

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
- Pagès, (2004), Analyse factorielle de données mixtes
- Husson et al. Package  : FactoMineR (fonction AFDM)
- Chavent et al., (2012), Orthogonal rotation in PCAMIX

⇒ Approche : SVD avec métrique

MFA : L'analyse factorielle de données multiples

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

⇒ Approche AFM avec mixité à l'intérieur des groupes

PLAN

- ① La méthode PCAmix avec métriques
- ② La méthode MFAmix
- ③ Utilisation de MFAmix dans 

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 : $M = \begin{pmatrix} M_1 & 0 \\ 0 & M_2 \end{pmatrix}$
 - ▶ $M_1 = \text{diag} (\omega_j, j = 1, \dots, p_1)$ 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 $M_2 = \text{diag} (\frac{n}{n_s}, s = 1, \dots, m)$

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 $(p_1 + m) \times r$ 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 : $V = \left(\begin{array}{c} V_1 \\ V_2 \end{array} \right) \quad \left. \begin{array}{l} \} p_1 \\ \} m \end{array} \right\}$

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 :

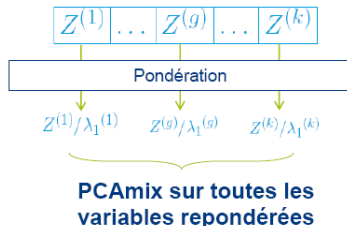
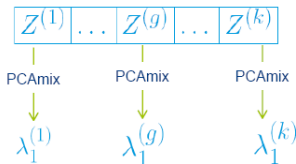
- Si ($p_2 = 0$) , PCAmix = ACP classique.
- Si ($p_1 = 0$) , on obtient les mêmes résultats qu'une ACM pour les coordonnées des modalités. Cependant, on obtient des valeurs propres p_2 fois plus grandes et donc des coordonnées factorielles multipliées par $\sqrt{p_2}$.
- Ainsi la propriété quasi-barycentrique 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 et/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
<u>PCAmix</u>	$F = U \Lambda$	$A_1 = V_1 \Lambda$	$A_2 = M_2 V_2 \Lambda$
<u>MFAmix</u>	$F^* = U^* \Lambda^*$	$A_1^* = V_1^* \Lambda^*$	$A_2^* = M_2 V_2^* \Lambda^*$

Utilisation de MFAmix dans :

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
- 35 variables ($p_1 = 26$, $p_2 = 9$, $m = 18$) (INSEE, Corine Land Cover)
- Variables séparées en 4 thématiques relatives à la Qualité de vie :
 - ▶ Logement et Services,
 - ▶ Emplois,
 - ▶ Condition Familiales,
 - ▶ Sécurité et Soins

Utilisation de MFAmix dans :

Appel de la méthode :

```
1 library(PCAmixdata)
2 data(INSEE)
3 class.var<-c(rep(1,15),rep(2,13),rep(3,5),rep(4,2))
4 nom.group<-c("Logmt_Services","Emplois","Condit_Famil","Secu_et_Soins")
5 res<-MFAmix(data=INSEE,group=class.var,name.group=nom.group,ndim=3,graph=F)
```

	name	description
1	"\$eig"	"eigenvalues"
2	"\$separate.analyses"	"separate analyses for each group of variables"
3	"\$group"	"results for all the groups"
4	"\$partial.axes"	"results for the partial axes"
5	"\$ind"	"results for the individuals"
6	"\$quanti.var"	"results for the quantitatives variables"
7	"\$quali.var"	"results for the categorials variables"
8	"\$global.pca"	"results for the global PCA"

Utilisation de MFAmix dans :

Summary de MFAmix :

```
1 summary(res)
```

Call:

```
MFAmix(data = INSEE, group = class.var, name.group = nom.group,      ndim = 3, graph = F)
```

Data:

```
number of observations: 302
number of variables: 35
  number of numerical variables: 26
  number of categorical variables: 9
    total number of categories: 18
```

Summary by groups :

	Logmt_Services	Emplois	Condit_Famil	Secu_et_Soins
Number of numerical variables	8	13	5	0
Number of categorical variables	7	0	0	2
Total number of categories	14	0	0	4

Utilisation de MFAmix dans :

Variables quantitatives :

```
1 summary(res$quanti.var)
```

	Length	Class	Mode
coord	78	-none-	numeric
contrib	78	-none-	numeric
cos2	78	-none-	numeric

```
1 res$quanti.var$coord[1:5,1:2]
```

	dim1	dim2
NbMoyPieceLog_99	-0.364	0.530
LogmtAp90_99	0.363	0.136
Logmt1548_99	-0.237	-0.346
Logmt4974_99	0.637	-0.297
Logmt7589_99	0.591	0.311

Variables qualitatives :

```
1 summary(res$quali.var)
```

	Length	Class	Mode
coord	54	-none-	numeric
contrib	54	-none-	numeric
cos2	54	-none-	numeric
contrib.quali	27	-none-	numeric

```
1 res$quali.var$contrib.quali  
[1:5,1:2]
```

	dim1	dim2
RPHMLoc_99	3.100	0.263
BanqueCE_R	3.451	1.367
CollPubPriv_R	2.925	1.010
CrecheFam_R	2.674	0.059
EcoleMatClassEnf_R	2.292	0.063

Utilisation de MFAmix dans :

Individus :

```
1 summary(res$ind)
```

	Length	Class	Mode
coord	906	-none-	numeric
contrib	906	-none-	numeric
cos2	906	-none-	numeric
coord.partiel	3	data.frame	list

```
1 res$ind$coord.partiel[1:5,1:2]
```

	Dim.1	Dim.2
17015.Logmt_Services	-1.929	0.947
17015.Emplois	-0.008	0.088
17015.Condit_Famil	0.877	1.927
17015.Secu_et_Soins	-2.199	1.029
17021.Logmt_Services	0.411	0.418

Groupes de variables :

```
1 summary(res$group)
```

	Length	Class	Mode
Lg	16	-none-	numeric
RV	16	-none-	numeric
coord	3	data.frame	list
contrib	3	data.frame	list
dist2	4	-none-	numeric
cos2	12	-none-	numeric

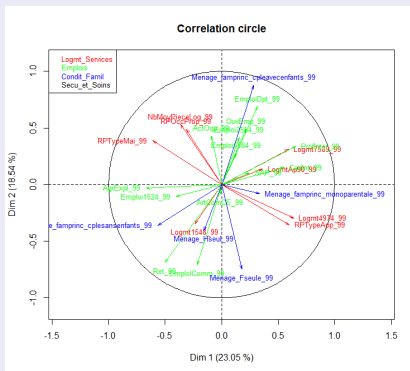
```
1 res$group$RV
```

	Log_Serv	Emplois	Cond_Fam	SecuSoin
Log_Serv	1.00	0.24	0.20	0.44
Emplois	0.24	1.00	0.35	0.12
Cond_Fam	0.20	0.35	1.00	0.06
SecuSoin	0.44	0.12	0.06	1.00

Utilisation de MFAmix dans :

Cercle des corrélations des variables quantitatives :

```
plot(res, choix="var", habillage="group", axes=c(1,2), col.hab=c("red", "green", "blue", "black"), cex=0.8)
```

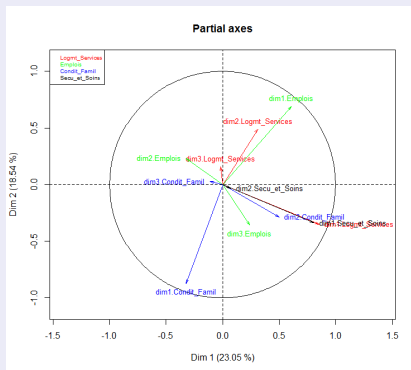


Forte corrélation entre l'axe 2 et la variable CoupleAvecEnfants

Utilisation de MFAmix dans :

Axes partiels des analyses séparées :

```
plot(res, choix="axes", habillage="group", axes=c(1,2), col.hab=c("red", "green", "blue", "black"), cex=0.8)
```

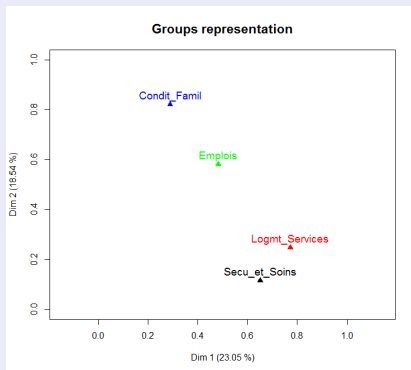


Forte corrélation entre le premier axe de l'analyse séparée du groupe "Conditions familiales" et le 2ème axe de l'AFM.

Utilisation de MFAmix dans :

Représentation des groupes :

```
plot(res, choix="group", axes=c(1,2), col.hab=c("red", "green", "blue", "black"),  
      cex=1.2)
```

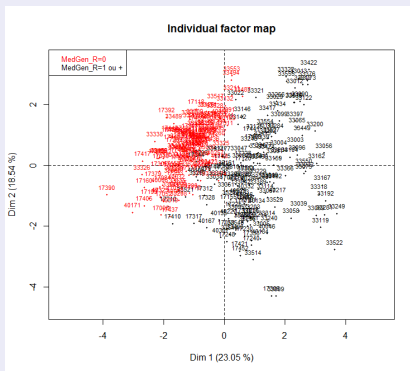


Le groupe "Logement_Services" est prépondérant dans la construction de l'axe 1.

Utilisation de MFAmix dans :

Représentation des individus :

```
plot(res, choix="ind", invisible="quali", habillage="MedGen_R", axes=c(1,2), col.
      hab=c("red", "black"), cex=0.7)
```

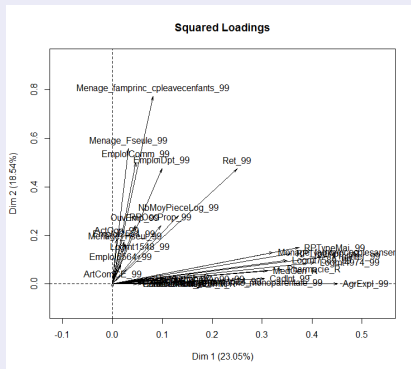


L'axe 1 discrimine bien les communes n'ayant pas de médecins des communes possédant un ou plusieurs médecins.

Utilisation de MFAmix dans :

Représentation des loadings au carré :

```
plot(res$global.pca, choice="var", axes=c(1,2), main="Squared Loadings")
```

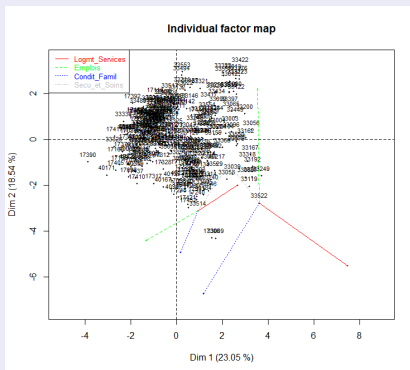


Coefficient de corrélation au carré pour les variables quantitatives et rapport de corrélation pour les variables qualitatives.

Utilisation de MFAmix dans :

Représentation des individus partiels :

```
plot(res, choix="ind", invisible="quali", partial=c("33522", "33514"), habillage="group", axes=c(1,2), col.hab=c("red", "green", "blue", "grey"), cex=0.7)
```



Ces deux communes sont proches d'un point de vue global mais ont des scores très éloignés sur la thématique de l'emploi.

CONCLUSION

- La méthode MFAmix permet d'effectuer une AFM avec des variables mixtes à l'intérieur des groupes, ce qui était nécessaire dans notre étude de cas
- 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
- Code  et sorties graphiques inspirées de FactoMineR
- Méthode PCAmix et MFAmix intégrée dans le package  PCAmixdata bientôt disponible sur le CRAN (déjà 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...