

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

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.*

² *Univ. Bordeaux, IMB, UMR 5251, F-33400 Talence, France.*

³ *INRIA, CQFD, F-33400 Talence, France.*



INTRODUCTION

PCAmix : L'analyse factorielle de données mixtes

- Escofier B, (1979), Traitement simultané de variables qualitatives et quantitatives en analyse factorielle
- Kiers, (1991), Simple structure in component analysis techniques for mixtures of qualitative and quantitative variables
- Pages, (2004), Analyse factorielle de données mixtes
- Husson et al. FactoMineR (fonction AFDM)
- Chavent et al., (2012), Orthogonal rotation in PCAMIX

⇒ Approche : SVD avec métrique

MFA : L'analyse factorielle de données multiples

- Escofier et Pages (1983), Méthode pour l'analyse de plusieurs groupes de variables
- Pages J (2002), Analyse factorielle multiple appliquée aux variables qualitatives et aux données mixtes
- FactoMineR : fonction MFA

⇒ Approche AFM avec mixité à l'intérieur des groupes

PLAN

- ① La méthode PCAmix avec métriques
- ② La méthode MFAmix
- ③ Application sur des données relatives à la qualité de vie

La méthode PCAmix (1/3)

Données

- n individus
- $\begin{cases} p_1 \text{ variables quantitatives stockées dans la matrice } X_1 \\ p_2 \text{ variables qualitatives (total de } m \text{ modal) stockées dans la matrice } X_2 \end{cases}$
- On recode les données de la manière suivante dans $Z = (Z_1|Z_2)$
 - ▶ $Z_1 = X_1$ **centrée-réduite**. $\dim(Z_1) = n \times p_1$
 - ▶ $Z_2 = \text{TDC de } X_2$ **centré**. $\dim(Z_2) = n \times m$

Métriques

- Métrique D des individus : $D = \text{diag}(d_i, i = 1 \dots n)$, en général $d_i = \frac{1}{n}$.
- Métrique M des variables : $M = \begin{pmatrix} M_1 & \\ & M_2 \end{pmatrix}$
 - ▶ $M_1 = \text{diag}(\omega_j, j = 1, \dots, p_1)$ pour les variables quantitatives.
 - ▶ $M_2 = \text{diag}(\omega_s, s = 1, \dots, m)$ pour les indicatrices des variables qualitatives.

Dans PCAmix, $M_1 = I_{p_1}$ et $M_2 = \text{diag}\left(\frac{n}{n_s}, s = 1, \dots, m\right)$

La méthode PCAmix (2/3)

SDV de Z avec les métriques D et M :

$$Z = U\Lambda V^t$$

- $\Lambda = \text{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_r})$. Où :
 - ▶ r est le rang de Z
 - ▶ $\lambda_1, \dots, \lambda_r$ sont les valeurs propres de $Z^t D Z M$ et $Z M Z^t D$.
- U est la matrice de dimension $n \times r$ dont les colonnes sont les vecteurs propres de $Z M Z^t D$ et $U^t D U = I_r$
- V est la matrice de dimension $p \times r$ dont les colonnes sont les vecteurs propres de $Z^t D Z M$ et $V^t M V = I_r$

On sépare la matrice V en deux parties : $V = \left(\begin{array}{c} V_1 \\ V_2 \end{array} \right) \quad \left. \begin{array}{c} \} p_1 \\ \} m \end{array} \right\}$

La méthode PCAmix (3/3)

Coordonnées factorielles des individus	Coordonnées factorielles des variables	
	Variables quantitatives	Modalités des variables qualitatives
$F = U \Lambda$ $\dim(F) = n \times r$	$A_1 = V_1 \Lambda$ $\dim(A_1) = p_1 \times r$	$A_2 = M_2 V_2 \Lambda$ $\dim(A_2) = m \times r$

Remarques :

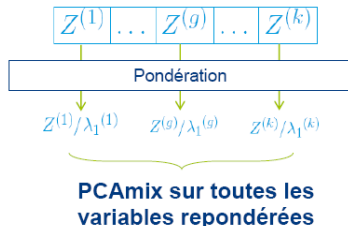
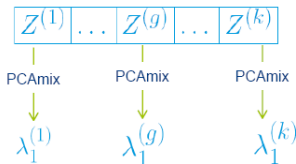
- Si $(p_2 = 0)$, PCAmix = ACP classique.
- Si $(p_1 = 0)$, on obtient les mêmes résultats qu'une ACM pour les coordonnées des modalités. Cependant, on obtient des valeurs propres p_2 fois plus grandes et donc des coordonnées factorielles multipliées par $\sqrt{p_2}$.
- Ainsi la propriété quasi-barycentrique de l'ACM reliant les coordonnées factorielles des modalités aux coordonnées factorielles des individus reste vraie.

La méthode MFAmix (1/4)

Les données :

- Les données sont séparées en k groupes de variables $X^{(1)} \dots X^{(k)}$. **Les variables à l'intérieur d'un groupe peuvent être quantitatives ou qualitatives.**
- On recode chaque $X^{(g)}$, ($g = 1, \dots, k$) en $Z^{(g)}$ comme dans PCAmix.

Principe de la méthode :



La méthode MFAmix (2/4)

Métrique M^* des variables : $M^* = \begin{pmatrix} M_1^* \\ M_2^* \end{pmatrix}$

- $M_1^* = M_1 P_1 = M_1 \times \begin{pmatrix} \ddots & & \\ & \alpha_j & \\ & & \ddots \end{pmatrix}$ avec $\alpha_j = \frac{1}{\lambda_1(g)}$ où g est le groupe auquel appartient la variable j .
- $M_2^* = M_2 P_2 = M_2 \times \begin{pmatrix} \ddots & & \\ & \alpha_s & \\ & & \ddots \end{pmatrix}$ avec $\alpha_s = \frac{1}{\lambda_1(g)}$ où g est le groupe auquel appartient la modalité s .

On effectue la SVD de Z avec les métriques D et M^* : $Z = U^* \Lambda^* V^{*t}$

	Coordonnées factorielles des individus	Coordonnées factorielles des variables	
		Variables quantitatives	Modalités des variables qualitatives
PCAmix	$F = U \Lambda$	$A_1 = V_1 \Lambda$	$A_2 = M_2 V_2 \Lambda$
MFAmix	$F^* = U^* \Lambda^*$	$A_1^* = V_1^* \Lambda^*$	$A_2^* = M_2 V_2^* \Lambda^*$

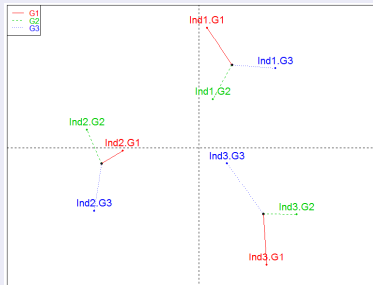
La méthode MFAmix (3/4)

Résultats spécifiques à l'AFM

Scores des individus partiels : $F^{(k)}$

- On note $\tilde{Z}^{(g)} = \begin{bmatrix} 0 & Z^{(g)} & 0 \end{bmatrix}$ le tableau de données partiel de dimension $n \times p$ correspondant au $g^{\text{ème}}$ groupe complété par des 0 pour les variables des autres groupes.
- Les individus partiels sont les individus décrits séparément par chaque groupe de variables. Les coordonnées des projections des individus partiels sur les axes de l'AFM sont :

$$F^{(g)} = k \tilde{Z}^{(g)} M^* V^*$$



La méthode MFAmix (4/4)

Coordonnées d'un groupe g sur l'axe α

La coordonnée d'un groupe est égale à l'inertie des projections des variables du groupe sur l'axe :

$$\sum_{j \in \text{Group}(g)} \omega_j^* a_{j\alpha}^{*2}$$

Ainsi, si la coordonnée d'un groupe est proche de 1 sur un axe, ce groupe détermine de façon prépondérante la construction de cet axe.

Mesure globale de liaison entre deux groupes g et g' : coefficient RV

Le coefficient RV permet de rendre compte de la proximité des structures de deux groupes de données :

$$RV = \frac{\text{trace}\{Z^{(g)t} Z^{(g')} Z^{(g')t} Z^{(g)}\}}{\sqrt{\text{trace}\{Z^{(g)t} Z^{(g)} Z^{(g)t} Z^{(g)}\} \text{trace}\{Z^{(g')t} Z^{(g')} Z^{(g')t} Z^{(g')}\}}}$$

Si ce coefficient est proche de 1, alors on conclut que les deux groupes possèdent très vraisemblablement de l'information commune.

Application sur des données relatives à la Qualité de vie

Problématique :

- Projet ANR ADAPTEAU : 
- Etablir un diagnostic socio-économique et environnemental des territoires par la notion de Qualité de vie
- Quels sont les thèmes prépondérants dans la notion de Qualité de vie ?

Les Données :

- 302 communes du Sud-Ouest
- 56 variables ($p_1 = 56$, $p_2 = 23$, $m = 50$) (INSEE, Corine Land Cover)
- Variables séparées en 8 thématiques relatives à la Qualité de vie :
 - ▶ Logement et Services,
 - ▶ Population,
 - ▶ Emplois,
 - ▶ Condition Familiales,
 - ▶ Education,
 - ▶ Revenus,
 - ▶ Environnement,
 - ▶ Sécurité et Soins

Application sur des données relatives à la Qualité de vie

Utilisation de MFAmix dans :

```
1 res<-MFAmix( data=X, group=y, name.group=nom, ndim=3)
2 summary( res )
```

Call:
MFAmix(DATA = DATA, group = group, name.group = name.group, ndim = 3, graph = F)

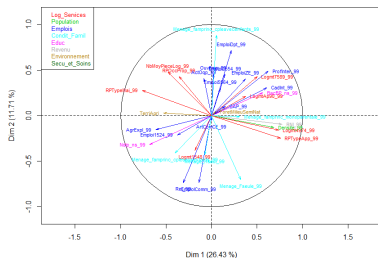
Data:
number of observations: 302
number of variables: 56
number of numerical variables: 33
number of categorical variables: 23
total number of categories: 50

Summary by groups :

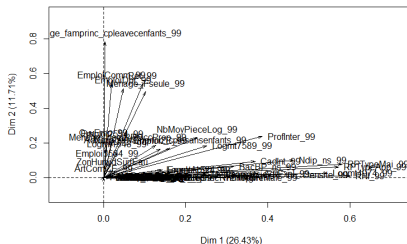
	Log_Services	Population	Emplois	Condit_Famil	Educ	Revenu	Environnement	Secu_et_Soins
Number of numerical variables	8	1	14	5	2	1	2	0
Number of categorical variables	18	0	1	0	0	0	2	2
Total number of categories	39	0	2	0	0	0	5	4

Application sur des données relatives à la Qualité de vie

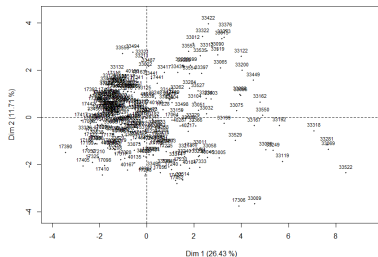
Correlation circle



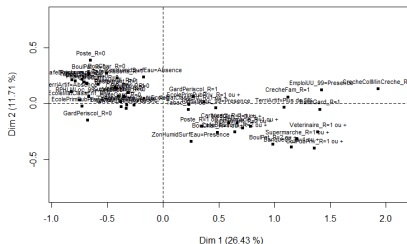
Squared Loadings



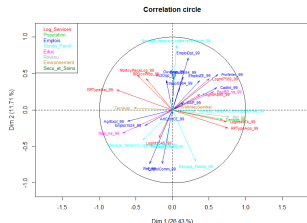
Individual factor map



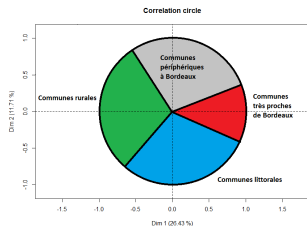
Categories factor map



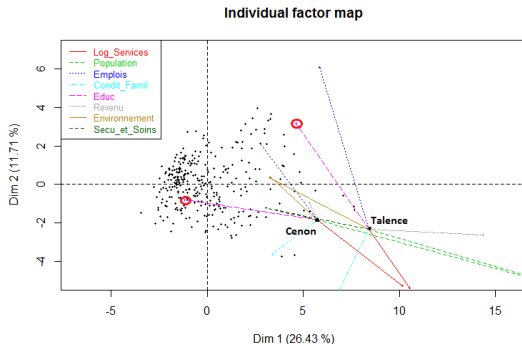
Application sur des données relatives à la Qualité de vie



- **Les communes très proches de Bordeaux** sont caractérisées par : forte densité, logements de type appartement anciens, des forts revenus, une population qualifiée et un accès à un grand nombre de services
- **Les communes périphériques à Bordeaux**, sont caractérisées par : forte proportion de ménages avec enfants, des logements récents, des emplois qualifiés hors de la commune mais à l'intérieur du département
- **Les communes rurales** sont caractérisées par une faible densité, un faible revenu, des habitations principales de type maison, une forte proportion de territoires et d'exploitants agricoles et une population dans l'ensemble peu diplômée
- **Les communes littorales** sont caractérisées par : une forte proportion de retraités, des ménages composés d'hommes ou de femmes seules, et des emplois à l'échelle de la commune, souvent des emplois dans le domaine des services



Application sur des données relatives à la Qualité de vie



- Ces deux individus sont proches sur le plan factoriel
- Cependant, leurs coordonnées partielles sur les différents groupes ne sont pas aussi proche
- Sur le groupe Education, forte coordonnée pour la ville de Talence et faible coordonnée pour la ville de Cenon

CONCLUSION

- La méthode MFAmix a permis de faire de l'AFM avec des variables mixtes à l'intérieur des groupes, ce qui était nécessaire afin d'analyser nos données
- La méthode a permis de mettre en évidence les thématiques importantes de la Qualité de vie et de caractériser les communes en fonction des différentes thématiques
- Poursuivre les applications sur un autre domaine d'étude plus étendu
- Code  et sorties graphiques inspirées de FactoMineR
- Méthode PCAmix et MFAmix intégrée dans le package PCAmixdata (disponible auprès des auteurs)

Bibliographie

- Chavent M, Kuentz-Simonet V, Saracco J, (2012), Orthogonal rotation in PCAMIX, *Advances in Data Analysis and Classification*, 6 : 131-146.
- Escofier B (1979), Traitement simultané de variables qualitatives et quantitatives en analyse factorielle, *Cahiers de l'Analyse des données*, 4(2) : 137-146.
- 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.
- Escofier B et Pagès J (1998), *Analyses factorielles simples et multiples*, Dunod, 3e ed.
- 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>.
- Pagès J (2002), Analyse factorielle multiple appliquée aux variables qualitatives et aux données mixtes, *Revue de statistique appliquée*, 50(4) : 5-37.
- Pagès J (2004), Analyse factorielle de données mixtes, *Revue de statistique appliquée*, 52(4) : 93-111.

MERCI DE VOTRE ATTENTION...