Une extension de l'analyse factorielle multiple pour des groupes de variables mixtes

Amaury Labenne [1,2] , Marie Chavent [2,3] , Vanessa Kuentz-Simonet [1] , Tina Rambonilaza [1] & Jérôme Saracco [2,3]

[1] *IRSTEA, UR ADBX, 33612 Cestas Cedex, France.*
{*amaury.labenne, vanessa.kuentz-simonet, tina.rambonilaza*}*@irstea.fr*
[2] *Univ. Bordeaux, IMB, UMR 5251, F-33400 Talence, France.*
{*marie.chavent, jerome.saracco*}*@math.u-bordeaux1.fr*
[3] *INRIA, CQFD, F-33400 Talence, France.*

Résumé. L'Analyse Factorielle Multiple (AFM) initialement proposée par Escofier et Pagès en 1982 est une méthode dédiée à l'étude d'un ensemble de n individus décrits par des groupes de variables quantitatives. Plus tard, cette méthode a été étendue pour prendre en compte des groupes de variables qualitatives (Pagès, 1983) puis simultanément des groupes quantitatifs et des groupes qualitatifs (Pagès, 2002). Cependant, cette méthode ne permet pas à l'heure actuelle de prendre en compte des groupes mixtes, c'est-à-dire contenant à la fois des variables quantitatives et qualitatives. Le but de notre étude étant de confectionner des indicateurs de développement durable en intégrant l'aspect de la qualité de vie, nous avons été confrontés à l'analyse de groupes de variables comportant des variables quantitatives et qualitatives. Dans ce travail, nous proposons une extension de l'AFM, appelée MFAMIX, pour l'analyse factorielle multiple de groupes de variables mixtes. Cette approche s'appuie sur une combinaison de l'AFM et de la méthode PCAMIX qui permet l'analyse de données mixtes. La méthode MFAMIX sera présentée à l'aide d'une décomposition en valeurs singulières et sera illustrée sur des données socio-économiques relatives à la qualité de vie.

Mots-clés. analyse factorielle multiple, méthode PCAMIX, analyse de la qualité de vie

Abstract. Multiple Factor Analysis (MFA) originally proposed by Escofier and Pagès in 1982 is a method dedicated to the study of a set of n individuals described by groups of quantitative variables. Later, this method was extended to take into account groups of qualitative variables (Pagès, 1983) then simultaneously quantitative groups and qualitative groups (Pagès, 2002). However, this method does not currently take into account mixed groups, that is to say containing both quantitative and qualitative variables. The aim of our study is to propose sustainable development indicators by integrating the aspect of quality of life. For that, we are confronted with the analysis of groups of variables with quantitative and qualitative variables. In this work, we propose an extension of the MFA method, called MFAMIX, for the multiple factor analysis of mixed groups of variables. This approach relies on a combination of AFM and PCAMIX method that allows

the analysis of mixed data. MFAMIX method will be presented using a singular value decomposition and will be illustrated on socio-economic data about the quality of life.

Keywords. multiple factor analysis, PCAMIX method, quality of life

1 Introduction

L'Analyse Factorielle Multiple (AFM) est une méthode de réduction de dimension qui permet de prendre en compte le fait que les individus sont décrits par des variables naturellement structurées en groupes ou thématiques. Initialement l'AFM a été mise en place pour l'analyse de variables quantitatives (Escofier et Pagès, 1983). Elle a ensuite été élargie à l'analyse de groupes de variables qualitatives (Escofier et Pagès, 1998) puis à l'étude d'un tableau de données que l'on qualifiera de "semi-mixte", où chaque groupe peut être soit de type quantitatif, soit de type qualitatif. Cette dernière extension de la méthode, proposée par Pagès (2002), permet la réduction de dimension dans un contexte où les bases de données deviennent de plus en plus composites.

A ce titre, nous sommes confrontés à une variété de données complexes dans les travaux relatifs à la construction d'indicateurs du développement durable, il faut entendre par là l'état de l'environnement, de l'économie, de la santé, des conditions sociales des individus comme des communautés. Pour cela, nous optons pour une approche en termes de qualité de vie car l'analyse et la mesure du bien être et de ses différentes composantes constituent un indicateur pertinent pour l'évaluation des états des sociétés. Face à la multitude de variables issues de thématiques différentes (environnement, social, économie, démographie, etc.) disponibles pour décrire la qualité de vie, les méthodes multi-tableaux telles l'AFM sont une réponse pertinente pour l'analyse de ces données structurées en groupes. Dans cette problématique, les variables au sein d'une même thématique ne sont pas homogènes, mais mixtes dans le sens où elles peuvent être quantitatives ou qualitatives. L'écriture actuelle de l'AFM et son implémentation dans le package R FactoMineR ne permettant pas d'intégrer des thématiques mixtes dans l'analyse, nous proposons une extension de l'AFM qui permet l'analyse de groupes mixtes via l'utilisation de la méthode PCAMIX (Chavent et al. 2012).

2 Rappels sur la méthode PCAMIX

La méthode PCAMIX (Chavent et al. 2012) est une méthode d'analyse factorielle pour des données mixtes, c'est à dire un mélange de variables quantitatives et qualitatives. Elle est similaire à la méthode d'analyse factorielle de données mixtes d'Escofier (1979) et Pagès (2004) dans la manière dont sont recodées les variables qualitatives et quantitatives. Cependant, Chavent et al. (2012) propose une formulation de PCAMIX à l'aide d'une décomposition en valeurs singulières (SVD) des données préalablement transformées. Cela permet d'obtenir directement les composantes principales, les "loadings" des variables

quantitatives (corrélation avec les composantes principales), les rapports de corrélation entre les variables qualitatives et les composantes, ainsi que les coordonnées principales des modalités des variables qualitatives.

On note n le nombre d'individus, p_1 le nombre de variables quantitatives, p_2 le nombre de variables qualitatives et $p = p_1 + p_2$ le nombre total de variables. On note $\mathbf{Z}_1$ la matrice des variables quantitatives que l'on considère centrée-réduite. On note m le nombre total de modalités des p_2 variables qualitatives et n_s le nombre d'individus possédant la modalité s. Soit $\mathbf{G}$ le tableau disjonctif complet associé aux variables qualitatives et soit $\mathbf{D}$ la matrice diagonale contenant les fréquences des m modalités : $\mathbf{D} = \text{diag}(n_s/n), s = 1 \dots m$.

On note $\mathbf{J} = \mathbf{I_n} - \mathbf{11'/n}$ l'opérateur de centrage, où $\mathbf{1}$ est le vecteur d'ordre n comprenant uniquement la valeur 1. La matrice des données qualitatives est recodée de la façon suivante : $\mathbf{Z}_2 = \mathbf{JGD}^{-1/2}$. On juxtapose les deux matrices et on note $\mathbf{Z} = \frac{1}{\sqrt{n}}(\mathbf{Z}_1|\mathbf{Z}_2)$.

On réalise alors la SVD de $\mathbf{Z}$:

$$\mathbf{Z} = \mathbf{U}\Lambda^{1/2}\mathbf{V}', \tag{1}$$

Où $\mathbf{U'U} = \mathbf{V'V} = \mathbf{I}_r$ avec r le rang de la matrice $\mathbf{Z}$ et Λ est la matrice diagonale des valeurs propres rangées par ordre décroissant.

La matrice des k premières composantes principales standardisées est définie par : $\mathbf{X} = \sqrt{n}\mathbf{U}_k$ où $\mathbf{U}_k$ est formée par les k premières colonnes de U.

On calcule la matrice $\mathbf{A} = \mathbf{V}_k\Lambda_k^{1/2}$ et on l'écrit comme la concaténation d'une matrice $\mathbf{A}_1$ de dimension $p_1 \times k$ et d'une matrice $\mathbf{A}_2$ de dimension $m \times k$: $\mathbf{A} = \begin{pmatrix}\mathbf{A_1}\\ \mathbf{A_2}\end{pmatrix}$. La matrice $\mathbf{A}_1$ contient les corrélations entre les variables quantitatives et les composantes principales, alors que la matrice $\mathbf{D}^{-1/2}\mathbf{A}_2$ contient les coordonnées principales des modalités des variables qualitatives. On peut également calculer la matrice $\mathbf{C}$ des "squared loadings" des p variables sur les k composantes principales. On la déduit de $\mathbf{A}$ de la façon suivante :

$$\begin{cases} c_{jl} = a_{jl}^2 & \text{si la variable } j \text{ est quantitative,} \\ c_{jl} = \sum_{s \in I_j} a_{sl}^2 & \text{si la variable } j \text{ est qualitative.} \end{cases}$$

Si la variable j est quantitative, il s'agit de la corrélation au carré avec la composante principale. Si la variable j est qualitative, le "squared loading" correspond au rapport de corrélation entre la variable et la composante principale.

3 La méthode MFAMIX

Cette méthode se distingue d'une ACP ou d'une ACM globale appliquée à l'ensemble des données dans la mesure où elle permet de prendre en compte la structure en groupes de l'ensemble des variables. Pour cela, l'AFM applique une pondération particulière aux variables selon leur appartenance aux différentes thématiques. Ainsi l'influence des groupes est équilibrée dans la construction des composantes principales globales. Au contraire une

méthode factorielle classique, qui ne considère pas la structure en groupes des variables, accorderait plus d'importance à un groupe avec une structure forte ou de grande dimension. L'obtention des composantes principales serait alors influencée de manière prépondérante par ce thème et celles-ci ne résumeraient pas de manière objective l'information apportée par l'ensemble des données. De plus, l'AFM permet de situer les différentes thématiques dans un même référentiel, en vue de leur comparaison, ce qui n'est pas permis par des analyses factorielles classiques qui seraient réalisées de manière indépendante sur chaque groupe.

Le principe général de l'AFM mixte repose essentiellement sur deux étapes. Tout d'abord, on analyse chaque thématique prise séparément avec la méthode PCAMIX. On obtient ainsi la plus grande valeur propre correspondant à chaque sous-tableau. Puis, on applique PCAMIX sur l'ensemble de toutes les variables prises en commun où chaque variable est pondérée par l'inverse de la première valeur propre de la thématique dont elle est issue. Ainsi l'influence de chaque groupe est équilibrée dans la construction des composantes principales globales. Une écriture sous forme de décomposition en valeurs singulières est proposée pour MFAMIX. Les codes R sont disponibles auprès des auteurs.

L'application de la méthode sera illustrée sur des données socio-économiques relatives à la qualité de vie d'un ensemble de communes.

Bibliographie

[1] Chavent M, Kuentz-Simonet V, Saracco J, (2012), Orthogonal rotation in PCAMIX, *Advances in Data Analysis and Classification*, 6 : 131-146.
[2] Escofier B (1979), Traitement simultané de variables qualitatives et quantitatives en analyse factorielle, *Cahiers de l'Analyse des données*, 4(2) : 137-146.
[3] Escofier B et Pagès J (1983), Méthode pour l'analyse de plusieurs groupes de variables. Application à la caractérisation des vins rouges du Val de Loire, *Revue de statistique appliquée*, 31(2) : 43-59.
[4] Escofier B et Pagès J (1998), *Analyses factorielles simples et multiples*, Dunod, 3e ed.
[5] Husson F, Josse J, Lê S et Mazet J (2012). FactoMineR : Multivariate Exploratory Data Analysis and Data Mining with R, *R package version 1.20. http ://CRAN.R-project.org/package=FactoMineR.*
[6] Pagès J (2002), Analyse factorielle multiple appliquée aux variables qualitatives et aux données mixtes, *Revue de statistique appliquée*, 50(4) : 5-37.
[7] Pagès J (2004), Analyse factorielle de données mixtes, *Revue de statistique appliquée*, 52(4) : 93-111.