Multivariate analysis of mixed data: The PCAmixdata R package

Marie Chavent
In collaboration with: Vanessa Kuentz, Amaury Labenne, Benoît Liquet, Jérôme Saracco

University of Bordeaux, France
Inria Bordeaux Sud-Ouest, CQFD Team
Irstea, UR ADBX, cestas, France
The University of Queensland, Australia
Multivariate analysis of a mixture of numerical and categorical data

Three main functions:

- Function **PCAmix** for principal component analysis (PCA) of mixed data.
  - Includes standard PCA and MCA (multiple component analysis) as special cases.

- Function **PCArot** for orthogonal rotation in PCAmix.
  - Includes standard varimax rotation and rotation in MCA as special cases.

- Function **MFAmix** for multiple factor analysis (MFA) for multiple-table mixed data.

https://github.com/chavent/PCAmixdata
Outline

1. PCAmix

2. PCArot

3. MFAmix
Principal component analysis of mixed data

Several implementations already in R:

- Function **FAMD** in the R package **FactoMineR**.
  - Implements the method designed by Pagès (2004).
- Function **dudi.mix** in the R package **ade4**.
- Function **PCAmix** in the R package **PCAmixdata**.
  - Implements a single PCA with metrics based on a GSVD of preprocessed data.

⇒ Three different coding scheme but identical numerical results.
A real data example

The **gironde** data are available in the package

```r
library(PCAmixdata)
data(gironde)
```

- They characterize **living conditions** in Gironde, a southwest region in France.
- 542 cities are described with 27 variables separated into 4 groups (Employment, Housing, Services, Environment).
  - Four datatables
A mixed data type example

The datatable `housing` is mixed.

➔ 3 numerical and 2 categorical variables.

```r
housing <- gironde$housing
```

```
head(housing)
```

<table>
<thead>
<tr>
<th>##</th>
<th>density</th>
<th>primaryres</th>
<th>owners</th>
<th>houses</th>
<th>council</th>
</tr>
</thead>
<tbody>
<tr>
<td>##</td>
<td>ABZAC</td>
<td>132</td>
<td>89</td>
<td>64</td>
<td>inf 90% sup 5%</td>
</tr>
<tr>
<td>##</td>
<td>AILLAS</td>
<td>21</td>
<td>88</td>
<td>77</td>
<td>sup 90% inf 5%</td>
</tr>
<tr>
<td>##</td>
<td>AMBARES</td>
<td>532</td>
<td>95</td>
<td>66</td>
<td>inf 90% sup 5%</td>
</tr>
<tr>
<td>##</td>
<td>AMBES</td>
<td>101</td>
<td>94</td>
<td>67</td>
<td>sup 90% sup 5%</td>
</tr>
<tr>
<td>##</td>
<td>ANDERNOS</td>
<td>552</td>
<td>62</td>
<td>72</td>
<td>inf 90% inf 5%</td>
</tr>
<tr>
<td>##</td>
<td>ANGLADE</td>
<td>64</td>
<td>81</td>
<td>81</td>
<td>sup 90% inf 5%</td>
</tr>
</tbody>
</table>
Two data sets:

→ a numerical data matrix $X_1$ of dimension $542 \times 3$.
→ a categorical data matrix $X_2$ of dimension $542 \times 2$.

```r
split<-splitmix(housing)
X1<-split$X.quanti
X2<-split$X.quali
```

<table>
<thead>
<tr>
<th>head(X1)</th>
<th>head(X2)</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>##</code></td>
<td><code>##</code></td>
</tr>
<tr>
<td><code>##</code></td>
<td><code>##</code></td>
</tr>
<tr>
<td><code>##</code></td>
<td><code>##</code></td>
</tr>
<tr>
<td><code>##</code></td>
<td><code>##</code></td>
</tr>
<tr>
<td><code>##</code></td>
<td><code>##</code></td>
</tr>
</tbody>
</table>

UseR!2015

June 30 - July 3, Aalborg, Denmark
The PCAmix method

An simple algorithm in three main steps

1. Preprocessing step.
2. GSVD (Generalized Singular Value Decomposition) step.

Some notations:

- Let $X_1$ be a $n \times p_1$ numerical data matrix.
- Let $X_2$ be a $n \times p_2$ categorical data matrix.
- Let $m$ be the total number of categories.
Preprocessing step

1. Build a **numerical data matrix** $Z = (Z_1|Z_2)$ of dimension $n \times (p_1 + m)$ with:
   - $Z_1$ the standardized version of the matrix $X_1$.
   - $Z_2$ the centered indicator matrix of the levels of $X_2$.

2. Build the diagonal matrix $N$ of the **weights of the rows**.
   - The $n$ rows are weighted by $\frac{1}{n}$.

3. Build the diagonal matrix $M$ of the **weights of the columns**.
   - The $p_1$ first columns are weighted by 1.
   - The $m$ last columns are weighted by $\frac{n}{n_s}$, with $n_s$ the number of observations with level $s$.

$\Rightarrow$ The **total variance** is $p_1 + m - p_2$. 
The GSVD (Generalized Singular Value Decomposition) of $Z$ with the metrics $N$ and $M$ gives the decomposition

$$Z = UDV^t$$  \hspace{1cm} (1)

where

- $D = \text{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_r})$ is the $r \times r$ diagonal matrix of the singular values of $ZMZ^tN$ and $Z^tNZM$, and $r$ denotes the rank of $Z$;

- $U$ is the $n \times r$ matrix of the first $r$ eigenvectors of $ZMZ^tN$ such that $U^tNU = I_r$;

- $V$ is the $p \times r$ matrix of the first $r$ eigenvectors of $Z^tNZM$ such that $V^tMV = I_r$. 

A real data example

The PCAmix method

Principal component prediction

GSVD step

UseR!2015
June 30 - July 3, Aalborg, Denmark
Scores processing step

1. The set of factor scores for rows is computed as:

\[ F = UD. \]

\( \rightarrow \) Principal Component scores

2. The set of factor scores for columns is computed as:

\[ A = MVD. \]

\( \rightarrow \) In standard PCA: \( A = VD. \)
The R function

```r
args(PCAmix)
```

```r
# function (X.quanti = NULL, X.quali = NULL, ndim = 5, rename.level = FALSE, 
#    weight.col = NULL, weight.row = NULL, graph = TRUE)
# NULL
```

```r
PCAmix(X.quanti=X1,X.quali=X2,ndim=2)
```

```r
# Call:
# PCAmix(X.quanti = X1, X.quali = X2, ndim = 2)
# Method = Principal Component of mixed data (PCAmix)
```

<table>
<thead>
<tr>
<th>name</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1,]</td>
<td>&quot;$eig&quot;           eigenvalues of the principal components (PC) &quot;</td>
</tr>
<tr>
<td>[2,]</td>
<td>&quot;$ind&quot;          results for the individuals (coord,contrib,cos2)</td>
</tr>
<tr>
<td>[3,]</td>
<td>&quot;$quanti&quot;       results for the quantitative variables (coord,contrib,cos2)</td>
</tr>
<tr>
<td>[4,]</td>
<td>&quot;$levels&quot;       results for the levels of the qualitative variables (coord,contrib,cos2)</td>
</tr>
<tr>
<td>[5,]</td>
<td>&quot;$quali&quot;        results for the qualitative variables (contrib,relative contrib)</td>
</tr>
<tr>
<td>[6,]</td>
<td>&quot;$sqload&quot;       squared loadings</td>
</tr>
<tr>
<td>[7,]</td>
<td>&quot;$coef&quot;         coef of the linear combinations defining the PC</td>
</tr>
</tbody>
</table>

UseR!2015 June 30 - July 3, Aalborg, Denmark
Principal Component scores of the observations

```r
obj <- PCAmix(X.quanti=X1,X.quali=X2,ndim=2)
```

```r
head(obj$ind$coord)
```

<table>
<thead>
<tr>
<th></th>
<th>dim 1</th>
<th>dim 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABZAC</td>
<td>2.36</td>
<td>0.024</td>
</tr>
<tr>
<td>AILLAS</td>
<td>-0.88</td>
<td>0.123</td>
</tr>
<tr>
<td>AMBARES</td>
<td>2.62</td>
<td>0.800</td>
</tr>
<tr>
<td>AMBES</td>
<td>0.93</td>
<td>0.919</td>
</tr>
<tr>
<td>ANDERNOS</td>
<td>1.18</td>
<td>-2.481</td>
</tr>
<tr>
<td>ANGLADE</td>
<td>-1.01</td>
<td>-0.424</td>
</tr>
</tbody>
</table>

```r
plot(obj,choice="ind")
```
Loadings of the numerical variables

```
head(obj$quanti.cor)
```

```
##       dim1 dim2
## density 0.704 0.25
## primaryres -0.019 0.97
## owners -0.858 0.13
```

```
plot(obj,choice="cor")
```

The (non standardized) loadings are correlations.
Scores of the levels of the categorical variables

```
head(obj$categ.coord)
```

```
## dim1  dim2
## inf 90% 1.63 -0.339
## sup 90% -0.42 0.087
## inf 5%  -0.40 -0.065
## sup 5%   1.52 0.245
```

```
plot(obj,choice="levels")
```

The barycentric property is still verified.
Contributions of the variables

The contribution $c_{j\alpha}$ of a variable $j$ to the component $\alpha$ is:

$$
\begin{align*}
    c_{j\alpha} &= a^2_{j\alpha} = \text{Squared correlation} & \text{if variable } j \text{ is numerical,} \\
    c_{j\alpha} &= \sum_{s \in I_j} \frac{n}{n_s} a^2_{s\alpha} = \text{Correlation ratio} & \text{if variable } j \text{ is categorical.}
\end{align*}
$$

Called squared loadings in varimax criterion for PC rotation.
Each principal component $f_\alpha$ writes as a linear combination of the columns of $X = (X_1|G)$ where $X_1$ is the numerical data matrix and $G$ is the indicator matrix of the levels of the matrix $X_2$:

$$f_\alpha = \beta_0 + \sum_{j=1}^{p_1+m} \beta_j x_j$$

with:

$$\beta_0 = -\sum_{k=1}^{p_1} v_{k\alpha} \bar{x}_k - \sum_{k=p_1+1}^{p_1+m} v_{k\alpha},$$

$$\beta_j = v_{j\alpha} \frac{1}{s_j}, \text{ for } j = 1, \ldots, p_1$$

$$\beta_j = v_{j\alpha} \frac{n}{n_j}, \text{ for } j = p_1 + 1, \ldots, p_1 + m$$
The method predict()

Coefficients of the PC found on the learning set (without the 5 first cities)

```r
test <- 1:5
obj2 <- PCAmix(X1[-test,],X2[-test,],graph=FALSE,ndim=3)
data.frame(obj2$coef)
```

<table>
<thead>
<tr>
<th></th>
<th>dim1</th>
<th>dim2</th>
<th>dim3</th>
</tr>
</thead>
<tbody>
<tr>
<td>const</td>
<td>3.64214</td>
<td>-9.02951</td>
<td>1.5950</td>
</tr>
<tr>
<td>density</td>
<td>0.00086</td>
<td>0.00048</td>
<td>0.0015</td>
</tr>
<tr>
<td>primaryres</td>
<td>-0.00129</td>
<td>0.09355</td>
<td>-0.0179</td>
</tr>
<tr>
<td>owners</td>
<td>-0.05065</td>
<td>0.01207</td>
<td>-0.0047</td>
</tr>
<tr>
<td>inf 90%</td>
<td>1.03579</td>
<td>-0.32092</td>
<td>-0.4617</td>
</tr>
<tr>
<td>sup 90%</td>
<td>-0.26076</td>
<td>0.08079</td>
<td>0.1162</td>
</tr>
<tr>
<td>inf 5%</td>
<td>-0.25129</td>
<td>-0.05706</td>
<td>0.2704</td>
</tr>
<tr>
<td>sup 5%</td>
<td>0.96441</td>
<td>0.21898</td>
<td>-1.0379</td>
</tr>
</tbody>
</table>

Scores of the 5 first cities on the principal components

```r
predict(obj2,X1[test,],X2[test,])
```

<table>
<thead>
<tr>
<th></th>
<th>dim1</th>
<th>dim2</th>
<th>dim3</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABZAC</td>
<td>2.39</td>
<td>0.011</td>
<td>-1.595</td>
</tr>
<tr>
<td>AILLAS</td>
<td>-0.87</td>
<td>0.122</td>
<td>0.084</td>
</tr>
<tr>
<td>AMBARES</td>
<td>2.65</td>
<td>0.795</td>
<td>-1.098</td>
</tr>
<tr>
<td>AMBES</td>
<td>0.94</td>
<td>0.895</td>
<td>-1.164</td>
</tr>
<tr>
<td>ANDERNOS</td>
<td>1.19</td>
<td>-2.466</td>
<td>0.800</td>
</tr>
</tbody>
</table>
Outline

1. PCAmix
2. PCArot
3. MFAmix
Let us introduce

- $\mathbf{T}$ an orthonormal rotation matrix: $\mathbf{T}\mathbf{T}' = \mathbf{T}'\mathbf{T} = \mathbf{I}_k$
- $k$ is the number of dimensions in the rotation procedure

$\Rightarrow$ Rotate the loading matrix and the principal components so that groups of variables appear: having high loadings on the same component and negligible ones on the remaining components.

$\Rightarrow$ In PCAmix rotated squared loadings $\tilde{c}_{j\alpha}$ are correlations or correlation ratios for categorical variables.

$\Rightarrow$ The varimax function writes:

$$f(\mathbf{T}) = \sum_{\alpha=1}^{k} \sum_{j=1}^{p} (\tilde{c}_{j\alpha})^2 - \frac{1}{p} \sum_{\alpha=1}^{k} \left( \sum_{j=1}^{p} \tilde{c}_{j\alpha} \right)^2.$$  \hspace{1cm} (2)
Find the optimal rotation matrix $T$.

The varimax rotation problem is formulated as

$$\max_T \{f(T)|TT' = T'T = I_k\}, \quad (3)$$

⇒ An iterative procedure based on successive planar rotations.
⇒ Direct solution for the optimal angle of rotation (Chavent & al. 2012).
⇒ Reduces to the Kaiser (1958) for numerical data.
⇒ Performs rotation in MCA for categorical data.
The R function

```
obj <- PCAmix(X.quanti=X1, X.quali=X2, rename.level=TRUE, graph=FALSE)
rot <- PCArot(obj, dim=3)
rot
```

```
## Call:
## PCArot(obj = obj, dim = 3)
##
## Method = rotation after Principal Component of mixed data (PCAmix)
## number of iterations: 4
##
## name description
## [1,] "$eig" variances of the 'ndim' first dimensions after rotation"
## [2,] "$ind" results for the individuals after rotation (coord)"
## [3,] "$quanti" results for the quantitative variables (coord) after rotation"
## [4,] "$levels" results for the levels of the qualitative variables (coord) after rotation"
## [5,] "$quali" results for the qualitative variables (coord) after rotation"
## [6,] "$sqload" squared loadings after rotation"
## [7,] "$coef" coef of the linear combinations defining the rotated PC"
## [8,] "$theta" angle of rotation if 'dim'=2"
## [9,] "$T" matrix of rotation"
```
The method `plot()`

```r
plot(obj, choice="sqload", coloring.var="type", leg=TRUE, axes=c(1,3), posleg="topright", main="Squared loadings before rotation")
```

```r
plot(rot, choice="sqload", coloring.var="type", leg=TRUE, axes=c(1,3), posleg="topright", main="Squared loadings after rotation")
```

Squared loadings before rotation

Squared loadings after rotation
The method `plot()`

```r
plot(obj, choice="ind", label=FALSE, axes=c(1,3), main="Observations before rotation")
```

```r
plot(rot, choice="ind", label=FALSE, axes=c(1,3), main="Observations after rotation")
```

Prediction of scores of new observations on the rotated principal components
The method `predict()`

Coefficients of the PC found on the learning set (without the 5 first cities)

```r
obj2 <- PCAmix(X1[-test,],X2[-test,],graph=FALSE)
rot2 <- PCArot(obj2,dim=3)
data.frame(rot2$coef)
```

<table>
<thead>
<tr>
<th></th>
<th>dim1.rot</th>
<th>dim2.rot</th>
<th>dim3.rot</th>
</tr>
</thead>
<tbody>
<tr>
<td>const</td>
<td>2.00163</td>
<td>-9.3e+00</td>
<td>1.3741</td>
</tr>
<tr>
<td>density</td>
<td>-0.00094</td>
<td>3.9e-05</td>
<td>0.0022</td>
</tr>
<tr>
<td>primaryres</td>
<td>0.00688</td>
<td>9.6e-02</td>
<td>-0.0014</td>
</tr>
<tr>
<td>owners</td>
<td>-0.03313</td>
<td>1.4e-02</td>
<td>-0.0228</td>
</tr>
<tr>
<td>inf 90%</td>
<td>1.23247</td>
<td>-2.2e-01</td>
<td>-0.1816</td>
</tr>
<tr>
<td>sup 90%</td>
<td>-0.31027</td>
<td>5.5e-02</td>
<td>0.0457</td>
</tr>
<tr>
<td>inf 5%</td>
<td>-0.44031</td>
<td>-1.2e-01</td>
<td>0.1958</td>
</tr>
<tr>
<td>sup 5%</td>
<td>1.68985</td>
<td>4.6e-01</td>
<td>-0.7514</td>
</tr>
</tbody>
</table>
```

Scores of the 5 first cities on the rotated principal components

```r
predict(rot2,X1[test,],X2[test,])
```

<table>
<thead>
<tr>
<th></th>
<th>dim1.rot</th>
<th>dim2.rot</th>
<th>dim3.rot</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABZAC</td>
<td>3.28</td>
<td>0.36</td>
<td>-0.865</td>
</tr>
<tr>
<td>AILLAS</td>
<td>-0.72</td>
<td>0.12</td>
<td>-0.223</td>
</tr>
<tr>
<td>AMBARES</td>
<td>2.90</td>
<td>0.98</td>
<td>-0.037</td>
</tr>
<tr>
<td>AMBES</td>
<td>1.73</td>
<td>1.15</td>
<td>-0.764</td>
</tr>
<tr>
<td>ANDERNOS</td>
<td>0.33</td>
<td>-2.64</td>
<td>0.868</td>
</tr>
</tbody>
</table>
Outline

1. PCAmix
2. PCArot
3. MFAmix
Multiply Factor Analysis for mixed data

⇒ Analyze a set of observations described by several groups of variables.

⇒ Make all the groups of variables comparable in the PCA analysis by introducing weights: the weight of a variable is the inverse of the variance of the first principal component of its group.

⇒ In the function MFA in the package FactoMineR, the nature of the variables (categorical or numerical) can vary from one group to another, but the variables should be of the same type within a given group.

⇒ MFAmix is able to handle mixed data within a group of variables.
The R function

```r
dat <- cbind(gironde$employment, gironde$housing, gironde$services, gironde$environment)
# definition of the groups of variables
class.var <- c(rep(1,9), rep(2,5), rep(3,9), rep(4,4))
# names of the groups of variables
names <- c("employment", "housing", "services", "environment")
# Perform MFAmix
obj3 <- MFAmix(data=dat, groups=class.var, name.groups=names, ndim=3, rename.level=TRUE, graph=FALSE)
```

## Results of the Multiple Factor Analysis for mixed data (MFAmix)

- **Results**: The analysis was performed on 542 individuals, described by 27 variables.

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;$eig&quot;</td>
<td>eigenvalues</td>
</tr>
<tr>
<td>&quot;$eig.separate&quot;</td>
<td>eigenvalues of the separate analyses</td>
</tr>
<tr>
<td>&quot;$separate.analyses&quot;</td>
<td>separate analyses for each group of variables</td>
</tr>
<tr>
<td>&quot;$groups&quot;</td>
<td>results for all the groups</td>
</tr>
<tr>
<td>&quot;$partial.axes&quot;</td>
<td>results for the partial axes</td>
</tr>
<tr>
<td>&quot;$ind&quot;</td>
<td>results for the individuals</td>
</tr>
<tr>
<td>&quot;$ind.partial&quot;</td>
<td>results for the partial individuals</td>
</tr>
<tr>
<td>&quot;$quanti&quot;</td>
<td>results for the quantitative variables</td>
</tr>
<tr>
<td>&quot;$levels&quot;</td>
<td>results for the levels of the qualitative variables</td>
</tr>
<tr>
<td>&quot;$quali&quot;</td>
<td>results for the qualitative variables</td>
</tr>
<tr>
<td>&quot;$sqload&quot;</td>
<td>squared loadings</td>
</tr>
<tr>
<td>&quot;$global.pca&quot;</td>
<td>results for the global PCA</td>
</tr>
</tbody>
</table>
Some graphical output

(a) Correlation circle

(b) Partial axes

(c) Groups representation

(d) Partial observations
Other packages working with mixed data

⇒ **ClustOfVar** for the clustering of variables.

⇒ **divclust** for the divisive and monothetic clustering of observations.

⇒ **ClustGeo** for the clustering with geographical constraints (very soon available for mixed data).
Some references


Chavent, M., Kuentz, V., Saracco, J. (2012), Orthogonal Rotation in PCAMIX. *Advances in Data Analysis and Classification* 6, 131-146.


THANK YOU!