Rotation orthogonale dans PCAMIX

Marie Chavent & Vanessa Kuentz
& Jérôme Saracco

IMB, University of Bordeaux, France
INRIA Bordeaux Sud-Ouest, CQFD Team
CEMAGREF, UR ADBX, Bordeaux, France

18èmes rencontres de la SFC
Université d'Orléans, 28-30 septembre 2011

Outline

Introduction

Principal Component Analysis of a mixture of quantitative and qualitative data

- PCAMIX (Kiers, 1991) and AFDM (Pagès, 2004)
- It includes PCA and MCA as special cases
- Function `AFDM` in the R package **FactoMineR**
- Rotation in PCAMIX
  $\hookrightarrow$ Singular Value Decomposition presentation of PCAMIX

Introduction

Orthogonal rotation in PCAMIX

- Kier's (1991) gives a matrix reformulation of the varimax function in the context of PCAMIX.
- We propose a computationaly effcient procedure for varimax rotation in PCAMIX and a direct solution for the optimal angle of rotation.
- We implemented the SVD based approach of PCAMIX and the rotation procedure in the R package **PCAmixdata**.

Outline

The PCAMIX method

- Let $\mathbf{X}_1$ be a $n \times p_1$ **quantitative** data matrix where n observations are described on p_1 quantitative variables
- Let $\mathbf{X}_2$ be a $n \times p_2$ **qualitative** data matrix where the same n observations are described on p_2 qualitative variables
- Let $p = p_1 + p_2$ be the total number of variables and m the total number of categories
- Let k the number of components required in PCAMIX with $2 \leq k \leq \min(n, p_1 + m - p_2)$

The PCAMIX method

The procedure is carried out according to the following steps:

1. Recoding of $\mathbf{X}_1$ and $\mathbf{X}_2$:
   - $\mathbf{Z}_1$ is the standardized version of the quantitative matrix $\mathbf{X}_1$
   - $\mathbf{Z}_2 = \mathbf{JGD}^{-1/2}$ is the standardized version of the indicator matrix $\mathbf{G}$ of the qualitative matrix $\mathbf{X}_2$, where $\mathbf{D}$ is the diagonal matrix of frequencies of the categories and $\mathbf{J} = \mathbf{I} - \mathbf{1}'\mathbf{1}/n$ is the centering operator

$\hookrightarrow \mathbf{Z} = \frac{1}{\sqrt{n}}(\mathbf{Z}_1|\mathbf{Z}_2)$ is the $n \times (p_1 + m)$ matrix of interest

The PCAMIX method

2. Singular Value Decomposition of $\mathbf{Z}$

$$\mathbf{Z} = \mathbf{U}\Lambda\mathbf{V}'$$

$\hookrightarrow \mathbf{F} = \sqrt{n}\mathbf{U}_k$ is the $n \times k$ matrix of the PC scores where $\mathbf{U}_k$ is the matrix of the k first columns of $\mathbf{U}$
$\hookrightarrow \mathbf{A} = \mathbf{V}_k\Lambda_k$ is the $n \times k$ matrix of the PC "loadings" where $\mathbf{V}_k$ is the matrix of the k first columns of $\mathbf{V}$ and Λ_k the diagonal matrix of the k first singular values

The PCAMIX method

3. Write $\mathbf{A} = \begin{pmatrix} \mathbf{A}_1 \\ \mathbf{A}_2 \end{pmatrix}$ with:
   - $\mathbf{A}_1$ the $p_1 \times k$ matrix of the loadings (the correlations) of the quantitative variables.
   - $\mathbf{DA}_2$ the $m \times k$ matrix of the principal coordinates of the categories.

$\hookrightarrow$ Correlation circle of the quantitative variables

$\hookrightarrow$ Plot of the categories

The PCAMIX method

4. Calculate the $p \times k$ matrix **C** of the **squared loadings**:

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

where I_j is the set of row indices of **A** associated with the categories of the qualitative variable j.

$\hookrightarrow$ c_{jl} is a **squared correlation** if j is quantitative

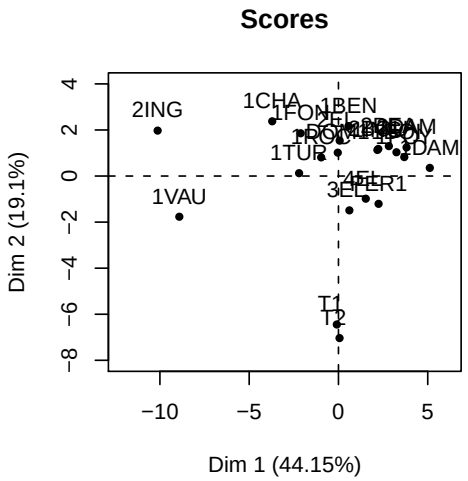
$\hookrightarrow$ c_{jl} is a **correlation ratio** if j is qualitative

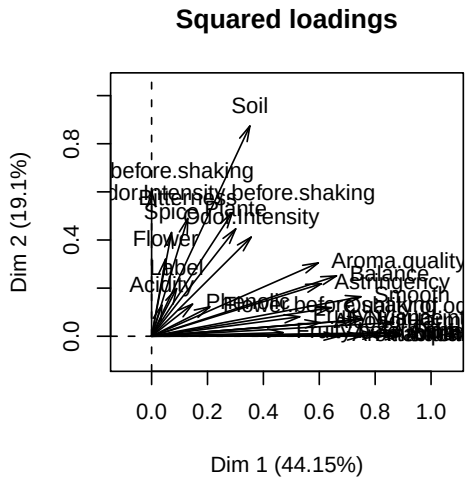
$\hookrightarrow$ quantitative and qualitative variables on the same plot

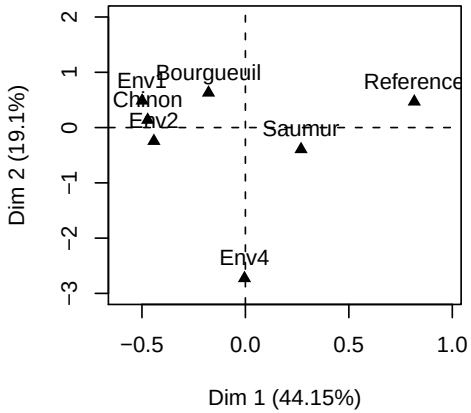
The PCAmixdata R package

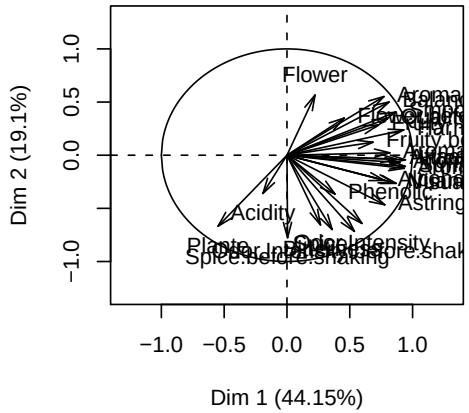
```
> require(PCAmixdata)
> data(wine)
> head(wine[,c(1:4)])
      Label        Soil       Odor.Intensity  Aroma.quality
2EL   Saumur       Env1       3.07            3.00
1CHA  Saumur       Env1       2.96            2.82
1FON  Bourgueuil   Env1       2.85            2.92
1VAU  Chinon       Env2       2.80            2.59
1DAM  Saumur       Reference  3.60            3.42
2BOU  Bourgueuil   Reference  2.85            3.11

> X.quanti <- wine[,c(3:29)]
> X.quali <- wine[,c(1,2)]
> pca <- PCAmix(X.quanti,X.quali,ndim=10)
```

The PCAmixdata R package









Outline

Rotation in PCAMIX

Why using rotation ?

$\mathbf{FA}'$ is a rank k least squares approximation of $\mathbf{Z}$

- Let $\mathbf{T}$ an orthonormal rotation matrix: $\mathbf{TT}\prime = \mathbf{T}'\mathbf{T} = \mathbf{I}_k$
- Let $\widetilde{\mathbf{F}} = \mathbf{FT}$ and $\widetilde{\mathbf{A}} = \mathbf{AT}$ the scores and "loadings" matrices after rotation

$\hookrightarrow \mathbf{FA}' = \widetilde{\mathbf{F}}\widetilde{\mathbf{A}}'$

$\hookrightarrow$ This approximation is not unique over orthogonal rotations

$\hookrightarrow$ Improve the interpretability: find $\mathbf{T}$ in such a way that squared loadings are either large (close to 1) or close to zero.

Rotation in PCAMIX

The varimax problem.

$$\begin{array}{ll} \max\limits_{\mathbf{T}} & f(\mathbf{T}), \\ \text{s.t.} & \mathbf{T}\mathbf{T}' = \mathbf{T}'\mathbf{T} = \mathbf{I}_k \end{array}$$

where f measures the simplicity of the interpretation of the components after rotation.

In PCA the varimax function is:

$$f(\mathbf{T}) = \sum_{l=1}^{k} \sum_{j=1}^{p} (\tilde{a}_{jl}{}^2)^2 - \frac{1}{p} \sum_{l=1}^{k} \left(\sum_{j=1}^{p} \tilde{a}_{jl}{}^2 \right)^2$$

Rotation in PCAMIX

In PCAMIX

- The terms $\tilde{a}_{jl}{}^2$ are replaced by $\tilde{c}_{jl} = \sum_{s \in I_j} \tilde{a}_{sl}{}^2$

  $\hookrightarrow$ The varimax function f is:

$$f(\mathbf{T}) = \sum_{l=1}^{k} \sum_{j=1}^{p} (\tilde{c}_{jl})^2 - \frac{1}{p} \sum_{l=1}^{k} \left(\sum_{j=1}^{p} \tilde{c}_{jl} \right)^2$$

- The squared loadings after rotation $\tilde{c}_{jl}$ are squared correlations or correlation ratios with the rotated components in $\widetilde{\mathbf{F}}$.

Planar rotation ($k = 2$)

- Let θ be a planar angle of rotation
- The rotation matrix is then:

$$\mathbf{T} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

$\hookrightarrow$ The varimax rotation problem is then rewritten as:

$$\max_{\theta \in \mathbb{R}} f(\theta)$$

where...

Planar rotation ($k = 2$)

... where

$$f(\theta) = \sum_{j=1}^{p} \left(\sum_{s \in I_j} \tilde{a}_{s1}{}^2 \right)^2 + \sum_{j=1}^{p} \left(\sum_{s \in I_j} \tilde{a}_{s2}{}^2 \right)^2 - \frac{1}{p} \left(\sum_{j=1}^{p} \sum_{s \in I_j} \tilde{a}_{s1}{}^2 \right)^2 - \frac{1}{p} \left(\sum_{j=1}^{p} \sum_{s \in I_j} \tilde{a}_{s2}{}^2 \right)^2$$

with $\widetilde{\mathbf{A}} = \mathbf{AT}$ which gives:

$$\begin{aligned} \tilde{a}_{s1} &= a_{s1} \cos\theta + a_{s2} \sin\theta \\ \tilde{a}_{s2} &= -a_{s1} \sin\theta + a_{s2} \cos\theta \end{aligned}$$

Planar rotation ($k = 2$)

We demonstrate that:

$$f(\theta) = f(0) + \frac{\rho}{4p}\left(\cos(4\theta - \psi) - \cos\psi\right)$$

where ρ and ψ are defined by :

$$\rho = (a^2 + b^2)^{1/2} \quad , \quad \cos\psi = b/\rho \quad , \quad \sin\psi = a/\rho$$

and where a and b are given by ...

Planar rotation ($k = 2$)

...where a and b are given by :

$$a = 2p\sum_{j=1}^{p} u_j v_j - 2\sum_{j=1}^{p} u_j \sum_{j=1}^{p} v_j$$

$$b = p\sum_{j=1}^{p}({u_j}^2 - {v_j}^2) - \left(\sum_{j=1}^{p} u_j\right)^2 + \left(\sum_{j=1}^{p} v_j\right)^2$$

and where u_j and v_j are defined by :

$$u_j = \sum_{s \in I_j}(a_{s1}^2 - a_{s2}^2) \quad \text{and} \quad v_j = 2\sum_{s \in I_j} a_{s1} a_{s2}$$

Planar rotation ($k = 2$)

$$f(\theta) = f(0) + \frac{\rho}{4p}\left(\cos(4\theta - \psi) - \cos\psi\right)$$

is maximum for

$$\cos(4\theta - \psi) = 1 \Leftrightarrow 4\theta - \psi = 2k\pi$$

$\hookrightarrow$ the optimal angles are :

$$\theta = \frac{\psi}{4} + k\frac{\pi}{2}, \quad k \in \mathbb{Z}$$

The iterative rotation procedure ($k > 2$)

1. Initialization :
   - Calculate **F** and **A** with PCAMIX
   - $\widetilde{\mathbf{F}} = \mathbf{F}$ and $\widetilde{\mathbf{A}} = \mathbf{A}$
2. For each pair of dimensions (l, t) :
   - Calculate $\theta = \Psi/4$ with

$$\psi = \begin{cases} \arccos\left(\frac{b}{\sqrt{a^2+b^2}}\right) & \text{if} \quad a \geq 0, \\ -\arccos\left(\frac{b}{\sqrt{a^2+b^2}}\right) & \text{if} \quad a \leq 0. \end{cases}$$

   - ...

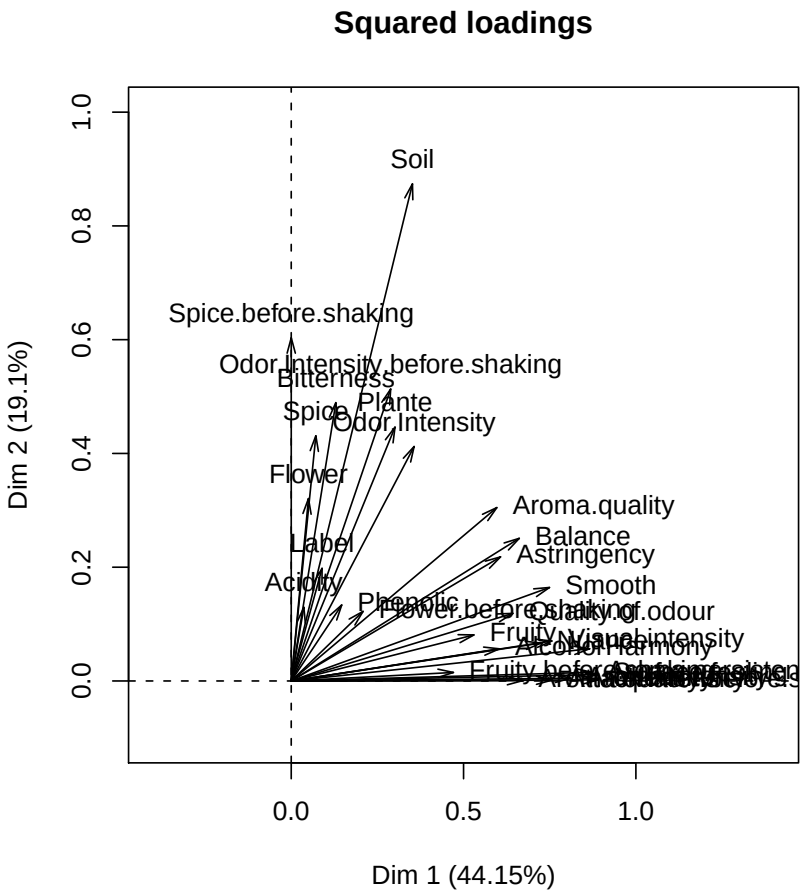
The iterative rotation procedure ($k > 2$)

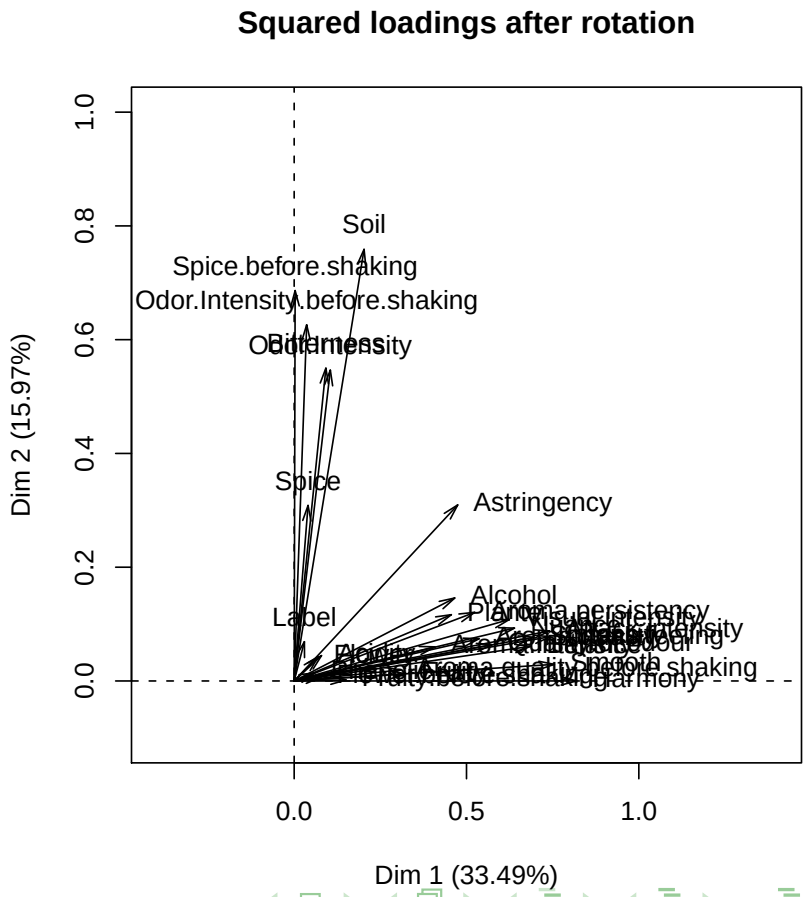
2. For each pair of dimensions (l, t) :
   - ...
   - $\mathbf{T} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$,
   - **Update $\widetilde{\mathbf{F}}$ and $\widetilde{\mathbf{A}}$** by rotation of their l-th and t-th columns
3. Repeat the previous step until the $k(k-1)/2$ angles θ are equal to zero.

$\hookrightarrow$ **Outputs after rotation**: $\widetilde{\mathbf{F}}$ (scores of the observations), $\widetilde{\mathbf{A}}_1$ (correlations of the quantitative variables), $\mathbf{D}\widetilde{\mathbf{A}}_2$ (coordinates of the categories), $\widetilde{\mathbf{C}}$ (squared loadings).

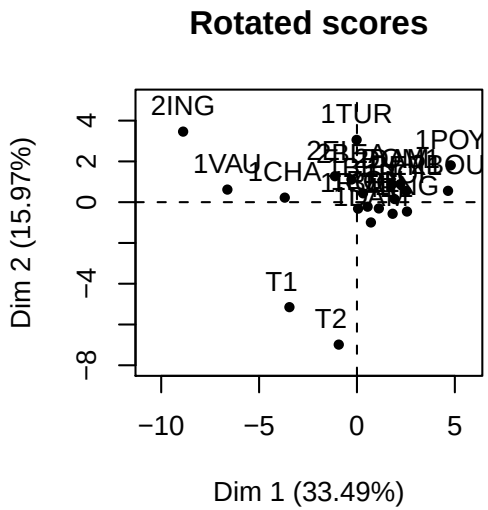
The PCAmixdata R package

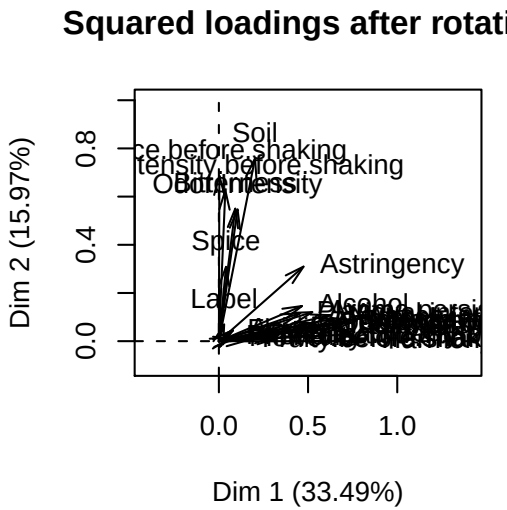
```
> rot<-PCArot(pca,dim=8)
```

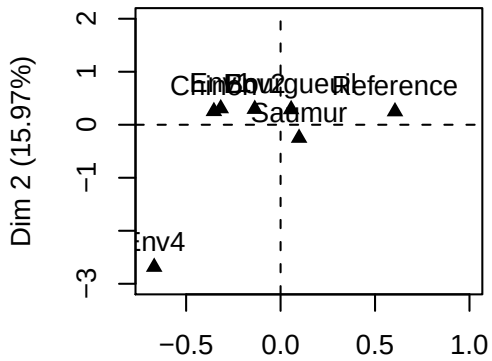


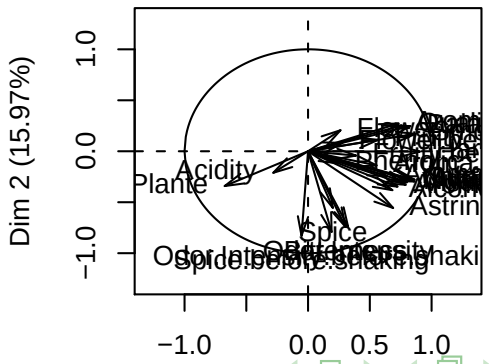


The PCAmixdata R package









Outline

A simulation study: comparison of computational times

Comparison with Kier's original approach:

- The same iterative rotation procedure but with:
  - Kiers' original PCAMIX approach in the initialization step
  - The coefficients a and b in step 2 calculated according to their expression obtained with Kiers' matrix reformulation of the varimax function.

  $\hookrightarrow$ Implemented in R for comparison purpose
- The rotation procedure based on the SVD approach of PCAMIX:

  $\hookrightarrow$ Implemented in the R package PCAmixdata

A simulation study: comparison of computational times

Simulated datasets with varying parameters:

- the number p of variables ($p/2$ quantitative and $p/2$ qualitative)
- the number n of observations.

$\hookrightarrow$ For each set of parameters (n, p), 20 simulations are drawn.

More precisely :

- A dataset is drawn from a multivariate normal distribution with a covariance matrix $\Sigma = \mathbf{Q}'\mathbf{Q}$ where $\mathbf{Q}$ is drawn from a uniform distribution on the interval $[0.2; 0.4]$.
- The $p/2$ last variables are distributed in three equal-count categories.

A simulation study: comparison of computational times

↪ The two procedures are applied on each dataset for $k = 2$

↪ The median computation times (over the 20 replications) are calculated

		p=10	p=50	p=100	p=200
n=50	Matrix reformulation	0.05	0.12	0.22	0.44
n=50	SVD	0.02	0.06	0.12	0.27
n=100	Matrix reformulation	0.14	0.33	0.56	1.04
n=100	SVD	0.02	0.09	0.17	0.34
n=200	Matrix reformulation	0.55	1.12	1.86	3.38
n=200	SVD	0.02	0.11	0.26	0.53
n=400	Matrix reformulation	2.15	4.32	7.1	12.65
n=400	SVD	0.03	0.16	0.37	0.89
n=800	Matrix reformulation	10.06	19.27	30.54	error
n=800	SVD	0.05	0.25	0.58	1.79

A simulation study: comparison of computational times

↪ Ratio between the median computation time of the two rotation procedures

	p=10	p=50	p=100	p=200
n=50	2.9	2.0	1.8	1.6
n=100	8.7	3.8	3.3	3.0
n=200	23.2	10.3	7.0	6.4
n=400	69.4	27.7	19.0	14.2
n=800	214.1	77.4	52.9	error

↪ From 2 two 214 times faster !

References

- Chavent, M., Kuentz, V., Saracco, J. (2011), Orthogonal rotation in PCAMIX, *in revision*.
- Kiers, H.A.L., (1991), Simple structure in Component Analysis Techniques for mixtures of qualitative and quantitative variables, *Psychometrika*, **56**, 197-212.
- Pagès, J., (2004), Analyse Factorielle de données mixtes [Factor Analysis for Mixed Data], *Revue de Statistique Appliquée*, **52**(4), 93-11.