Sélection de variables pour la construction d'indicateurs de qualité de vie pour des données mixtes structurées en groupes

A. Labenne[a], **M. Chavent**[b], **V. Kuentz-Simonet**[a], **T. Rambonilaza**[a] and **J. Saracco**[b]

[a]IRSTEA, UR ADBX, 33612 Cestas Cedex, France
amaury.labenne@irstea.fr

[b]Univ. Bordeaux, IMB & INRIA Bordeaux Sud-Ouest, CQFD, F-33400 Talence

Mots clefs : analyse factorielle multiple, données mixtes, stabilité, closest submodel selection

1 Introduction

L'analyse et la mesure de la qualité de vie peuvent se faire via deux approches différentes et complémentaires. La première est tournée vers l'analyse des niveaux de satisfaction de la vie à l'aide d'enquêtes auprès des individus. La seconde, à laquelle nous nous intéressons ici, vise à analyser les conditions de vie des personnes et s'appuie sur des données nationales. L'enjeu de cette approche consiste à créer des indices composites des conditions de vie. Dans cet objectif, les méthodes de réduction de dimension sont particulièrement adaptées car elles permettent de construire de nouvelles variables qui résument "au mieux" l'information contenue dans les variables initiales. Selon la littérature, les composantes de la qualité de vie sont reliées à différents thèmes (groupes de variables) dont les plus souvent cités sont "Les conditions familiales", "Les conditions économiques", "Les conditions de logement", "L'accès aux services", "L'environnement". L'Analyse Factorielle Multiple (AFM), initialement développée par Escofier et Pagès [1], est une méthode d'analyse factorielle conçue pour traiter les données structurées en groupes de variables quantitatives. L'idée principale de la méthode est de diviser chaque variable d'un groupe par la première valeur propre issue de l'Analyse en Composantes Principales (ACP) de ce groupe, puis d'effectuer une ACP sur l'ensemble des variables ainsi pondérées. Nous allons utiliser la méthode d'analyse factorielle multiple mixte (MFAmix) que nous avons développée [2] afin d'analyser ces données structurées en groupes de variables quantitatives et/ou qualitatives. Ainsi les composantes principales issues de MFAmix (en tant que combinaisons linéaires des variables d'origine) constitueront nos indices composites de mesure de qualité de vie. Cependant, la création de ces indices composites soulève plusieurs questions. Combien de composantes principales faut-il retenir pour la création d'indices composites ? Peut-on obtenir des indices composites semblables en ne sélectionnant qu'un nombre restreint de variables ou de groupes de variables afin de faciliter l'interprétation ?

2 Choix du nombre de composantes principales

Nous allons utiliser ici un critère de stabilité dans le but de choisir un nombre de dimensions pertinent. Il existe différents critères de choix de dimension en analyse factorielle [3, 4]. Le critère que nous utilisons ici est celui initialement développé par Besse [4] dans le cadre de l'ACP. Il sera réutilisé ici dans le cadre de MFAmix mais peut être appliqué de la même manière sur toutes les méthodes d'analyse factorielle basée sur des Décompositions en Valeurs Singulières

Généralisées (DVSG). Ce critère de stabilité utilise la notion de distance entre projecteurs et est construit comme une fonction de risque estimée par bootstrap des individus.
La méthode MFAmix [2] est basée sur la DVSG de $\mathbf{Z}$ (la matrice $(n \times p)$ des données brutes précédemment recodées) avec les métriques $\mathbf{D}$ pour les individus et $\mathbf{M}$ pour les variables. On a ainsi :

$$\mathbf{Z} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^t.$$

On définit ensuite la matrice de projection suivante $\widehat{\mathbf{P}_q} = \mathbf{V}_q\mathbf{V}_q^t\mathbf{M}$ pour $1 < q < r$, qui est la matrice de projection $\mathbf{M}$-orthogonale des lignes de $\mathbf{Z}$ sur $E_q = Im(\mathbf{V}_q)$, le sous espace engendré par les q premières colonnes de $\mathbf{V}$. La fonction de perte reposant sur la distance euclidienne entre deux projecteurs orthogonaux est donnée par :

$$\mathcal{L}_q = \mathcal{Q}(E_q, \widehat{E_q}) = \frac{1}{2}\|\mathbf{P}_q - \widehat{\mathbf{P}_q}\|_2^2 = q - Tr(\mathbf{P}_q\widehat{\mathbf{P}_q}).$$

Finalement, le risque est défini comme l'espérance de la fonction de perte : $R_q = E[\mathcal{L}_q]$. L'idée est d'estimer R_q par $\widehat{R_q}$ au moyen du bootstrap des individus et de déterminer le nombre de composantes par $\hat{q} = \underset{q}{\operatorname{argmin}}\ \widehat{R_q}$.

3 Sélection de variables

Afin de faciliter l'interprétation des composantes principales de MFAmix en tant que combinaisons linéaires des variables d'origine, nous allons chercher des sous-espaces de $\mathbb{R}^n$ engendrés par les q premières composantes principales de MFAmix obtenues avec un nombre réduit p_0 de variables. Nous allons nous inspirer de la méthode CSS (Closest Submodel Selection) utilisée initialement pour la régression SIR, voir Coudret et al [5]. Les meilleurs sous-espaces doivent être le plus proche du sous espace de référence (composantes principales de MFAmix sur les p variables de $\mathbf{Z}$). Pour cela, il est nécessaire de définir une mesure de proximité entre les sous espaces engendrés par les q premières composantes de MFAmix réalisée sur un sous ensemble a de variables et celles de MFAmix réalisée sur l'ensemble des p variables. Les variables qui apparaissent le plus souvent dans la construction des meilleurs sous-espaces sont naturellement considérées comme calculées sur les variables les plus importantes. Ainsi, nos indices composites parcimonieux seront les composantes principales construites sur les variables sélectionnées. Des exemples de résultats seront illustrés au travers des différentes fonctions du package `PCAmixdata`.

Références

[1] 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.
[2] Chavent, M., Kuentz-Simonet, V., Labenne, A.,Saracco, J. (2013).Une extension de l'analyse factoriellemultiple pour des groupes de variables mixtes : MFAmix. *45èmes Journées de la Statistique, Toulouse.*
[3] Besse, P. (1992). PCA stability and choice of dimensionality. *Statistics and Probability Letters*, **13**, 405-410.
[4] Josse, J., Husson, F. (2012). Selecting the number of components in PCA using cross-validation approximations. *Computational Statistics and Data Analysis*, **56**, pp. 1869-1879.
[5] Coudret, R., Liquet, B., Saracco, J. (2014). Comparison of sliced inverse regression method approaches for undeterminated cases.*Journal de la Société française de Statistique*, in press.