Rotation orthogonale dans PCAMIX

Marie Chavent, V. Kuentz, Jérôme Saracco

in : 18èmes Rencontres de la Société Francophone de Classification, Orléans, 28-30 septembre 2011

Rotation orthogonale dans PCAMIX

Marie Chavent [*,**], Vanessa Kuentz[***]
Jérôme Saracco[*,**]

[*]Université de Bordeaux, IMB, CNRS, UMR 5251, France
{marie.chavent,jerome.saracco}@math.u-bordeaux1.fr
[**]INRIA Bordeaux Sud-Ouest, CQFD team, France
[***]Cemagref, UR ADBX, F-33612 Cestas Cedex, France
vanessa.kuentz@cemagref.fr

Résumé. La rotation orthogonale dans PCAMIX a été initialement introduite par Kiers (1991). PCAMIX est une méthode d'analyse en composantes principales pour un mélange de variables quantitatives et qualitatives qui inclut comme cas particuliers l'Analyse en Composantes Principales (ACP) et l'Analyse des Correspondances Multiples (ACM). Dans ce papier, nous donnons une nouvelle présentation de PCAMIX où les composantes principales et les loadings sont obtenues à l'aide d'une Décomposition en Valeurs Singulières. Dans ce contexte, nous proposons une nouvelle expression analytique directe pour l'angle varimax de rotation dans PCAMIX. L'algorithme de rotation qui en résulte est simple et relativement peu coûteux en temps de calculs. Une application sur un jeu de données réel illustre l'intérêt pratique de la rotation. L'ensemble des codes sera prochainement disponible dans un package R nommé *PCAmixdata*.

Mots-clés: mélange de données quantitatives et qualitatives, analyse en composantes principales, analyse des correspondances multiples, rotation.

1 Introduction

Différents critères ont été proposés pour la rotation en ACP. Le plus célèbre est varimax, introduit par Kaiser en 1958. L'idée est de maximiser la variance des colonnes de la matrice des loadings qui contient les corrélations au carré des variables aux composantes principales. Des groupes de variables se forment, facilitant ainsi l'interprétation. En Analyse des Correspondances, différents travaux ont été récemment réalisés (voir par exemple van de Velden and Kiers, 2005). D'autre part, Kiers (1991) a considéré la rotation orthogonale dans la méthode PCAMIX. Cette méthode d'analyse factorielle pour un ensemble de données qualitatives et quantitatives inclut comme cas particuliers l'ACP et l'ACM. Dans cet article, nous proposons une écriture de PCAMIX sous forme de Décomposition en Valeurs Singulières qui facilite l'utilisation de la rotation.

L'idée de la rotation orthogonale dans PCAMIX utilise la définition des loadings au carré (voir Chavent et al., 2011 pour plus de détails). Pour une variable quantitative, il s'agit du carré de sa corrélation avec la composante principale. Pour une variable qualitative, c'est le rapport de corrélation qui est utilisé. La fonction varimax est alors appliquée à la matrice des

in : 18èmes Rencontres de la Société Francophone de Classification, Orléans, 28-30 septembre 2011

loadings au carré, conduisant à un nouveau problème d'optimisation. En deux dimensions, la définition de la matrice de rotation orthogonale optimale s'obtient en résolvant un problème non contraint. Nous utilisons différentes astuces et formules trigonométriques pour annuler la dérivée et obtenir une nouvelle écriture explicite de l'angle optimal de rotation (planaire). Cette écriture est plus simple que celle proposée par Kiers et moins coûteuse en temps de calcul. Dans le cas de plus de deux dimensions, nous utilisons l'algorithme proposé par Kaiser (1958) qui consiste à réaliser des rotations successives de paires de facteurs jusqu'à convergence.

2 Un exemple illustratif

Nous appliquons l'approche de rotation sur le jeu de données "Prostate" utilisé entre autres par Hunt et al. (2003). Il concerne 506 patients atteints du cancer de la prostate ayant suivi un essai clinique aléatoire visant à comparer quatre traitements. Ces données sont disponibles dans le package R "Hmisc" développé par Harrell (2010). Les données sont mixtes : 8 variables sont quantitatives ("age", "poids", "pression sanguine systolique", "pression sanguine diastolique", "sérum hémoglobine", "taille de la tumeur", "indice de niveau de la tumeur", "sérum prostatique") et 4 variables sont qualitatites ("niveau de performance", "historique cardiovasculaire", "électrocardiogramme", "métastases"). Voici un exemple de code R suivi de quelques sorties graphiques et numériques obtenues avec le package "PCAmixdata" (prochainement disponible).

Les sorties numériques fournies par le package comprennent les scores des composantes principales avant et après rotation (standardisées dans ce cas), les coordonnées des modalités des variables qualitatives avant et après rotation, le cercle des corrélations pour les variables quantitatives avant et après rotation ainsi que les loadings au carré de chaque variable (carré de sa corrélation linéaire si elle est quantitative, rapport de corrélation si elle est qualitative) avec la composante principale.
Les lignes de code,

```
require(Hmisc)
getHdata(prostate)
require(PCAmixdata)
res<-PCAmix(X.quanti=prostate.quanti,X.quali=prostate.quali,ndim=4)
```

permettent de lancer la méthode PCAMIX et de conserver 4 composantes principales. Les graphiques obtenus sont présentés dans la Figure 1. Pour ne pas surcharger le résumé, nous nous sommes limités aux deux premières composantes principales.

La ligne suivante,

```
rot<-PCArot(res,dim=4)
```

permet d'effectuer une rotation sur les 4 composantes principales obtenues avec PCAMIX (voir Figure 1).

La rotation des composantes principales conduit à une meilleure association entre les variables (voir Tableau 2 et Figure 1) : le poids est associé avec les pressions systolique et diastolique, le sérum hémoglobine avec la taille et le niveau de la tumeur, l'âge ne s'associe pas avec les autres variables et le sérum prostatique se distingue également.

	Before rotation				*After rotation*			
	1	2	3	4	1	2	3	4
age	-0.06	0.10	-0.58	0.15	-0.03	-0.06	**-0.61**	-0.09
wt	-0.44	0.20	0.29	0.01	-0.26	**0.46**	0.18	-0.08
sbp	-0.36	0.77	0.12	0.06	0.05	**0.84**	-0.15	-0.01
dbp	-0.43	0.65	0.30	0.00	-0.05	**0.83**	0.06	-0.04
hg	-0.51	-0.09	0.32	0.02	**-0.46**	0.25	0.28	-0.13
sz	0.42	0.32	0.00	-0.40	**0.65**	0.06	0.06	-0.08
sg	0.57	0.27	0.15	-0.38	**0.72**	0.00	0.21	0.04
ap	0.53	0.14	0.28	0.56	0.18	-0.02	0.03	**0.82**
pf	0.23	0.16	0.22	0.51	0.18	0.02	0.17	**0.76**
hx	0.06	0.05	0.26	0.09	0.04	0.02	**0.40**	0.00
ekg	0.04	0.20	0.31	0.13	0.09	0.15	**0.41**	0.04
bm	0.48	0.09	0.00	0.00	**0.46**	0.00	0.00	0.11

TAB. 1 – *Loadings au carré avec les 4 premières composantes avant et après rotation*

References

Chavent, M., Kuentz, V., Saracco, J., (2011), Orthogonal rotation in PCAMIX, *Under review*.

Harrell F. E., (2010), The Hmisc package, CRAN R Project.

Hunt, L.A. and Jorgensen, M.A., (2003), Mixture model clustering for mixed data with missing information, *Computational Statistics and Data Analysis*, **41**, 429-440.

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.

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

Summary

Orthogonal rotation in PCAMIX has been initially introduced by Kiers (1991). PCAMIX is a factorial method for a mixture of quantitative and qualitative variables. It includes as ordinary Principal Component Analysis (PCA) and Multiple Correspondence Analysis (MCA) as special cases. In this paper, we give a new presentation of PCAMIX where the principal components and the squared loadings are obtained from a Singular Value Decomposition. In this context we give a new analytic expression of the varimax angle for rotation in PCAMIX. The resulting rotation algorithm is simple and computationnaly efficient. An application on real data shows the benefits of using rotation. All source codes will be soon available in the R package *PCAmixdata*.

Mots-clés: mixture of qualitative and quantitative data, principal component analysis, multiple correspondance analysis, rotation.

Rotation orthogonale dans PCAMIX

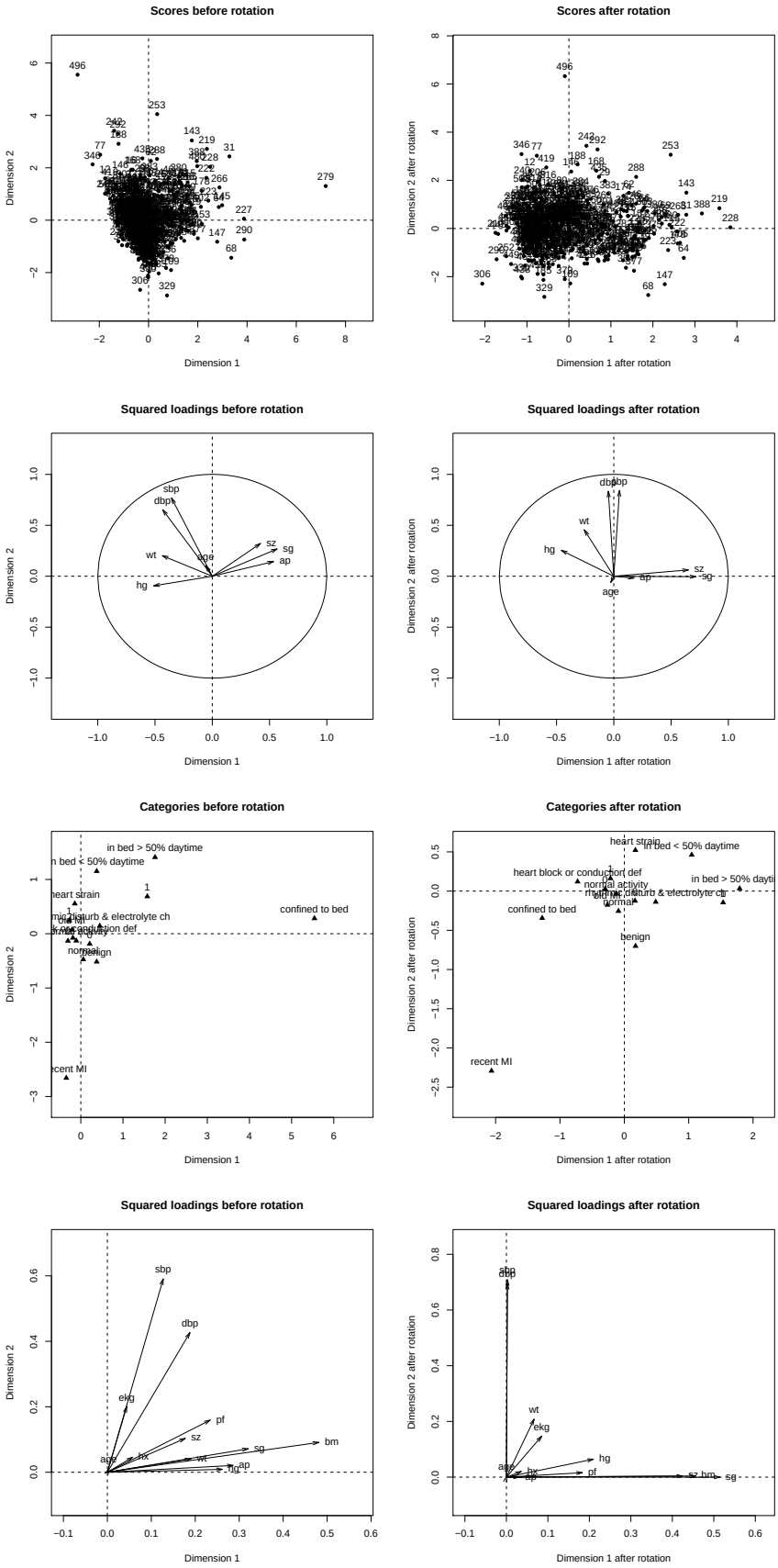

FIG. 1 – *Scores, cercles de corrélation, modalités et loadings au carré avant et après rotation.*