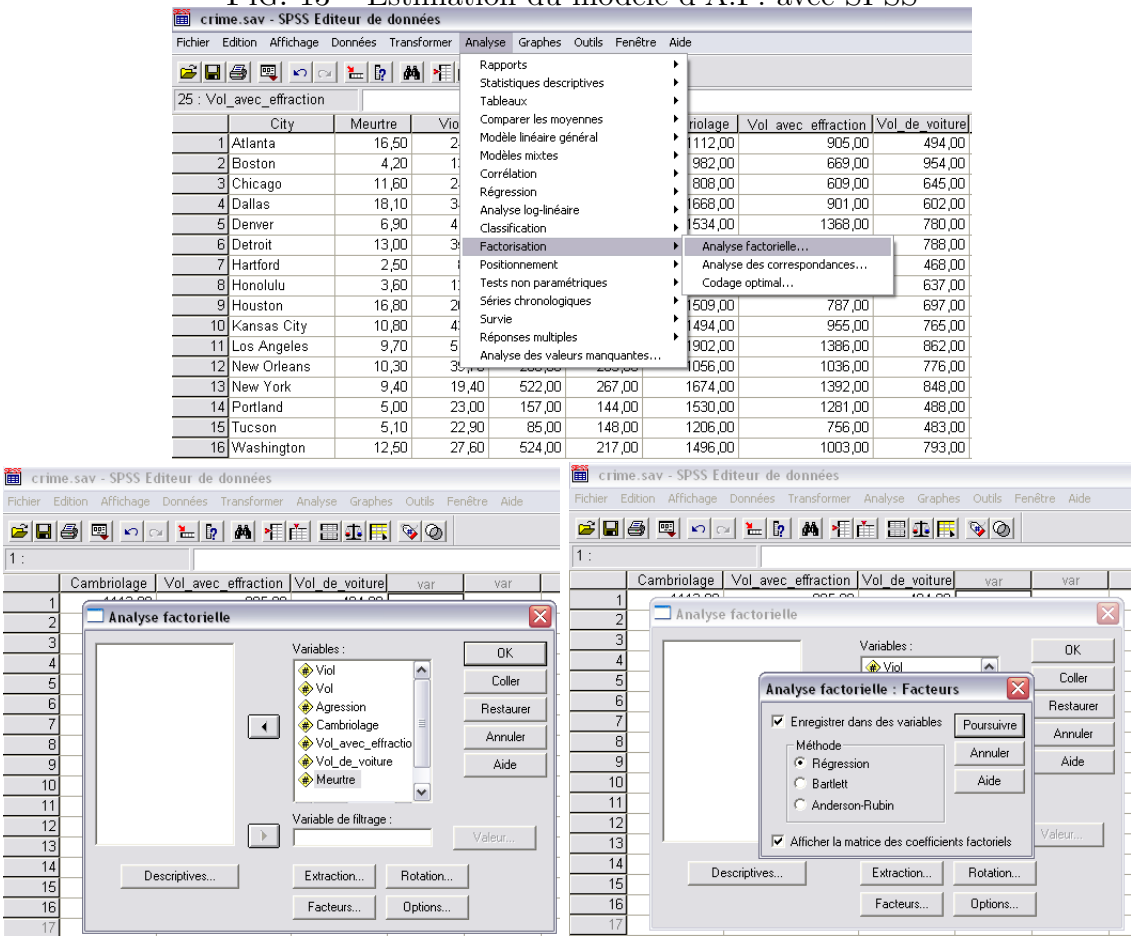

FIG. 14 – Matrice $\hat{F}_q$ et V_q^*

FAC1_1	FAC2_1	FAC3_1
-,65136	1,02409	-1,33121
-1,14686	-,09341	2,25970
-,46915	1,86631	,72773
,60730	1,02715	-1,60512
,52648	-1,28472	-,26113
,99029	,06705	,47158
-1,79772	-,20949	-,11075
-1,18358	-1,64921	-,06817
,16228	1,03625	-,24782
,55351	,18998	-,16374
1,79691	-,87833	-,49665
,42147	,62572	,36436
1,04904	-,77618	1,23927
-,43827	-1,37912	-1,06418
-1,12146	-,02193	-,73196
,70110	,45583	1,01809

Matrice des coefficients des coordonnées des composantes

	Composante		
	1	2	3
Viol	,233	,050	-,287
Vol	,208	,168	,464
Agression	,247	,230	-,092
Cambriolage	,209	-,350	-,300
Vol_avec_effraction	,186	-,517	-,129
Vol_de_voiture	,170	-,080	,737
Meurtre	,153	,522	-,297

Méthode d'extraction : Analyse en composantes principales.

Scores composante.

nous retrouvons les matrices $\hat{G}_q$ et $\hat{B}_q$ (figure 16) de la procédure FACTOR du logiciel SAS (figures 11 et 12).

FIG. 15 – Matrice $\hat{A}_q$ de saturation

Matrice des composantes[a]

	Composante		
	1	2	3
Meurtre	,529	,694	-,285
Viol	,808	,067	-,275
Vol	,721	,223	,445
Agression	,853	,306	-,088
Cambriolage	,722	-,465	-,288
Vol_avec_effraction	,644	-,688	-,123
Vol_de_voiture	,588	-,106	,706

Méthode d'extraction : Analyse en composantes principales.

a. 3 composantes extraites.

FIG. 16 – Matrice $\hat{G}_q$ et $\hat{B}_q$ après la rotation

FAC1_2	FAC2_2	FAC3_2
,82096	-,67793	-1,45313
-1,62745	-1,39610	1,35373
,73091	-1,87478	,42868
1,69615	,17476	-1,04538
-,47039	1,33206	,01498
,47122	,39043	,91273
-1,20077	-,89935	-1,01846
-1,84110	,49391	-,70128
,91275	-,56319	-,10617
,53229	,25396	,14646
,67721	1,88853	,47103
,55250	-,31933	,54286
-,37102	,77146	1,58301
-,81850	1,08963	-1,17017
-,42063	-,41362	-1,20246
,35587	-,25043	1,24357

Matrice des composantes après rotation[a]

	Composante		
	1	2	3
Viol	,645	,534	,178
Vol	,425	,124	,755
Agression	,767	,327	,367
Cambriolage	,228	,870	,111
Vol_avec_effraction	-,037	,927	,207
Vol_de_voiture	,016	,192	,904
Meurtre	,913	-,082	,042

Méthode d'extraction : Analyse en composantes principales.
Méthode de rotation : Varimax avec normalisation de Kaiser.

a. La rotation a convergé en 5 itérations.

5.4 Logiciel SPAD

Parmi les méthodes d'analyse factorielle du logiciel SPAD, l'A.C.P. est proposée mais l'A.F. n'est pas disponible. Afin d'effectuer une A.C.P., on insère la méthode des composantes principales appelée "COPRI" (figure 17).

Le logiciel réalise alors l'analyse des points-variables (voir l'ouvrage de Lebart et al., 1997) et propose comme résultats les coordonnées des variables sur les composantes calculées. Cette matrice est égale à la matrice des corrélations variable-facteur, il s'agit de l'estimation de la matrice de saturation, notée $\hat{A}_q$ (figure 18). Sur cette figure, on retrouve également la matrice des "anciens axes unitaires" qui correspond en fait à la matrice V_q des vecteurs propres de R.

Pour obtenir la matrice des composantes principales Ψ_q (figure 19) , il faut demander à SPAD, lors du paramétrage de la filière, d'afficher les résultats pour les individus.

FIG. 17 – Filière SPAD

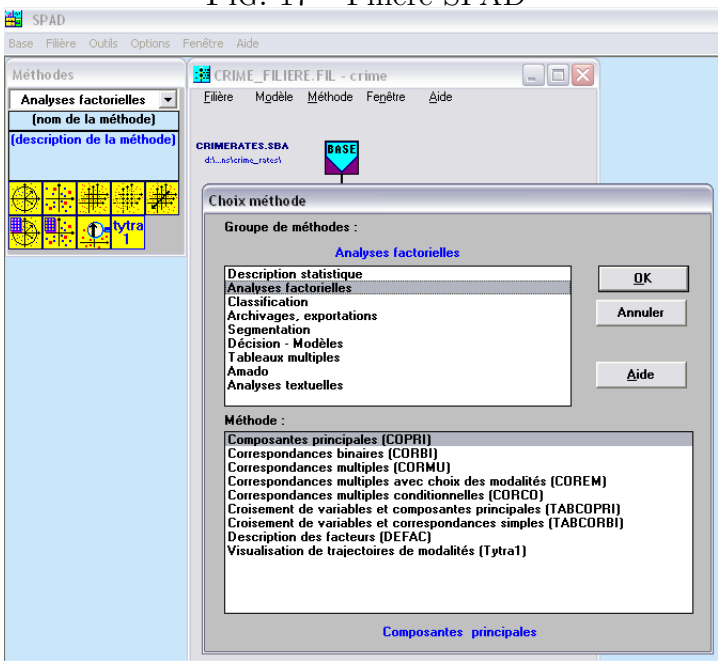

FIG. 18 – Matrice $\hat{A}_q$ de saturation

COORDONNEES DES VARIABLES SUR LES AXES 1 A 3
VARIABLES ACTIVES

VARIABLES	COORDONNEES					CORRELATIONS VARIABLE-FACTEUR					ANCIENS AXES UNITAIRES				
IDEN - LIBELLE COURT	1	2	3	0	0	1	2	3	0	0	1	2	3	0	0
C2 - Meurtre	0.53	-0.69	-0.28	0.00	0.00	0.53	-0.69	-0.28	0.00	0.00	0.28	-0.60	-0.29	0.00	0.00
C3 - Viol	0.81	-0.07	-0.28	0.00	0.00	0.81	-0.07	-0.28	0.00	0.00	0.43	-0.06	-0.28	0.00	0.00
C4 - Vol	0.72	-0.22	0.44	0.00	0.00	0.72	-0.22	0.44	0.00	0.00	0.39	-0.19	0.45	0.00	0.00
C5 - Agression	0.85	-0.31	-0.09	0.00	0.00	0.85	-0.31	-0.09	0.00	0.00	0.46	-0.27	-0.09	0.00	0.00
C6 - Cambriolage	0.72	0.47	-0.29	0.00	0.00	0.72	0.47	-0.29	0.00	0.00	0.39	0.40	-0.29	0.00	0.00
C7 - Vol_avec_effraction	0.64	0.69	-0.12	0.00	0.00	0.64	0.69	-0.12	0.00	0.00	0.35	0.60	-0.13	0.00	0.00
C8 - Vol_de_voiture	0.59	0.11	0.71	0.00	0.00	0.59	0.11	0.71	0.00	0.00	0.32	0.09	0.72	0.00	0.00

FIG. 19 – Matrice Ψ_q des composantes principales

INDIVIDUS			COORDONNEES				
IDENTIFICATEUR	P.REL	DIST0	1	2	3	0	0
01	6.25	6.03	-1.25	-1.22	-1.35	0.00	0.00
02	6.25	11.95	-2.20	0.11	2.28	0.00	0.00
03	6.25	7.18	-0.90	-2.22	0.74	0.00	0.00
04	6.25	6.43	1.17	-1.22	-1.62	0.00	0.00
05	6.25	4.60	1.01	1.53	-0.26	0.00	0.00
06	6.25	4.66	1.90	-0.08	0.48	0.00	0.00
07	6.25	12.69	-3.45	0.25	-0.11	0.00	0.00
08	6.25	9.77	-2.27	1.96	-0.07	0.00	0.00
09	6.25	3.26	0.31	-1.23	-0.25	0.00	0.00
10	6.25	2.29	1.06	-0.23	-0.17	0.00	0.00
11	6.25	14.73	3.45	1.05	-0.50	0.00	0.00
12	6.25	3.65	0.81	-0.75	0.37	0.00	0.00
13	6.25	9.14	2.02	0.92	1.25	0.00	0.00
14	6.25	5.31	-0.84	1.64	-1.08	0.00	0.00
15	6.25	5.88	-2.16	0.03	-0.74	0.00	0.00
16	6.25	4.42	1.35	-0.54	1.03	0.00	0.00

5.5 Tableau récapitulatif des différents résultats

Le tableau ci-dessous récapitule un certain nombre de résultats et permet de faire le lien entre les facteurs ou composantes, obtenus avec les trois logiciels. Pour faciliter la

lecture, les matrices ne sont pas indicées par q.

TAB. 3 – Tableau comparatif des différents résultats

	SAS **Proc** PRINCOMP	**SAS** **Proc** FACTOR **(estimation par A.C.P.)**	**SPAD**	**SPSS**
Méthode factorielle	A.C.P.	A.F.	A.C.P.	A.F.
Estimateur de la variance	non biaisé	non biaisé	biaisé	non biaisé
Facteurs/composantes	$\Psi_{sas} = \tilde{X}V$	$F_{sas} = \tilde{X}V\Lambda^{-1}$	$\Psi_{spad} = \tilde{X}V$	$F_{spss} = \tilde{X}V\Lambda^{-1}$
Norme au carré des facteurs	$(n-1)\lambda_\alpha$	$n-1$	$n\lambda_\alpha$	$n-1$
Variance des facteurs	λ_α	1	λ_α	1
Matrice de saturation ? Nom donné par le logiciel	non fournie	$\hat{A} = V\Lambda$ "factor pattern"	$\hat{A} = V\Lambda$ "corrélations variables facteurs"	$\hat{A} = V\Lambda$ "matrice des composantes"
Rotation possible ?	non	oui	non	oui
Lien	ψ^α_{sas}	$f^\alpha_{sas} = \frac{\psi^\alpha_{sas}}{\sqrt{\lambda_\alpha}}$	$\psi^\alpha_{spad} = \sqrt{\frac{n}{n-1}}\psi^\alpha_{sas}$	$f^\alpha_{spss} = f^\alpha_{sas}$

Pour mettre en place un modèle d'A.F., on peut donc utiliser la procédure FACTOR de SAS ou le logiciel SPSS.

Si on utilise la procédure PRINCOMP de SAS ou le logiciel SPAD, la méthode réalisée est une A.C.P. et il faut standardiser les composantes principales pour obtenir l'estimation des facteurs communs. Ceci peut se faire facilement sous le logiciel SAS, en spécifiant l'option "*standard*" dans le code de la procédure PRINCOMP (figure 2). On obtient alors la matrice des composantes principales standardisées présentée dans la figure (20), qui correspond bien à la matrice $\hat{F}_q$ de la procédure FACTOR (figure 7).

FIG. 20 – Matrice Ψ_q des composantes principales standardisées

Prin1	Prin2	Prin3
-0.651362384	1.024085956	-1.331208242
-1.146864496	-0.093407542	2.2596975134
-0.469146166	1.8663102248	0.72772815
0.6073044668	1.0271526611	-1.605120706
0.5264824771	-1.28471954	-0.261125953
0.9902856339	0.0670465686	0.4715758913
-1.797715978	-0.209487103	-0.110750694
-1.183575955	-1.649210454	-0.068168887
0.162283657	1.0362490339	-0.247819347
0.5535123752	0.1899808466	-0.163741925
1.7969105388	-0.878328404	-0.496646318
0.4214663279	0.6257232916	0.3643619061
1.0490445493	-0.776179155	1.2392669249
-0.438269594	-1.379120376	-1.064178926
-1.121459921	-0.021929146	-0.731957565
0.701104468	0.4558331377	1.0180881778

De plus, la procédure PRINCOMP de SAS ne fournit pas l'estimation de la matrice de saturation. Il faut la calculer : $\hat{A}_q = V_q\Lambda_q$. Cependant, si elle contient des valeurs très dispersées entre 0 et 1, il sera difficile d'associer des variables entre elles. Ce problème peut également se rencontrer avec le logiciel SPAD qui ne propose de rotation.

On peut cependant souligner deux avantages du logiciel SPAD par rapport à SAS et SPSS : l'interactivité et la possibilité de réaliser facilement des graphiques.

Ainsi, on voit l'avantage d'utiliser la procédure FACTOR de SAS ou le logiciel SPSS pour estimer le modèle d'A.F., car ils fournissent l'estimation de la matrice A_q de saturation et proposent une rotation des facteurs.

6 Conclusion

Le modèle d'A.F. est une méthode factorielle linéaire. Cette technique écrit un ensemble de p variables aléatoires comme une combinaison linéaire de q facteurs non corrélés, communs à toutes les variables, et de p facteurs spécifiques à chaque variable. L'ensemble de ces facteurs communs et uniques reproduit les covariances des variables aléatoires initiales. Ainsi, le modèle d'A.F. permet de résumer au "mieux" l'information contenue dans p variables aléatoires, ou de détecter des facteurs sous-jacents communs dans une problématique particulière. Comme toute méthode factorielle, le point stratégique du modèle d'A.F. réside dans le choix du nombre q de facteurs communs, difficulté à laquelle nous avons souhaité apporter une aide. Cet aide n'est que partielle car nous avons vu que seule une interprétation attentive des résultats et des objectifs de l'étude permet de répondre au problème du choix de q.

Après une présentation synthétique du modèle d'A.F., nous avons décrit les techniques d'estimation et nous avons vu que lorsqu'on estime le modèle d'A.F. via les composantes principales, cela revient à faire une A.C.P.

L'accent a ensuite été mis sur les techniques de rotation des facteurs, qui peuvent s'avérer très utiles. Nous avons montré que, contrairement à ce qu'on peut lire dans certains travaux, la rotation en A.C.P. est également possible à condition d'effectuer convenablement la transformation.

Une application numérique a ensuite été mise en place sur des données concernant la criminalité de villes américaines. Ainsi, l'estimation du modèle d'A.F. couplée à une rotation de type Varimax nous a permis de résumer la criminalité des villes américaines à l'aide de trois facteurs communs : les crimes violents contre la personne, les crimes en rapport avec la maison et les crimes commis à l'extérieur. De plus, cette application a permis de clarifier le vocabulaire utilisé par les logiciels SAS, SPAD et SPSS, réelle source de confusion. L'exemple accompagné de nombreuses illustrations pourra servir de guide, tant pour l'implémentation que pour la lecture des résultats numériques.

7 Annexes

7.1 Annexe 1 : Etude des propriétés des facteurs communs estimés par la méthode des composantes principales

- Par construction, les facteurs communs estimés sont centrés.

- L'hypothèse (H_1) est vérifiée car $\hat{F}_q' M \hat{F}_q = U_q' M^{-1} U_q = I_q$.

- L'hypothèse (H_2) n'est pas nécessairement vérifiée. En effet, cette méthode d'estimation de la matrice Ξ ne garantit pas que $\hat{\Xi} = \hat{E}_q \hat{E}_q'$ soit diagonale. Cependant, en pratique, les termes en dehors de la diagonale de la matrice $\hat{\Xi}$ sont souvent négligeables. Ainsi, la solution trouvée avec cette méthode est souvent une approximation convenable, ce qui explique que cette méthode d'estimation du modèle d'A.F. est très utilisée. On pourrait préconiser à l'utilisateur d'examiner attentivement la matrice $\hat{E}_q \hat{E}_q'$ et de recommencer

l'estimation du modèle avec une autre méthode si les valeurs en dehors de la diagonale de cette matrice sont trop grandes.

• L'hypothèse (H_3) est vérifiée. En effet, en ne retenant que les vecteurs propres associés aux q plus grandes valeurs propres, on a :

$$\tilde{X} = \underbrace{M^{-1}U_q}_{\hat{F}_q}\underbrace{\Lambda_q V_q'}_{\hat{A}_q'} + \hat{E}_q$$

où $\hat{E}_q = M^{-1}U_e\Lambda_e V_e'$, avec U_e, Λ_e et V_e les matrices contenant respectivement les $r-q$ dernières colonnes de U, Λ et V.

On a alors :

$$\begin{aligned} \hat{E}_q' M \hat{F}_q &= \hat{E}_q' U_q \\ &= V_e \Lambda_e U_e' M^{-1} U_q \\ &= 0 \end{aligned} \tag{47}$$

car la matrice U est orthonormée.

• On peut vérifier que les valeurs de la matrice $\hat{A}_q$, notés $\hat{a}_j^\alpha$ sont les corrélations empiriques entre les variables x^j et les facteurs f^α :

$$\begin{aligned} \hat{a}_j^\alpha &= \sum_{i=1}^n z_i^j u_i^\alpha \\ &= \sum_{i=1}^n z_i^j \frac{1}{\sqrt{m}} f_i^\alpha \\ &= \sum_{i=1}^n \left(\frac{x_i^j - \overline{x}^j}{\sqrt{m} s^j} \right) \left(\frac{f_i^\alpha - \overline{f}^\alpha}{\sqrt{m}\sqrt{var(f^\alpha)}} \right) \\ &= corr(x^j, f^\alpha) \end{aligned} \tag{48}$$

7.2 Annexe 2 : Démonstration de la propriété des facteurs et des "loadings" après rotation

• Les facteurs après rotation sont toujours centrés.

• L'hypothèse (H_1) est vérifiée car $\hat{G}_q' M \hat{G}_q = T' \hat{F}_q' M \hat{F}_q T = I_q$.

• L'hypothèse (H_2) est vérifiée car la matrice des erreurs $\hat{E}_q$ n'est pas modifiée.

• L'hypothèse (H_3) est vérifiée car $\hat{E}_q' M \hat{G}_q = \hat{E}_q' M \hat{F}_q T = 0$.

• Les "loadings" après rotation, notés $\hat{b}_j^\alpha$, représentent les corrélations des variables x^j aux facteurs g^α après la rotation.

On a : $\hat{B}_q = \hat{A}_q T = Z'(U_q T) = Z'\breve{U}_q$ où $\breve{U}_q = U_q T$.
De plus, on a :

$$\hat{g}^\alpha = \sqrt{m}\breve{u}^\alpha$$

où $\hat{g}^\alpha$ est la $\alpha^{\text{ème}}$ colonne de la matrice $\hat{G}_q$, et $\breve{u}^\alpha$ est la $\alpha^{\text{ème}}$ colonne de la matrice $\breve{U}_q$.
On en déduit :

$$\begin{aligned}
\hat{b}_j^\alpha &= \sum_{i=1}^{n} z_i^j \breve{u}_i^\alpha \\
&= \sum_{i=1}^{n} z_i^j \frac{1}{\sqrt{m}} g_i^\alpha \\
&= \sum_{i=1}^{n} \left(\frac{x_i^j - \overline{x}^j}{\sqrt{m}s^j}\right)\left(\frac{g_i^\alpha - \overline{g^\alpha}}{\sqrt{m}\sqrt{var(g^\alpha)}}\right) \\
&= corr(x^j, g^\alpha). \qquad (49)
\end{aligned}$$

• La matrice de saturation après la rotation reproduit toujours le modèle de structure de covariance défini par (8) :

$$\hat{B}_q\hat{B}'_q + \hat{\Xi} = \hat{A}_q T T' \hat{A}'_q + \hat{\Xi} = \hat{A}_q\hat{A}'_q + \hat{\Xi} = R. \qquad (50)$$

• Les communalités sont inchangées :

$$\hat{h}_j^2 = \sum_{\alpha=1}^{q} (\hat{b}_j^\alpha)^2 = \sum_{\alpha=1}^{q} (\hat{a}_j^\alpha)^2 \qquad (51)$$

car $\hat{B}_q\hat{B}'_q = (\hat{A}_q T)(\hat{A}_q T)' = \hat{A}_q\hat{A}'_q$.

• La variance totale expliquée par les q facteurs communs n'est pas modifiée :

$$\sum_{j=1}^{p}\sum_{\alpha=1}^{q} (\hat{b}_j^\alpha)^2 = \sum_{i=1}^{p}\sum_{\alpha=1}^{q} (\hat{a}_j^\alpha)^2. \qquad (52)$$

Références

[1] Baccini A., Besse P. (2005), "Data mining I, Exploration Statistique" *http ://www.lsp.ups-tlse.fr/Besse/pub/Explo_stat.pdf.*

[2] Bouveyron C., (2006), *Modélisation et classification des données de grande dimension - Application à l'analyse d'images*, p 45-47, Thèse, Université Joseph Fourier - Grenoble 1.

[3] Fine J. (1993), "Problèmes d'indétermination en analyse en facteurs et analyse en composantes principales optimale", *Revue de Statistique Appliquée*, tome 41, n°4, p 45-72.

[4] Garnett J.-C.(1919), "General ability, cleverness and purpose", *British Journal of Psychiatry*, **9**, p 345-366.

[5] Harman H. H. (1960), *Modern Factor Analysis*, University of Chicago Press.

[6] Heywood H.B. (1931), "On finite sequences of real numbers", *Proceedings of the Royal Society, Series A*, **134**, p 486-501.

[7] Hotelling H. (1933), "Analysis of a complex of statistical variables into principal components", *Journal of Educational Psychology*, **24**, p 417-441.

[8] Jobson J.D. (1992), *Applied Multivariate Data Analysis, Volume II : Categorical and Multivariate Methods*, Springer-Verlag.

[9] Lawley D.N., Maxwell A.E. (1963), *Factor Analysis as a statistical method*, Butterworths London.

[10] Lebart L., Morineau A., Piron M. (1997), *Statistique exploratoire multidimensionnelle, 2e cycle, 2e édition*, Editions Dunod.

[11] Pearson K. (1901), "On lines and planes of closest fit to systems of points in space", *Philosophical Magazine*, **2**, p 559-572.

[12] Rencher, A.C. (2002), *Methods of Multivariate Analysis, Second Edition*, Wiley Series in Probability and Statistics.

[13] Seber G.A.F. (1984), *Multivariate observations*, Wiley Series in Probability and Mathematical Statistics.

[14] Spearman C. (1904), "General intelligence, objectively determined and measured", *American Journal of Psychology*, **15**, p 201-293.

[15] Tipping M.E, Bishop C.M. (1999), "Probabilistic Principal Component Analysis", *Journal of the Royal Statistical Society, Series B*, **61**, Part 3, p 611-622.

4.3 PCA and PMF based methodology for air pollution sources identification and apportionment

ENVIRONMETRICS
Environmetrics (2008)
Published online in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/env.963

PCA- and PMF-based methodology for air pollution sources identification and apportionment

Marie Chavent[1,2*,†], Hervé Guégan[3], Vanessa Kuentz[1,2], Brigitte Patouille[1] and Jérôme Saracco[1,2,4]

[1]*Université de Bordeaux, IMB, CNRS, UMR 5251, France*
[2]*INRIA Bordeaux Sud-Ouest, CQFD team, France*
[3]*ARCANE-CENBG, Gradignan, France*
[4]*Université Montesquieu – Bordeaux IV, GREThA, CNRS, UMR 5113, France*

SUMMARY

Air pollution is a wide concern for human health and requires the development of air quality control strategies. In order to achieve this goal pollution sources have to be accurately identified and quantified. The case study presented in this paper is part of a scientific project initiated by the French Ministry of Ecology and Sustainable Development. For the following study measurements of chemical composition data for particles have been conducted on a French urban site. The first step of the study consists in the identification of the sources profiles which is achieved through principal component analysis (PCA) completed by a rotation technique. Then the apportionment of the sources is evaluated with a receptor modeling using positive matrix factorization (PMF) as estimation method. Finally the joint use of these two statistical methods enables to characterize and apportion five different sources of fine particulate emission. Copyright © 2008 John Wiley & Sons, Ltd.

KEY WORDS: pollution sources; principal component analysis; receptor modeling; positive matrix factorization

1. INTRODUCTION

Particulate pollution, also known as particulate matter or PM, comes from various sources such as factory and utility smokestacks, vehicle exhaust, wood burning, mining, construction activity, or agriculture. This air pollution is a complex mixture of extremely small particles and liquid droplets suspended in the air we breathe. High concentrations of particles have been found to present a serious danger to human health. Particles of special concern to the protection of lung health are those known as fine particles (PM2.5), less than 2.5 microns in diameter. Development of PM2.5 control strategies is then a wide preoccupation of environmental protection agencies. Since strategies to improve ambient air quality involve the reduction of emissions from primary sources, it is important to be able to identify and apportion the contributions of these sources.

*Correspondence to: M. Chavent, IMB, Université Bordeaux1, 351 Cours de la libération, 33051 Talence, France.
†E-mail: chavent@math.u-bordeaux1.fr

Received 25 June 2007
Accepted 2 September 2008

Receptor modeling, using measurements of chemical composition data for particles on a sample site, is often a reliable way to provide information regarding source characteristics (Gabriel and Zamir, 1979) (Hopke, 1991). Some multivariate receptor models are based on the analysis of the correlations between measured concentrations of chemical species, assuming that highly correlated compounds come from the same source. One commonly used multivariate receptor model is principal component analysis (PCA) (Gabriel and Zamir, 1979) (Jolliffe, 2002), successfully applied to identify sources in several studies. However, PCA is not a convenient tool for quantifying sources contributions. Therefore specific methods, such as positive matrix factorization (PMF) Paatero and Tapper (1994), have been specifically developed in order to address this problem.

The case study presented in this paper is a statistical part of the scientific program PRIMEQUAL,[‡] initiated by the MEDD[§] and the ADEME,[‖] about atmospheric pollution and its impact. We propose and apply a methodology for determining particulate emission sources and their concentrations at the urban site of Anglet located in the southwest of France. The following three-step process has been implemented:

1. PM2.5 were collected with sequential fine particle samplers on the receptor site and the chemical composition of each sampler was measured with particle induced X-ray emission (PIXE) method. After several pre-treatments a data matrix of chemical compounds concentrations in each sampler was selected.
2. PCA was applied to this data matrix and the standardized principal components were rotated, in order to identify possible sources.
3. PMF was applied to the same data matrix and the results were normalized in order to find components with physical interpretations (concentration of each source in each particle sampler).

Steps 2 and 3 are independent but results of step 2 will be used to validate results of step 3.

2. DATA

The air pollution receptor modeling (n, p) data matrix consists of the measurements of p chemical species in n particle samplers. In this application, $n = 61$ samplers of PM2.5 were collected with sequential fine particle samplers by AIRAQ[¶] in the French urban site of Anglet, every 12 h, in December 2005. There are two samples per 24 h: one for the day (7AM:7PM) and one for the night (7PM:7AM). The mass and volume, represented by the concentration C in ng/m^3, of each particle sampler were measured with the PIXE method by ARCANE–CENBG,[**] as well as the concentrations of $p = 15$ chemical elements (Al, Si, P, S, Cl, K, Ca, Ti, Mn, Fe, Ni, Cu, Zn, Br, Pb). Table 1 gives a subset of the data in their initial form.

First Ni and Ti elements which were frequently present in concentrations below the detection limits (BDL) were excluded and only 13 elements were selected. Then the few BDL data remaining in this selected dataset were replaced by values corresponding to one-half of the appropriate analytical detection limit. Al, Si, S, and Fe elements were respectively replaced by the compounds Al_2O_3, SiO_2, SO_4, Fe_2O_3.

[‡]Projet de Recherche Interorganisme pour une MEilleure QUalité de l'Air à l'échelle Locale.
[§]French ministry of Ecology and Sustainable Development.
[‖]French Environment and Energy Management Agency.
[¶]Réseau de surveillance de la qualité de l'air en Aquitaine.
[**]Atelier Régional de Caractérisation par Analyse Nuclaire Elémentaire – Centre d'Etudes Nucléaires de Bordeaux Gradignan.

DOI: 10.1002/env

Table 1. Subset of the original data table

Date	C	Al	Si	...	K	Ca	...	Br	Pb
23-11-05 day	7300	92	75	...	163	35	...	7	10
23-11-05 night	9600	135	90	...	211	23	...	7	77
24-11-05 day	11000	175	137	...	241	69	...	8	19
24-11-05 night	5300	36	31	...	94	44	...	9	7
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
24-12-05 day	21000	< 2	< 1	...	266	< 1	...	7	18
24-12-05 night	18100	18	< 1	...	307	< 1	...	7	19
25-12-05 day	23300	37	22	...	311	12	...	7	14
25-12-05 night	36100	< 2	< 1	...	277	< 1	...	10	19

Then the remaining concentration, called C_{org}, which was not measured with the previous compounds, was calculated for each particle sampler:

$$C_{\text{org}} = C - (Al_2O_3 + SiO_2 + P + SO_4 + Fe_2O_3 + Cl + K + Ca + Mn + Cu + Zn + Br + Pb)$$

Finally the (n, p) concentration matrix $\mathbf{X} = (x_{ij})$ used in the receptor model has $n = 61$ rows and $p = 14$ columns (Al_2O_3, SiO_2, P, SO_4, Cl, K, Ca, Mn, Fe_2O_3, Cu, Zn, Br, Pb, C_{org}). The coefficient x_{ij} is the concentration of the jth chemical compound in the ith sampler. One can observe that C_{org} represents the largest concentration in the particle samplers and then the largest part (almost all) of PM2.5. The discovery of its origin is a key point in the results.

3. SOURCES IDENTIFICATION

In order to identify the sources of fine particulate emission we applied PCA to the concentration matrix $\mathbf{X}$ and completed it by an orthogonal rotation of the standardized principal components. Then we have associated groups of correlated chemical compounds to air pollution sources.

First, we will give a short theoretical reminder of factor analysis with PCA estimation method. Then we will interpret the corresponding results on the air pollution data.

3.1. Factor analysis with PCA estimation method

3.1.1. Notations. We consider a (n, p) numerical data matrix $\mathbf{X}$ where n objects are described on $p < n$ variables $x_1, \ldots, x_p$. We will note $\mathbf{x}_j$ a column of $\mathbf{X}$. Let $\tilde{\mathbf{X}} = (\tilde{x}_{ij})_{n,p}$ be the standardized data matrix with $\tilde{x}_{ij} = (x_{ij} - \bar{x}_j)/s_j$ where $\bar{x}_j$ and s_j are respectively the empirical mean and the standard deviation of x_j.

Let $\mathbf{R} = \tilde{\mathbf{X}}'\mathbf{M}\tilde{\mathbf{X}}$ be the empirical correlation matrix of $x_1, \ldots, x_p$, where $\mathbf{M} = \frac{1}{m}\mathbf{I}_n$ with $m = n$ or $n - 1$ depending on the choice of the denominator of s_j. The correlation matrix can also be written $\mathbf{R} = \mathbf{Z}'\mathbf{Z}$ with $\mathbf{Z} = \mathbf{M}^{1/2}\tilde{\mathbf{X}}$.

Let us denote by $r \leq p$ the rank of $\mathbf{Z}$ and consider the singular value decomposition (SVD) of $\mathbf{Z}$:

$$\mathbf{Z} = \mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{V}' \tag{1}$$

DOI: 10.1002/env

where:

- Λ is the (r, r) diagonal matrix of the r nonnull eigenvalues λ_k, $k = 1, \ldots, r$, of the matrix $\mathbf{Z}'\mathbf{Z}$ (or $\mathbf{Z}\mathbf{Z}'$), ordered from largest to smallest;
- $\mathbf{U}$ is the (n, r) orthonormal matrix of the r eigenvectors $\mathbf{u}_k$, $k = 1, \ldots, r$ of $\mathbf{Z}\mathbf{Z}'$ associated with the first r eigenvalues;
- $\mathbf{V}$ is the (p, r) orthonormal matrix of the r eigenvectors $\mathbf{v}_k$, $k = 1, \ldots, r$ of $\mathbf{Z}'\mathbf{Z} = \mathbf{R}$ associated with the first r eigenvalues.

From the SVD of $\mathbf{Z}$ we deduce the following decomposition of $\tilde{\mathbf{X}}$:

$$\tilde{\mathbf{X}} = \mathbf{M}^{-1/2}\mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{V}' \tag{2}$$

3.1.2. Factor analysis model. The basic idea underlying factor analysis (using correlation matrix) is that the p observed standardized variables $\tilde{x}_1, \ldots, \tilde{x}_p$ can be expressed, to the exception of an error term, as linear functions of $q < p$ unobserved variables or common factors $f_1, \ldots, f_q$. The observed standardized matrix $\tilde{\mathbf{X}}$ being given, factor analysis model can be expressed in its simplified form as

$$\tilde{\mathbf{X}} = \mathbf{F}\mathbf{A}' + \mathbf{E} \tag{3}$$

where $\mathbf{F}$ is the (n, q) matrix of unobserved values of the factors and $\mathbf{A}$ is the (p, q) matrix of unknown loadings providing the information relating the factors f_k to the original variables $x_1, \ldots, x_p$. The (n, p) matrix $\mathbf{E}$ is the rest of the approximation of $\tilde{\mathbf{X}}$ with $\hat{\tilde{\mathbf{X}}} = \mathbf{F}\mathbf{A}'$.

Several approaches were developed to estimate the model (principal factor, maximum likelihood ...) but PCA is often used in practice.

3.1.3. PCA. In PCA, when $q = r$, Equation (2) is rewritten as

$$\tilde{\mathbf{X}} = \mathbf{\Psi}\mathbf{V}' \tag{4}$$

where $\mathbf{\Psi} = \mathbf{M}^{-1/2}\mathbf{U}\mathbf{\Lambda}^{1/2}$ is the principal component scores matrix. The columns of $\mathbf{\Psi}$ are the r principal components $\boldsymbol{\psi}_k = (m\lambda_k)^{1/2}\mathbf{u}_k$, $k = 1, \ldots, r$. Since $\mathbf{U}$ and $\mathbf{V}$ are orthonormal we have $\boldsymbol{\psi}_k = \tilde{\mathbf{X}}\mathbf{v}_k$ and $\mathrm{Var}(\boldsymbol{\psi}_k) = \lambda_k$.

3.1.4. Estimation of the factor model using PCA. When $q = r$ Equation (2) is rewritten as

$$\tilde{\mathbf{X}} = \mathbf{F}\mathbf{A}' \tag{5}$$

where $\mathbf{F} = \mathbf{M}^{-1/2}\mathbf{U}$ is the factor scores matrix and $\mathbf{A} = \mathbf{V}\mathbf{\Lambda}^{1/2}$ is the loadings matrix. The columns $\mathbf{f}_k = m^{1/2}\mathbf{u}_k$ of the matrix $\mathbf{F}$ are realizations of the r factors f_k, $k = 1, \ldots, r$. The coefficient a_{jk} of the matrix $\mathbf{A}$ is equal to the empirical correlation between $\mathbf{x}_j$ and $\mathbf{f}_k$. Since $\mathbf{U}$ and $\mathbf{V}$ are orthonormal we have $\mathbf{f}_k = \lambda_k^{-1/2}\tilde{\mathbf{X}}\mathbf{v}_k = \lambda_k^{-1/2}\boldsymbol{\psi}_k$ for $k = 1, \ldots, r$ and $\mathrm{Var}(\mathbf{f}_k) = 1$. Then $\mathbf{f}_k$ is also the standardized principal component $\boldsymbol{\psi}_k$.

DOI: 10.1002/env

When the user only retains the first $q < r$ eigenvalues of $\mathbf{\Lambda}$ the corresponding approximation of $\tilde{\mathbf{X}}$ in Equation (3) is then

$$\hat{\tilde{\mathbf{X}}}_q = \mathbf{F}_q \mathbf{A}'_q$$

where $\mathbf{F}_q$ and $\mathbf{A}_q$ are the matrices $\mathbf{F}$ and $\mathbf{A}$ reduced to their first q columns. $\mathbf{F}_q$ is then the matrix of the first q standardized principal components.

3.1.5. Rotation of the standardized principal components. Let $\mathbf{T}$ be an orthogonal transformation matrix corresponding to an orthogonal rotation of the q axes in a p-dimensional space: $\mathbf{T}\mathbf{T}' = \mathbf{T}'\mathbf{T} = \mathbf{I}_q$.

The orthogonal rotation is applied to the standardized principal components:

$$\hat{\breve{\mathbf{X}}}_q = \mathbf{F}_q \mathbf{T} (\mathbf{A}_q \mathbf{T})'$$

The q rotated standardized principal components $\breve{\mathbf{f}}_k^q$ are the q columns of the matrix $\breve{\mathbf{F}}_q = \mathbf{F}_q \mathbf{T}$. They have the property of being mutually orthogonal and of variance equal to 1.

In order to be able to interpret the $\breve{\mathbf{f}}_k^q$'s (also called rotated factors) let us remark that the coefficients $\breve{a}_{jk}^q$ of the matrix $\breve{\mathbf{A}}_q = \mathbf{A}_q \mathbf{T}$ are equal to the empirical correlations between the rotated factors $\breve{\mathbf{f}}_k^q$ and $\mathbf{x}_j$.

From a practical point of view the orthogonal transformation matrix $\mathbf{T}$ is defined in order to construct a matrix $\breve{\mathbf{A}}_q$ such that each variable x_j is clearly correlated to one of the rotated factor $\breve{\mathbf{f}}_{k^*}^q$ (that is $\breve{a}_{jk^*}^q$ close to 1) and not to the other rotated factors (that is $\breve{a}_{jk}^q$ close to 0 for $k \neq k^*$). The most popular rotation technique is varimax which seeks rotated loadings maximizing the variance of the squared loadings in each column of $\breve{\mathbf{A}}_q$.

3.2. Results

We applied the FACTOR procedure of SAS to the data matrix $\mathbf{X}$ introduced in Section 2. The following options were used: METHOD = PRIN, ROTATE = VARIMAX, and NFACTORS = 5. The number $q = 5$ of factors was chosen both because it allows to explain 90, 93% of the total variance and because decompositions in a larger number of factors did not give satisfactory interpretations. Table 2 gives the matrix $\breve{\mathbf{A}}_5$ of the loadings after rotation.

This matrix can be used to associate, when possible, sources with the rotated factors. Indeed we observe for each factor the strongly correlated compounds. For instance Zn and Pb are strongly correlated to $\breve{\mathbf{f}}_3^5$. Because Zn and Pb are known to have industrial origin this rotated factor is associated to the industrial pollution source. The same way, the element Cl is strongly correlated to $\breve{\mathbf{f}}_5^5$ which is then associated with seasalt pollution. Possible associations between the five rotated factors and five pollution sources are given in Table 3.

In order to confirm these associations we have confronted the rotated factors $\breve{\mathbf{f}}_k^5$ with external parameters such as meteorological data (temperatures and wind directions) and the periodicity night/day of the sampling. The coefficient $\breve{f}_{ik}^5$ represents a "relative" contribution of the source k to the particle sampler i. Figure 1(a) gives for instance the evolution of the relative contribution of the "vehicle" source associated with $\breve{\mathbf{f}}_4^5$. The night samplers have been distinguished from the day ones, enabling to notice that the contribution of this source is stronger during the day than during the night. It then confirms that this source corresponds to vehicle pollution. The same way, Figure 1(b) gives the evolution of the

DOI: 10.1002/env

Table 2. Correlations between the chemical compounds and the rotated factors

	$\breve{\mathbf{f}}_1^5$	$\breve{\mathbf{f}}_2^5$	$\breve{\mathbf{f}}_3^5$	$\breve{\mathbf{f}}_4^5$	$\breve{\mathbf{f}}_5^5$
Al_2O_3	0.981	0.087	−0.042	0.070	−0.038
SiO_2	0.979	0.012	−0.055	0.104	−0.074
P	0.972	0.090	−0.017	0.071	−0.092
SO_4	−0.028	0.765	0.247	0.180	−0.345
Cl	−0.153	−0.274	−0.136	−0.181	0.879
K	0.597	0.716	0.111	0.233	0.031
Ca	0.608	0.091	−0.113	0.560	0.272
Mn	−0.279	0.119	0.604	0.582	−0.238
Fe_2O_3	0.198	0.282	0.289	0.848	−0.112
Cu	0.213	0.359	0.161	0.816	−0.149
Zn	−0.029	0.053	0.977	0.129	−0.044
Br	0.490	0.615	0.097	0.281	0.392
Pb	0.004	0.163	0.969	0.126	−0.054
C_{org}	−0.018	0.893	0.021	0.222	−0.160

Table 3. Factor–source associations

Factor 1	Soil dust
Factor 2	Combustion
Factor 3	Industry
Factor 4	Vehicle
Factor 5	Sea

relative contribution of the "combustion" source associated with $\breve{\mathbf{f}}_2^5$. We can notice an increase in the contribution of this source in the middle of the sampling period, which corresponds to a decrease in the temperature measured on the sampling site, see Figure 1(c). This confirms that this source corresponds to combustion and heating pollution.

The identification of the sources using PCA is only the first step of a more complex process which consists in quantifying the sources. Although it is essential to identify the sources, the true challenge is to define, in percentage of total fine dust mass, the quantity of each of these sources.

4. SOURCES APPORTIONMENT

In order to apportion the sources of fine particulate emission we have applied PMF to the concentration matrix $\mathbf{X}$ and then normalized the results to find components with physical interpretation.

4.1. Receptor modeling with PMF estimation method

The basic problem is to estimate, from the data matrix $\mathbf{X}$, the number q of sources, their compositions, and their contributions. To address this problem we consider the mass balance equation:

$$x_{ij} = \sum_{k=1}^{q} g_{ik} b_{jk} \tag{6}$$

DOI: 10.1002/env

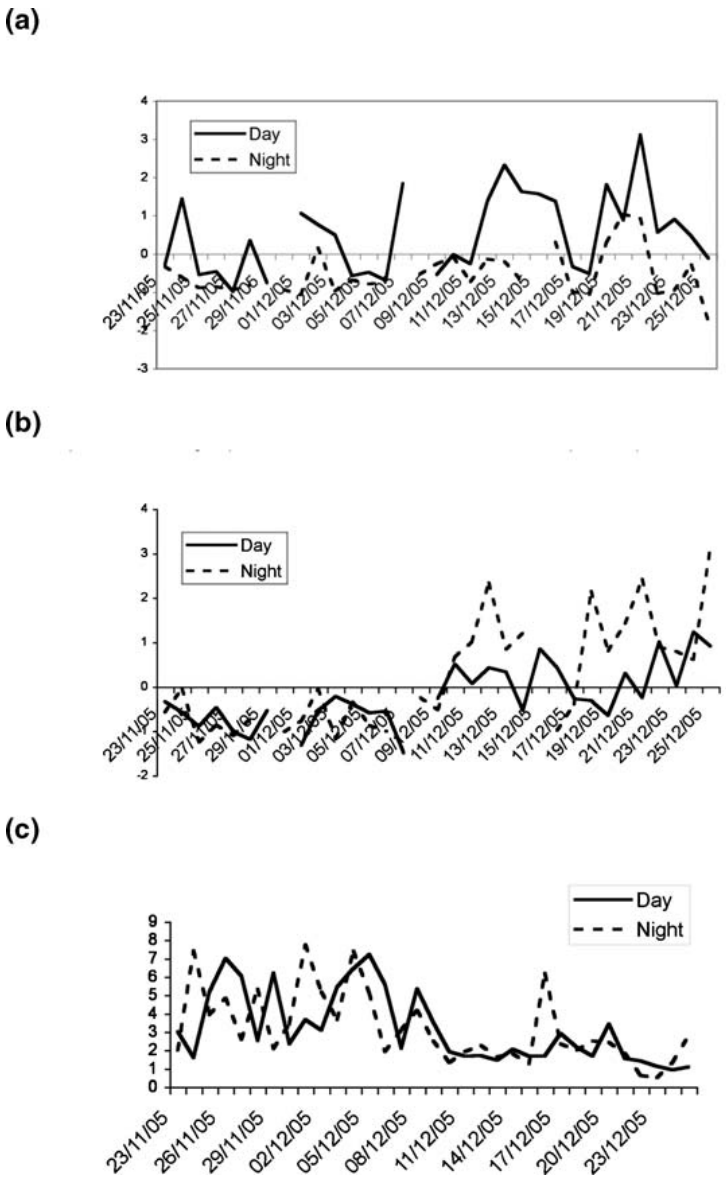

Figure 1. Evolution of (a) factor 4 associated to car pollution, (b) factor 2 associated to heating pollution, and (c) the temperatures

where

– g_{ik} is the concentration in particles from source k in the particle sampler i;
– b_{jk} is the mass fraction (percentage) of species j in source k.

In the receptor modeling vocabulary the b_{jk}'s are the sources compositions (or sources profiles) and the g_{ik}'s are the sources contributions. The product $g_{ik}b_{jk}$ is then the approximation of the concentration in the sampler i in particles from the jth species coming from the source k. Let m_{ijk} be the mass in the sampler i of species j from source k, and let m_{ik} be the mass in the sampler i from source k. Then

DOI: 10.1002/env

$b_{jk} = \frac{m_{ijk}}{m_{ik}}$ is the percentage of species j emitted by source k when sampler i was collected. Since the mass fraction b_{jk} is independent from i the sources profiles are assumed to be constant during the sampling period.

In matrix form Equation (6) can be written as

$$\mathbf{X} = \mathbf{G}\mathbf{B}' \tag{7}$$

where $\mathbf{G}$ is a (n, q) matrix of sources contributions and $\mathbf{B}$ is a (p, q) matrix of sources compositions. Approximations of $\mathbf{G}$ and $\mathbf{B}$ are obtained from the data matrix $\mathbf{X}$ and a previously selected number q of sources by applying the two following steps:

- *PMF step*. The matrix $\mathbf{X}$ is factorized in a product $\mathbf{H}\mathbf{C}'$ of rank q under constraints of positivity of the coefficients. This condition is required by physical reality of non-negativity of sources compositions and contributions: $g_{ik} \geq 0$ and $b_{jk} \geq 0$.
- *Scaling step*. The columns of the approximations $\hat{\mathbf{H}}$ and $\hat{\mathbf{C}}$ obtained in the previous step are scaled in order to get the approximations $\hat{\mathbf{G}}$ and $\hat{\mathbf{B}}$. The scaling coefficients are defined to fulfill other physical constraints of the sources compositions and contributions.

Let us assume now that there are at least as many species as sources.

4.1.1. PMF step. Given a matrix $\mathbf{X}$ and a previously selected rank $q \leq p$ the aim of PMF (or non-negative matrix factorization) is to approximate $\mathbf{X}$ by a product of two matrices $\mathbf{H}\mathbf{C}'$ (with $\mathbf{H}$ of dimension (n, q), and $\mathbf{C}$ of dimension (p, q)) subject to $h_{ik} \geq 0$ and $c_{jk} \geq 0$. Matrices $\mathbf{H}$ and $\mathbf{C}$ are obtained by minimization of a least squares function $Q(\mathbf{H}, \mathbf{C})$ under constraints of positivity.

When the constraints of positivity are ignored the ordinary SVD of $\mathbf{X}$, that is, $\mathbf{X} = \mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{V}'$ with $\mathbf{U}'\mathbf{U} = \mathbf{I}_r$ and $\mathbf{V}'\mathbf{V} = \mathbf{I}_r$, provides a sequence of approximations $\mathbf{H}\mathbf{C}'$ of rank $q = 1, \ldots, r$ which minimizes the square of the Euclidean norm of the residual matrix $\mathbf{L} = \mathbf{X} - \mathbf{H}\mathbf{C}'$:

$$Q_1(\mathbf{H}, \mathbf{C}) = \sum_{i=1}^{n}\sum_{j=1}^{p} l_{ij}^2 = \sum_{i=1}^{n}\sum_{j=1}^{p}\left(x_{ij} - \sum_{k=1}^{q} h_{ik}c_{jk}\right)^2 \tag{8}$$

In Equation (8) the rows and the columns of $\mathbf{X}$ have the same weight. Let us denote now ω_i the weight of the ith row and ϕ_j the weight of the jth column of $\mathbf{X}$. Let Ω and Φ be two diagonal matrices respectively with elements ω_i, $i = 1, \ldots, n$ and ϕ_j, $j = 1, \ldots, p$. The generalized SVD of $\mathbf{X}$, that is $\mathbf{X} = \mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{V}'$ with $\mathbf{U}'\Omega\mathbf{U} = \mathbf{I}_r$ and $\mathbf{V}'\Phi\mathbf{V} = \mathbf{I}_r$, provides a sequence of approximations $\mathbf{H}\mathbf{C}'$ which minimizes

$$Q_2(\mathbf{H}, \mathbf{C}) = \sum_{i=1}^{n}\sum_{j=1}^{p} \omega_i\phi_j\left(x_{ij} - \sum_{k=1}^{q} h_{ik}c_{jk}\right)^2 \tag{9}$$

note that this generalized SVD of $\mathbf{X}$ is obtained by finding the ordinary SVD of $\Omega^{1/2}\mathbf{X}\Phi^{1/2}$.

A third type of approximation of $\mathbf{X}$ is defined by minimizing

$$Q_3(\mathbf{H}, \mathbf{C}) = \sum_{i=1}^{n}\sum_{j=1}^{p} w_{ij}\left(x_{ij} - \sum_{k=1}^{q} h_{ik}c_{jk}\right)^2 \tag{10}$$

DOI: 10.1002/env

but this approximation cannot be obtained by SVD unless the w_{ij}'s can be written as products $w_{ij} = \omega_i \phi_j$. Gabriel and Zamir (1979) suggest a number of ways in which special cases of this weighted least squares analysis may be used.

The PMF algorithm developed by Paatero and Tapper (1994) in the context of receptor modeling minimizes Equation (10) with $w_{ij} = 1/\sigma_{ij}^2$. The coefficient σ_{ij} is a measure of uncertainty of the observation x_{ij}. Given the σ_{ij}'s this method searches **H** and **C** minimizing

$$Q_4(\mathbf{H}, \mathbf{C}) = \sum_{i=1}^{n} \sum_{j=1}^{p} \left(\frac{x_{ij} - \sum_{k=1}^{q} h_{ik} c_{jk}}{\sigma_{ij}} \right)^2 \tag{11}$$

subject to $h_{ik} \geq 0$ and $c_{jk} \geq 0$.

Polissar *et al.* (1998) propose several definitions for calculating the σ_{ij}'s from a matrix **X** of chemical species concentrations. The one used in the PMF program of the US Environment Protection Agency[††] is the following:

$$\sigma_{ij} = \begin{cases} 2\text{LD} & \text{if } x_{ij} \leq \text{LD} \\ \sqrt{(\theta_j x_{ij})^2 + \text{LD}^2} & \text{if } x_{ij} > \text{LD} \end{cases} \tag{12}$$

where LD is the limit of detection for the jth species and θ_j is a percentage of uncertainty associated with the jth species. One can note the subjectivity of this definition which changes from an article to another using this PMF method for sources apportionment.

In this case study, we have made a different choice for the σ_{ij}'s. Indeed, dealing with variables measured on very different scales is a problem when approximating **X** globally on all the variables. Minimizing the unweighed quadratic error Q_1 in Equation (8) gives better approximations for the columns of **X** corresponding to variables with large dispersion. Hence we have chosen to use Q_4 with $\sigma_{ij} = s_j$, the empirical standard deviation of the jth variable.

4.1.2. Scaling step. Let $\hat{\mathbf{X}} = \hat{\mathbf{H}}\hat{\mathbf{C}}'$ be the product calculated by PMF. Since $\hat{x}_{ij} = \sum_{k=1}^{q} \hat{h}_{ik}\hat{c}_{jk} = \sum_{k=1}^{q} \hat{h}_{ik} \frac{\beta_k}{\beta_k} \hat{c}_{jk}$ the matrix $\hat{\mathbf{X}}$ can be written as

$$\hat{\mathbf{X}} = \breve{\mathbf{H}}\breve{\mathbf{C}}' \tag{13}$$

with $\breve{h}_{ik} = \hat{h}_{ik}\beta_k$ and $\breve{c}_{jk} = \frac{\hat{c}_{jk}}{\beta_k}$.

The aim of scaling is then to define the scaling constants β_k, $k = 1, \ldots, q$ such that $\breve{\mathbf{H}}$ and $\breve{\mathbf{C}}$ verify the physical conditions of the matrices **G** and **B** of the mass balance Equation (6). We are going to use the two following conditions:

– Let γ_i be the concentration in the ith sampler:

$$\gamma_i = \sum_{k=1}^{q} g_{ik} \tag{14}$$

[††]E.P.A. PMF 1.1 Users's guide, http://www.epa.gov/heasd/products/pmf/pmf.htm

DOI: 10.1002/env

In other words, the sum of the concentrations of the sources adds up to the total concentration of the samplers.

– If the sum of the concentrations of the observed species adds up to (resp. is lower than) the total concentration of the samplers, then the sum of all species in each source profile is equal to (resp. lower than) unity:

$$\begin{cases} \sum_{j=1}^{p} b_{jk} = 1 & \text{if } \sum_{j=1}^{p} x_{ij} = \gamma_i, \\ \sum_{j=1}^{p} b_{jk} < 1 & \text{otherwise} \end{cases} \tag{15}$$

First, we consider the case where $\forall i, \sum_{j=1}^{p} x_{ij} = \gamma_i$. From the physical constraints (15) the scaling coefficients β_k can be calculated in two different ways.

– Directly from $\hat{\mathbf{X}} = \hat{\mathbf{H}}\hat{\mathbf{C}}'$ we get

$$\gamma_i = \sum_{j=1}^{p} x_{ij} = \sum_{j=1}^{p} \left(\sum_{k=1}^{q} \hat{h}_{ik}\hat{c}_{jk} + \hat{l}_{ij} \right) = \sum_{k=1}^{q} \hat{h}_{ik} \left(\sum_{j=1}^{p} \hat{c}_{jk} \right) + \sum_{j=1}^{p} \hat{l}_{ij}$$

we can set

$$\hat{\beta}_k = \sum_{j=1}^{p} \hat{c}_{jk} \tag{16}$$

Then we have the following approximations $\hat{\mathbf{G}}$ and $\hat{\mathbf{B}}$ of $\mathbf{G}$ and $\mathbf{B}$: $\hat{b}_{jk} = \hat{c}_{jk} / \sum_{j=1}^{p} \hat{c}_{jk}$ which satisfies constraint (15), and $\hat{g}_{ik} = \hat{h}_{ik}(\sum_{j=1}^{p} \hat{c}_{jk})$ which satisfies constraint (14) with an error sum of squares equal to $\sum_{i=1}^{n}(\sum_{j=1}^{p} \hat{l}_{ij})^2$.

– Considering the linear approximation of γ_i

$$\gamma_i = \sum_{k=1}^{q} \beta_k \hat{h}_{ik} + e_i \tag{17}$$

we search $\beta = (\beta_1, \ldots, \beta_q)'$ minimizing the error sum of squares

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(\gamma_i - \sum_{k=1}^{q} \beta_k \hat{h}_{ik} \right)^2$$

A well-known solution to this minimization problem is

$$\hat{\hat{\beta}} = (\hat{\mathbf{H}}'\hat{\mathbf{H}})^{-1}\hat{\mathbf{H}}\gamma \tag{18}$$

with $\gamma = (\gamma_1, \ldots, \gamma_n)'$. The corresponding approximations $\hat{\hat{\mathbf{G}}}$ and $\hat{\hat{\mathbf{B}}}$ are such that $\hat{\hat{b}}_{jk} = \hat{c}_{jk} / \hat{\hat{\beta}}_k$ does not satisfy constraint (15), and $\hat{\hat{g}}_{ik} = \hat{h}_{ik}\hat{\hat{\beta}}_k$ satisfies constraint (14) with an error sum of squares equal to $\sum_{i=1}^{n}(\hat{e}_{ij})^2$ with $\hat{e}_{ij} = \gamma_i - \sum_{k=1}^{q} \hat{\hat{\beta}}_k \hat{h}_{ik}$.

DOI: 10.1002/env

Since $\sum_{i=1}^{n}(\hat{e}_{ij})^2$ is the minimum error sum of squares we have:

$$\sum_{i=1}^{n}(\hat{e}_{ij})^2 \leq \sum_{i=1}^{n}\left(\sum_{j=1}^{p}\hat{l}_{ij}\right)^2 \tag{19}$$

Obviously, in case of equality in Equation (19), we get $\hat{\beta} = \hat{\hat{\beta}}$ which means that we have simultaneously the sum of all species in each source profile which is unity and the sum of the concentrations of the sources best fitting (for the least sum of squares error) the total concentration of the samples.

Comparing $\sum_{i=1}^{n}(\hat{e}_{ij})^2$ with $\sum_{i=1}^{n}(\sum_{j=1}^{p}\hat{l}_{ij})^2$ or equivalently comparing $\hat{\beta}$ with $\hat{\hat{\beta}}$ provides a confirmation that the information given by the columns of $\hat{\mathbf{H}}$ are coherent with the physical model we try to approximate. It is hence a first good way to validate the results.

If we consider now the case where $\sum_{j=1}^{p} x_{ij} < \gamma_i$ the scaling coefficients cannot be directly calculated from $\hat{\mathbf{X}} = \hat{\mathbf{H}}\hat{\mathbf{C}}'$ since $\sum_{j=1}^{p} b_{jk} < 1$. They are then evaluated with Equation (18).

A second way to validate the results is based on the regression of γ_i either on $\hat{\hat{\gamma}}_i = \sum_{k=1}^{q}\hat{\hat{\beta}}_k\hat{h}_{ik}$ or $\hat{\gamma}_i = \sum_{k=1}^{q}\hat{\beta}_k\hat{h}_{ik}$, depending on the choice of the scaling coefficients.

4.2. The results

We have applied the PMF algorithm to the concentration matrix $\mathbf{X}$ with $q = 5$ sources. The choice of the number of sources rises from the PCA results. The introduction of C_{org} yields $\sum_{j=1}^{p} x_{ij} = \gamma_i$, then the scaling coefficients $\hat{\beta}_k$ have been calculated from Equation (16).

We thus have the following numerical results:

- the (61, 5) matrix $\hat{\mathbf{G}}$ of the approximated concentrations of the five sources in the 61 samples,
- the (14, 5) matrix $\hat{\mathbf{B}}$ of the approximated compositions (profiles) of the five sources on the 14 compounds.

4.2.1. Quality of the model approximation. Since we are in the case where $\forall i$, $\sum_{j=1}^{p} x_{ij} = \gamma_i$ we can evaluate the quality of the approximation of $\mathbf{X}$ by $\hat{\mathbf{G}}\hat{\mathbf{B}}'$ using the two methods mentioned above. We can compare the scaling coefficients $\hat{\beta}_k$ and $\hat{\hat{\beta}}_k$. Table 4 clearly shows that the $\hat{\beta}_k$'s are close to the $\hat{\hat{\beta}}_k$'s. Moreover Figure 2 also shows a good fitting of the γ_i's by the $\hat{\gamma}_i$'s.

Table 4. The scaling coefficients

	$\hat{\beta}_k$	$\hat{\hat{\beta}}_k$
$k = 1$	147.1	158.6
$k = 2$	91.5	89.6
$k = 3$	73.5	76.8
$k = 4$	251.9	251.9
$k = 5$	51.1	73.2

DOI: 10.1002/env

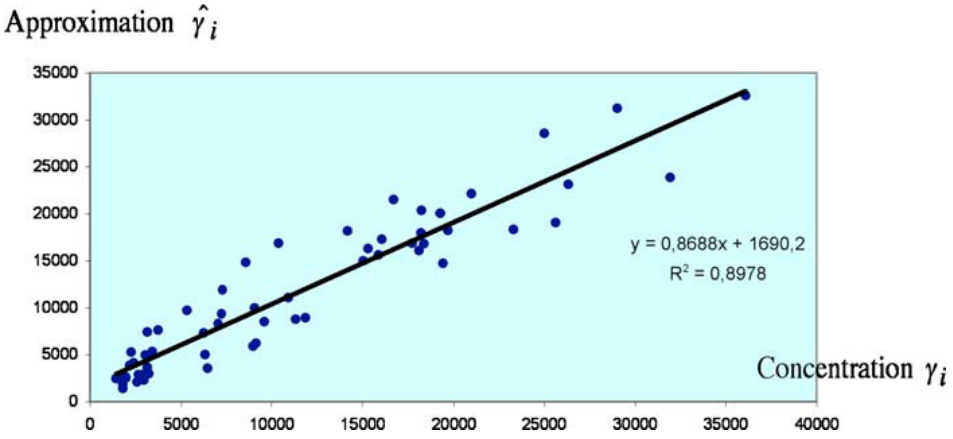

Figure 2. Adjustment of γ by $\hat{\gamma}$. This figure is available in color online at www.interscience.wiley.com/journal/env

Table 5. Relative contributions of the sources to the chemical compounds

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
Al_2O_3	100.0	0.0	0.0	0.0	0.0
$Si0_2$	100.0	0.0	0.0	0.0	0.0
P	81.5	0.5	3.9	8.2	6.0
SO_4	4.5	9.5	10.7	67.9	7.5
Cl	0.0	0.0	0.0	0.0	100.0
K	38.8	0.0	4.4	56.7	0.2
Ca	42.0	39.6	0.0	0.0	18.4
Mn	0.0	54.9	33.1	8.5	3.5
Fe_2O_3	19.0	59.2	14.4	7.4	0.0
Cu	18.5	56.8	9.1	15.6	0.0
Zn	9.0	0.5	87.5	0.0	3.1
Br	19.4	12.1	5.7	33.4	29.4
Pb	10.7	0.0	81.4	7.9	0.0
C_{org}	0.0	8.0	0.0	92.0	0.0

4.2.2. Sources identification. In practice the knowledge of $\hat{\mathbf{G}}$ and $\hat{\mathbf{B}}$ does not give direct indications on the nature of the sources. To try to discover the nature of the five sources we want to calculate their relative contribution to each of the 14 chemical compounds. In order to do that we need to work with the masses instead of the concentrations. Then we calculate, from $\hat{\mathbf{G}}$, the approximation of the total mass of particulate emitted from source k in the 61 samplers. This mass is multiplied by $\hat{b}_{jk}$ hence resulting in the percentages reported in Table 5.

Table 5 is used to identify the nature of the sources. For instance Al_2O_3 and SiO_2 are emitted exclusively by source 1. Because Al_2O_3 and SiO_2 are known to have natural origin this source is associated to the soil dust pollution source. We proceed the same way for the other sources. We deduce possible identifications of the five pollution sources, see Table 6.

One can notice that the sources identified in Table 6 are the same than those found with PCA in Table 3. To verify the coherence of these sources identifications we have calculated, in Table 7, the correlations between the factors (the columns of $\breve{\mathbf{F}}_5$) and the sources obtained by receptor

DOI: 10.1002/env

Table 6. Receptor model sources identification

$k = 1$	Soil dust
$k = 2$	Vehicles
$k = 3$	Industry
$k = 4$	Combustion
$k = 5$	Sea

Table 7. Correlations between the sources of the receptor model and the factors of PCA after rotation

	Source 1	Source 2	Source 3	Source 4	Source 5
Factor 1	**0.98**	−0.18	−0.11	−0.02	−0.18
Factor 2	0.11	0.12	0.06	**0.95**	−0.30
Factor 3	−0.05	−0.09	**0.98**	0.02	−0.15
Factor 4	0.12	**0.96**	0.10	0.11	−0.22
Factor 5	−0.02	−0.13	−0.10	−0.27	**0.88**

Table 8. The sources profiles

	Soil dust	Vehicles	Industry	Combustion	Sea
Al_2O_3	41.6	0.0	0.0	0.0	0.0
$Si0_2$	18.5	0.0	0.0	0.0	0.0
P	6.2	0.0	0.7	0.0	0.6
SO_4	10.1	15.3	59.6	12.2	22.6
Cl	0.0	0.0	0.0	0.0	74.5
K	12.9	0.0	3.6	1.5	0.1
Ca	2.4	1.6	0.0	0.0	1.4
Mn	0.0	0.2	0.3	0.0	0.0
Fe_2O_3	6.7	15.0	12.7	0.2	0.0
Cu	0.3	0.7	0.4	0.0	0.0
Zn	0.7	0.0	16.3	0.0	0.3
Br	0.2	0.1	0.2	0.0	0.5
Pb	0.3	0.0	6.2	0.0	0.0
C_{org}	0.0	67.1	0.0	85.9	0.0

modeling (the columns of $\hat{\mathbf{G}}$). We observe that the factors match well with the receptor model sources.

4.2.3. Sources descriptions. The matrix $\hat{\mathbf{B}}$ of the sources profiles is reported in Table 8. We notice that, according to these profiles, C_{org}, which represents almost the total concentration in PM2.5, is only emitted by the vehicle and combustion sources.

4.2.4. Sources apportionments. From matrix $\hat{\mathbf{G}}$ of the source contributions we can deduce some interesting comments. First, we can focus on the relative contribution of each source in each particle sampler. For instance, Figure 3 represents the relative contributions of the combustion source in the 61 particle

DOI: 10.1002/env

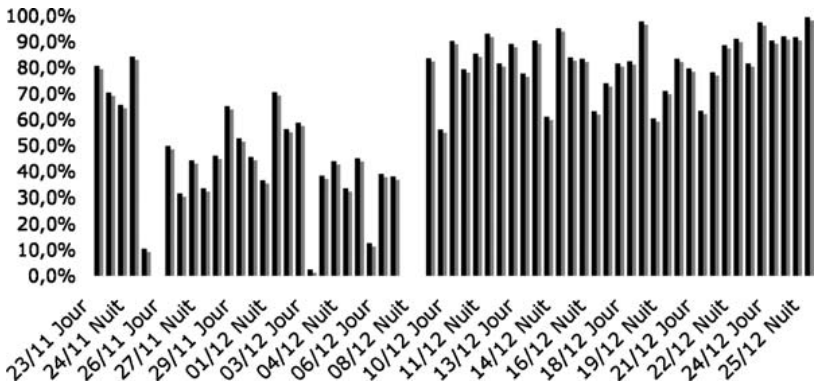

Figure 3. Relative contribution of the source combustion to the samples

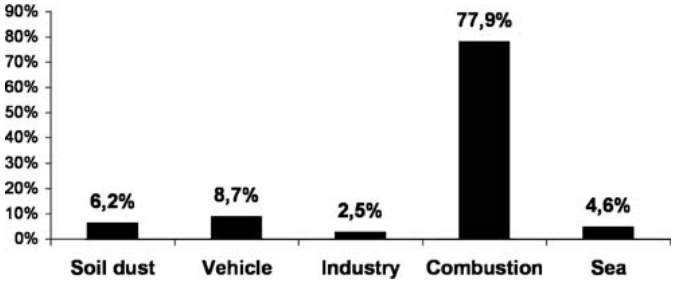

Figure 4. Global sources contributions to the PM2.5 dust contamination

samplers. We can notice the increase in the percentage of this source in the second period of sampling, corresponding to a decrease in the temperature (see Figure 1(c)).

We can also focus on the contribution of the sources to the PM2.5 dust contamination during the sampling period. Figure 4 shows the predominance of the combustion source during this winter sampling period.

5. CONCLUSION

In this case study, we propose a methodology for identifying and apportioning air pollution sources in a French urban site. The first step consists of factor analysis followed by a rotation technique and enables to identify the profiles of five principal sources: soil dust, vehicles, industry, combustion, and sea. Then a receptor modeling approach, based on PMF, is used to evaluate their contributions to the fine particles dust contamination. Thus we highlight, during winter, the predominance of combustion source over dust pollution. The interest of the approach lies in the fact that we do not use prior knowledge on the sources (number, nature, profiles), which means that this work can be applied to more complex sampling site. Finally this methodology is not specific to pollution and can be used for other sources detection problems.

DOI: 10.1002/env

REFERENCES

Hopke PK. 1991. *Receptor Modeling for Air Quality Management*. Elsevier, Amsterdam.
Jolliffe IT. 2002. *Principal Component Analysis.* Springer Verlag, New York.
Gabriel KR, Zamir S. 1979. Lower rank approximation of matrices by least squares with any choice of weights. *Technometrics* **21**: 489–498.
Paatero P, Tapper U. 1994. Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics* **5**: 111–126.
Polissar AV, Hopke PK, Malm WC, Sisler JF. 1998. Atmospheric aerosol over Alaska:2. Elemental composition and sources. *Journal of Geophysical Research* **103**: 19 045–19 057.

DOI: 10.1002/env

Table des figures

Bibliographie

Abdallah, H., Saporta, G. (1998). Classification d'un ensemble de variables qualitatives. *Revue de Statistique Appliquée*, **46**(4), 5-26.

Barrios, M.P., Velilla, S. (2007). A bootstrap method for assessing the dimension of a general regression problem. *Statistics & Probability Letters*, **77**, 247-255.

Benzécri, J. P. (1973). *L'analyse des données : T. 2, l'analyse des correspondances.* Paris : Dunod.

Breiman, L. (1996). Bagging Predictors. *Machine Learning*, **24**, 123-140.

Brillinger, D.R. (1983). A generalized linear model with "gaussian" regressor variables. *Festschrift for Erich L. Lehmann in Honor of His Sixty-Fifth Birthday (P. J. Bickel, K. A. Doksum, and J. L. Hodges, eds.)*, 97-114, Wadsworth, Belmont, Calif.

Bühlmann, P. (2004). Bagging, boosting and ensemble methods. *Handbook of Computational Statistics*, 877-907, Springer, Berlin.

Camden, M. (1989). The Data Bundle, Wellington : New Zealand Statistical Association.

Cook, R.D., Nachtsheim, C.J. (1994). Re-weighting to achieve elliptically contoured covariates in regression. *Journal of the American Statistical Association*, **89**, 592-599.

Cook, R.D., Weisberg, S. (1991). Comment on "Sliced Inverse Regression for Dimension Reduction", K.C. Li. *Journal of the American Statistical Association*, **86**, 328-332.

De Leeuw, J., Pruzansky, S. (1978). A new computational method to fit the weighted Euclidean distance model. *Psychometrika*, **43**, 479-490.

Derquenne, C. (1997). Classification de variables qualitatives. *XXIXe Journées ASU*, Carcassonne.

DeSarbo, W.S., Jedidi, K., Cool, K., Schendel, D. (1990). Simultaneous multidimensional unfolding and cluster analysis : An investigation of strategic groups. *Marketing Letters*, **2**, 129-146.

De Soete, G., Carroll, J.D., (1994). K-means clustering in a low-dimensional Euclidean space. *In : Diday, E., et al. (Eds.), New Approaches in Classification and Data Analysis. Springer, Heidelberg*, 212-219.

Dhillon, I.S, Marcotte, E.M., Roshan, U. (2003). Diametrical clustering for identifying anti-correlated gene clusters. *Bioinformatics*, **19**(13), 1612-1619.

Duan, N., Li, K.C. (1987). Distribution-free and link-free estimation method for the sample selection model. *Journal of Econometrics*, **35**(1), 25-35.

Eckart, C., Young, G., (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, **1**, 211-218.

Efron, B., Tibshirani, R. J. (1979). Bootstrap methods : another look at the jackknife. *The Annals of Statistics*, **7**(1), 1-26.

Efron, B., Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*, Chapman & Hall.

Escofier, B., Pagès, J. (2008). *Analyses factorielles simples et multiples*. Dunod.

Ferré, L. (1998). Determining the dimension in Sliced Inverse Regression and related methods. *Journal of the American Statistical Association*, **93**(441), 132-140.

Gifi, A. (1990). *Nonlinear Multivariate Analysis*, John Wiley & Sons.

Greenacre, M.J. (1984). *Theory and Applications of Correspondence Analysis*, London : Academic Press.

Greenacre, M.J., Blasius, J. (2006). *Multiple Correspondence Analysis and Related Methods*, Chapman & Hall/CRC Press, London.

Hall, P., Li, K.C. (1993). On almost linearity of low dimensional projections from high dimensional data. *The Annals of Statistics*, **21**, 867-889.

Hastie, T., Tibshirani, R., Eisen, M.B., Alizadeh, A., Levy, R., Staudt, L., Chan, W.C., Botstein, D., Brown, P. (2000). 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biology*, **1**(2), 1-21.

Hayashi, C. (1954). Multidimensional quantification–with applications to analysis of social phenomena. *Annals of the Institute of Statistical Mathematics*, **5**(2), 121-143.

Hubert, L., Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 193-208.

Kaiser, H.F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, **23**(3), 187-200.

Kiers, H.A.L. (1991). Simple structure in Component Analysis Techniques for mixtures of qualitative and quantitative variables. *Psychometrika*, **56**, 197-212.

Lebart, L., Morineau, A., Warwick, K. M. (1984). *Multivariate descriptive analysis : Correspondence analysis and related techniques for large matrics*, New York, Wiley-Interscience.

Lerman, I.C. (1990). Foundations of the likelihood linkage analysis (LLA) classification method. *Applied Stochastics Models and Data Analysis*, **7**(1), 63-76.

Lerman, I.C. (1993). Likelihood linkage analysis (LLA) classification method : An example treated by hand. *Biochimie*, **75**,(5) 379-397.

Li, K. C. (1991). Sliced inverse regression for dimension reduction, with discussion. *Journal of the American Statistical Association*, **86**, 316-342.

Li, L., Cook R.D., Nachtsheim, C.J. (2004). Cluster-based estimation for sufficient dimension reduction. *Computational Statistics & Data Analysis*, **47**, 175-193.

Liquet, B., Saracco, J. (2008). Application of the bootstrap approach to the choice of dimension and the α parameter in the SIR_α method. *Communications in Statistics - Simulation and Computation*, **37**(6), 1198-1218.

Mirkin, B., (2005), *Clustering for Data Mining. A Data Recovery Approach.*, Chapman & Hall, CRC Press, London, Boca Raton, FL.

Nishisato, S. (1980). *Analysis of categorical data : Dual Scaling and its applications*, Toronto : University of Toronto Press.

Nishisato, S. (1994). *Elements of Dual Scaling : An Introduction to Practical Data Analysis*, Hillsdale, NJ : Lawrence Erlbaum.

Saracco, J. (1997). An asymptotic theory for Sliced Inverse Regression. *Communications in Statistics - Theory and Methods*, **26**, 2141-2171.

Schott, J.R. (1994). Determining the dimensionality in Sliced Inverse Regression. *Journal of the American Statistical Association*, **89**, 141-148.

Shao, J., Tu, D.S. (1995). *The Jackknife and Bootstrap*, Springer Series in Statistics, Springer-Verlag, New York.

Smilde, A., Bro, R., Geladi, P. (2004). *Multi-way Analysis*, John Willey and Sons, England.

Soffritti, G. (1999). Hierarchical clustering of variables : a comparison among strategies of analysis. *Communications in Statistics - Simulation and Computation*, **28**(4), 977-999.

ten Berge, J.M.F. (1984). A joint treatment of VARIMAX rotation and the problem of diagonalizing symmetric matrices simultaneously in the least-squares sense. *Psychometrika*, **49**, 347-358.

Tenenhaus, M., Young, F. W. (1985). An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. *Psychometrika*, **50**, 91-119.

Vichi, M., Kiers, H.A.L. (2001). Factorial k-means analysis for two way data. *Computational Statistics and Data Analysis*, **37**, 49-64.

Vichi, M., Saporta, G. (2009). Clustering and disjoint principal component analysis. *Computational Statistics and Data Analysis*, **53**, 3194-3208.

Vigneau, E., Qannari, E.M. (2003). Clustering of Variables Around Latent Components. *Communications in Statistics - Simulation and Computation*, **32**(4), 1131-1150.

Résumé

Cette thèse est consacrée au problème de la réduction de dimension. Cette thématique centrale en Statistique vise à rechercher des sous-espaces de faibles dimensions tout en minimisant la perte d'information contenue dans les données.

Tout d'abord, nous nous intéressons à des méthodes de statistique multidimensionnelle dans le cas de variables qualitatives. Nous abordons la question de la rotation en Analyse des Correspondances Multiples (ACM). Nous définissons l'expression analytique de l'angle de rotation planaire optimal pour le critère de rotation choisi. Lorsque le nombre de composantes principales retenues est supérieur à deux, nous utilisons un algorithme de rotations planaires successives de paires de facteurs. Nous proposons également différents algorithmes de classification de variables qualitatives qui visent à optimiser un critère de partitionnement basé sur la notion de rapports de corrélation. Un jeu de données réelles illustre les intérêts pratiques de la rotation en ACM et permet de comparer empiriquement les différents algorithmes de classification de variables qualitatives proposés.

Puis nous considérons un modèle de régression semiparamétrique, plus précisément nous nous intéressons à la méthode de régression inverse par tranchage (SIR pour Sliced Inverse Regression). Nous développons une approche basée sur un partitionnement de l'espace des covariables, qui est utilisable lorsque la condition fondamentale de linéarité de la variable explicative est violée. Une seconde adaptation, utilisant le bootstrap, est proposée afin d'améliorer l'estimation de la base du sous-espace de réduction de dimension. Des résultats asymptotiques sont donnés et une étude sur des données simulées démontre la supériorité des approches proposées.

Enfin les différentes applications et collaborations interdisciplinaires réalisées durant la thèse sont décrites.

Mots-clés : Statistique multidimensionnelle, données qualitatives, rotation, classification de variables, régression semiparamétrique, méthode de régression inverse par tranchage, condition de linéarité, bootstrap.

Abstract

This thesis concentrates on dimension reduction approaches, that seek for lower dimensional subspaces minimizing the loss of statistical information.

First, we focus on multivariate analysis for categorical data. The rotation problem in Multiple Correspondence Analysis (MCA) is treated. We give the analytic expression of the optimal angle of planar rotation for the chosen criterion. If more than two principal components are to be retained, this planar solution is used in a practical algorithm applying successive pairwise planar rotations. Different algorithms for the clustering of categorical variables are also proposed to maximize a given partitioning criterion based on correlation ratios. A real data application highlights the benefits of using rotation in MCA and provides an empirical comparison of the proposed algorithms for categorical variable clustering.

We then examine the semiparametric regression method called SIR (Sliced Inverse Regression). We propose an extension based on the partitioning of the predictor space that can be used when the crucial linearity condition of the predictor is not verified. We also introduce bagging versions of SIR to improve the estimation of the basis of the dimension reduction subspace. Asymptotic properties of the estimators are obtained and a simulation study shows the good numerical behaviour of the proposed methods.

Finally applied multivariate data analysis on various areas is described.

Keywords : Multivariate analysis, categorical data, rotation, variable clustering, semiparametric regression, Sliced Inverse Regression, llinearity condition, bootstrap.