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
University of Pau and de l'Adour, France

Multivariate analysis of a mixture of numerical and categorical data

Three main functions:

- Function `PCAmix` for principal component analysis (PCA) of mixed data.
  $\hookrightarrow$ Includes standard PCA and MCA (multiple component analysis) as special cases.
- Function `PCArot` for orthogonal rotation in PCAmix.
  $\hookrightarrow$ 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

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**.
  ↪ Implements the method of Hill & Smith (1976).
- 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

```
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).

  $\hookrightarrow$ Four datatables

A mixed data type example

The datatable housing is mixed.

↪ 3 numerical and 2 categorical variables.

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

```
head(housing)
```

```
##          density primaryres owners   houses council
## ABZAC        132          89     64  inf 90%  sup 5%
## AILLAS        21          88     77  sup 90%  inf 5%
## AMBARES      532          95     66  inf 90%  sup 5%
## AMBES        101          94     67  sup 90%  sup 5%
## ANDERNOS     552          62     72  inf 90%  inf 5%
## ANGLADE       64          81     81  sup 90%  inf 5%
```

Two data sets:

↪ a numerical data matrix $\mathbf{X}_1$ of dimension 542×3.

↪ a categorical data matrix $\mathbf{X}_2$ of dimension 542×2.

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

```
head(X1)
```

```
##           density primaryres owners
## ABZAC         132          89     64
## AILLAS         21          88     77
## AMBARES       532          95     66
## AMBES         101          94     67
## ANDERNOS      552          62     72
## ANGLADE        64          81     81
```

```
head(X2)
```

```
##              houses council
## ABZAC     inf 90%  sup 5%
## AILLAS    sup 90%  inf 5%
## AMBARES   inf 90%  sup 5%
## AMBES     sup 90%  sup 5%
## ANDERNOS  inf 90%  inf 5%
## ANGLADE   sup 90%  inf 5%
```

The PCAmix method

An simple algorithm in three main steps

1. Preprocessing step.
2. GSVD (Generalized Singular Value Decomposition) step.
3. Scores and loadings processing step.

Some notations:

- Let $\mathbf{X}_1$ be a $n \times p_1$ **numerical** data matrix.
- Let $\mathbf{X}_2$ be a $n \times p_2$ **categorical** data matrix.
- Let m be the total number of levels of the categorical variables.

Preprocessing step

1. Build a **numerical data matrix** $\mathbf{Z} = (\mathbf{Z}_1|\mathbf{Z}_2)$ of dimension $n \times (p_1 + m)$ with:
   - $\hookrightarrow$ $\mathbf{Z}_1$ the standardized version of the matrix $\mathbf{X}_1$.
   - $\hookrightarrow$ $\mathbf{Z}_2$ the centered indicator matrix of the levels of $\mathbf{X}_2$.
2. Build the diagonal matrix $\mathbf{N}$ of the weights of the rows.
   - $\hookrightarrow$ The n rows are weighted by $\frac{1}{n}$.
3. Build the diagonal matrix $\mathbf{M}$ of the weights of the columns.
   - $\hookrightarrow$ The p_1 first columns are weighted by 1.
   - $\hookrightarrow$ The m last columns are weighted by $\frac{n}{n_s}$, with n_s the number of observations with level s.

$\hookrightarrow$ The total variance is $p_1 + m - p_2$.

GSVD step

The GSVD (Generalized Singular Value Decomposition) of $\mathbf{Z}$ with the metrics $\mathbf{N}$ and $\mathbf{M}$ gives the decompositon

$$\mathbf{Z} = \mathbf{U}\mathbf{D}\mathbf{V}^T \tag{1}$$

where

- $\mathbf{D} = \text{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_r})$ is the $r \times r$ diagonal matrix of the singular values of $\mathbf{Z}\mathbf{M}\mathbf{Z}^T\mathbf{N}$ and $\mathbf{Z}^T\mathbf{N}\mathbf{Z}\mathbf{M}$, and r denotes the rank of $\mathbf{Z}$;
- $\mathbf{U}$ is the $n \times r$ matrix of the first r eigenvectors of $\mathbf{Z}\mathbf{M}\mathbf{Z}^T\mathbf{N}$ such that $\mathbf{U}^T\mathbf{N}\mathbf{U} = \mathbb{I}_r$;
- $\mathbf{V}$ is the $p \times r$ matrix of the first r eigenvectors of $\mathbf{Z}^T\mathbf{N}\mathbf{Z}\mathbf{M}$ such that $\mathbf{V}^T\mathbf{M}\mathbf{V} = \mathbb{I}_r$.

Scores and loadings processing step

1. The principal coordinates of the rows are computed as:

$$\mathbf{F} = \mathbf{UD}.$$

$\hookrightarrow$ principal component scores of the observations,

2. The principal coordinates of the colums are computed as:

$$\mathbf{A} = \mathbf{MVD}.$$

$\hookrightarrow$ loadings of the numerical variables and of the levels of the caregorical variables,

$\hookrightarrow$ in standard PCA: $\mathbf{A} = \mathbf{VD}$.

The R function

```
args(PCAmix)


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

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


##
## Call:
## PCAmix(X.quanti = X1, X.quali = X2, ndim = 2, graph = FALSE)
##
## Method = Principal Component of mixed data (PCAmix)
##
##
##      name      description
## [1,] "$eig"    "eigenvalues of the principal components (PC) "
## [2,] "$ind"    "results for the individuals (coord,contrib,cos2)"
## [3,] "$quanti" "results for the quantitative variables (coord,contrib,cos2)"
## [4,] "$levels" "results for the levels of the qualitative variables (coord,contrib,cos2)"
## [5,] "$quali"  "results for the qualitative variables (contrib,relative contrib)"
## [6,] "$sqload" "squared loadings"
## [7,] "$coef"   "coef of the linear combinations defining the PC"
```

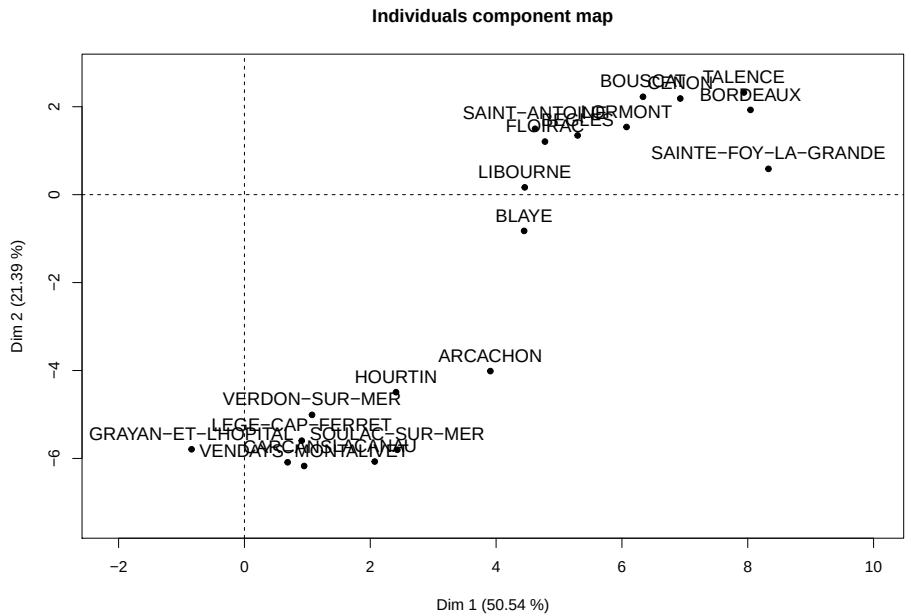
Scores of the observations

```
obj <- PCAmix(X.quanti=X1,X.quali=X2,
              ndim=2,rename.level = TRUE,graph=FALSE)
```

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

```
##           dim 1  dim 2
## ABZAC      2.36  0.024
## AILLAS    -0.88  0.123
## AMBARES    2.62  0.800
## AMBES      0.93  0.919
## ANDERNOS   1.18 -2.481
## ANGLADE   -1.01 -0.424
```

```
plot(obj,choice="ind",lim.contrib.plot = 1,cex=1.2)
```

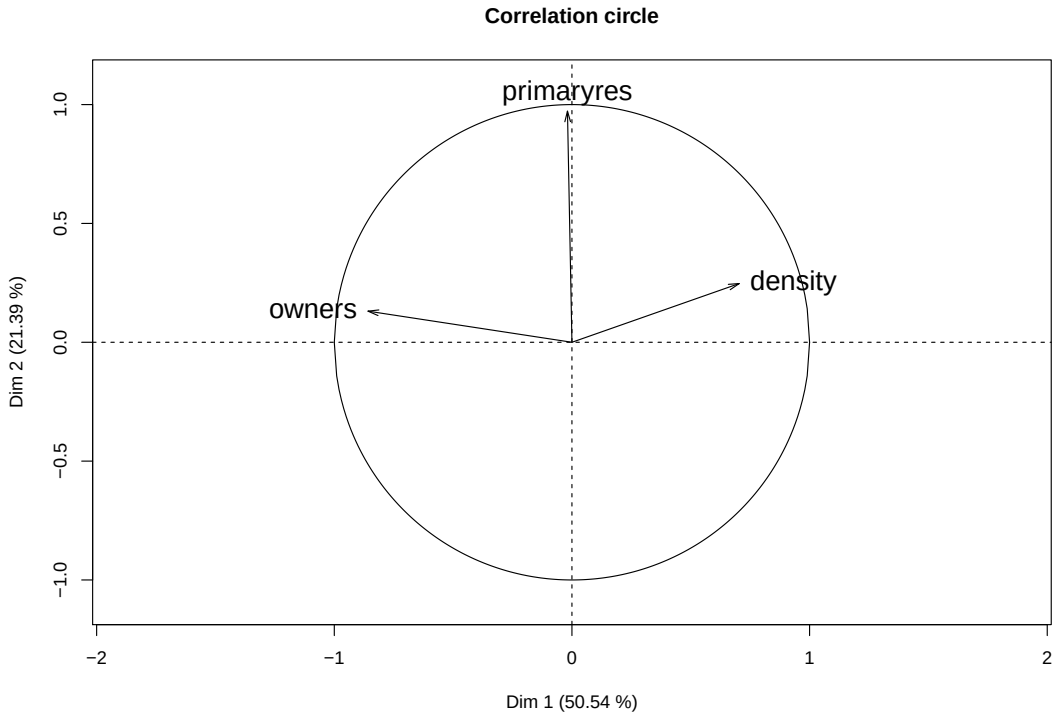


Scores of the numerical variables

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

```
##              dim 1 dim 2
## density      0.704  0.25
## primaryres -0.019  0.97
## owners     -0.858  0.13
```

```
#Component map with factor scores of the numerical columns
plot(obj,choice="cor",cex=1.5)
```



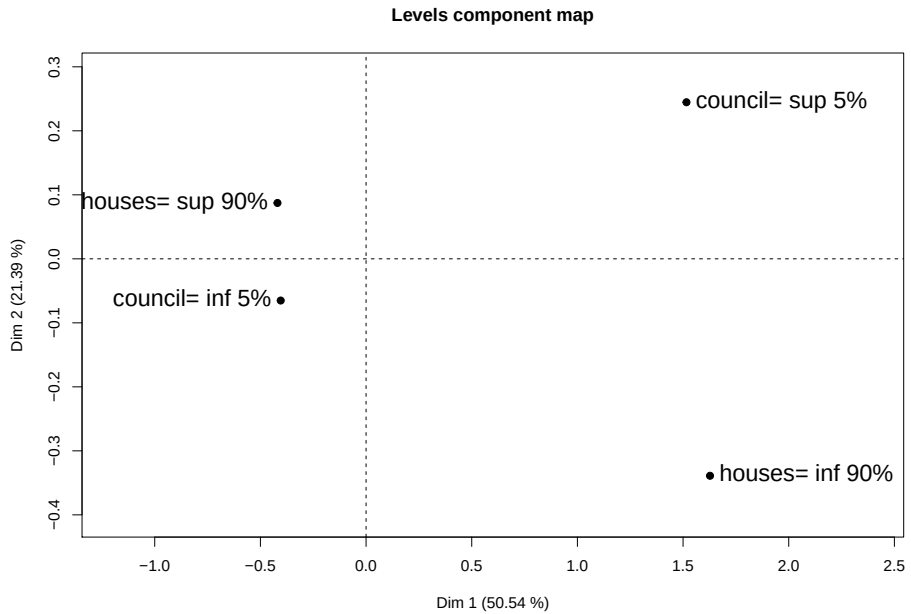
↪ The (non standardized) loadings are correlations.

Scores of the levels of the categorical variables

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

```
##                  dim 1  dim 2
## houses= inf 90%  1.63 -0.339
## houses= sup 90% -0.42  0.087
## council= inf 5% -0.40 -0.065
## council= sup 5%  1.52  0.245
```

```
plot(obj,choice="levels",cex=1.5,xlim=c(-1,2))
```



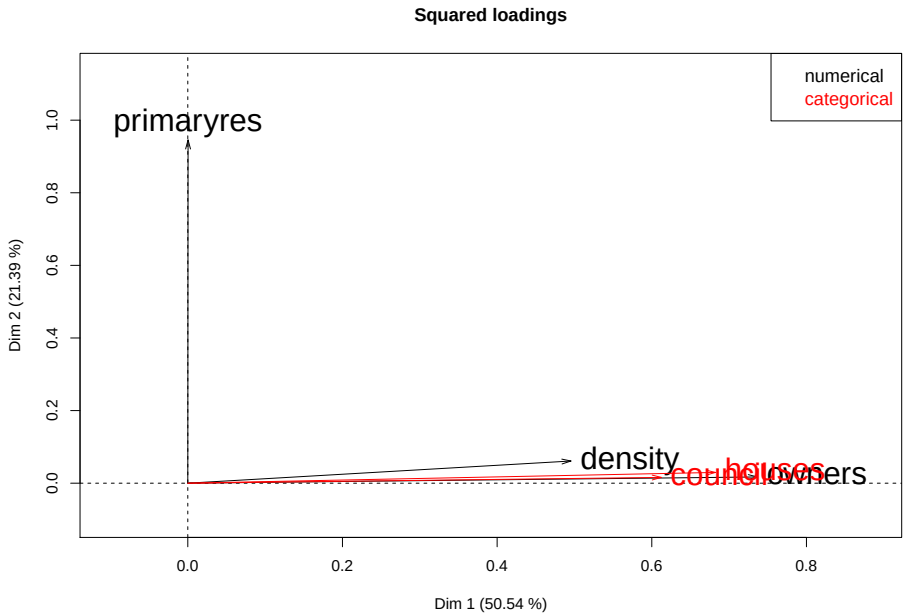
↪ The barycentric property is still verified.

Contributions of the variables

```
#contributions of the variables
head(obj$sqload)
```

```
##                dim 1 dim 2
## density      0.49550 0.061
## primaryres   0.00035 0.946
## owners       0.73651 0.017
## houses       0.68226 0.030
## council      0.61226 0.016
```

```
plot(obj,choice="sqload",coloring.var=TRUE,
     cex=2,posleg = "topright",cex.leg=1.2)
```



Squared loadings

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

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

Principal components prediction

Each principal component $\mathbf{f}_\alpha$ writes as a linear combination of the columns of $\mathbf{X} = (\mathbf{X}_1|\mathbf{G})$ where $\mathbf{X}_1$ is the numerical data matrix and $\mathbf{G}$ is the indicator matrix of the levels of the matrix $\mathbf{X}_2$:

$$\mathbf{f}_\alpha = \beta_0 + \sum_{j=1}^{p_1+m} \beta_j \mathbf{x}_j$$

with:

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

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

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

The method predict()

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

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

```
##                dim1     dim2    dim3
##             3.64214 -9.02951  1.5950
## density     0.00086  0.00048  0.0015
## primaryres -0.00129  0.09355 -0.0179
## owners     -0.05065  0.01207 -0.0047
##  inf 90%    1.03579 -0.32092 -0.4617
##  sup 90%   -0.26076  0.08079  0.1162
##  inf 5%    -0.25129 -0.05706  0.2704
##  sup 5%     0.96441  0.21898 -1.0379
```

Scores of the 5 first cities on the principal components

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

```
##           dim1   dim2   dim3
## ABZAC     2.39  0.011 -1.595
## AILLAS   -0.87  0.122  0.084
## AMBARES   2.65  0.795 -1.098
## AMBES     0.94  0.895 -1.164
## ANDERNOS  1.19 -2.466  0.800
```

Varimax type rotation in PCAmix

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 axes of PCAmix so that variables appear more clearly associated with the principal components.

$\Rightarrow$ Association between variables and principal components are measured by the squared loadings : rotated squared loadings $\tilde{c}_{j\alpha}$ are squared correlations or correlation ratios.

$\Rightarrow$ The varimax function writes:

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

Find the optimal rotation matrix $\mathbf{T}$

The varimax rotation problem is formulated as

$$\max_{\mathbf{T}} \quad \{f(\mathbf{T}) | \mathbf{T}\mathbf{T}' = \mathbf{T}'\mathbf{T} = \mathbf{I}_k\}, \tag{3}$$

$\Rightarrow$ An iterative procedure based on successive planar rotations.

$\Rightarrow$ Direct solution for the optimal angle of rotation (Chavent & al. 2012).

$\Rightarrow$ Reduces to the Kaiser (1958) for numerical data.

$\Rightarrow$ Performs rotation in MCA for categorical data.

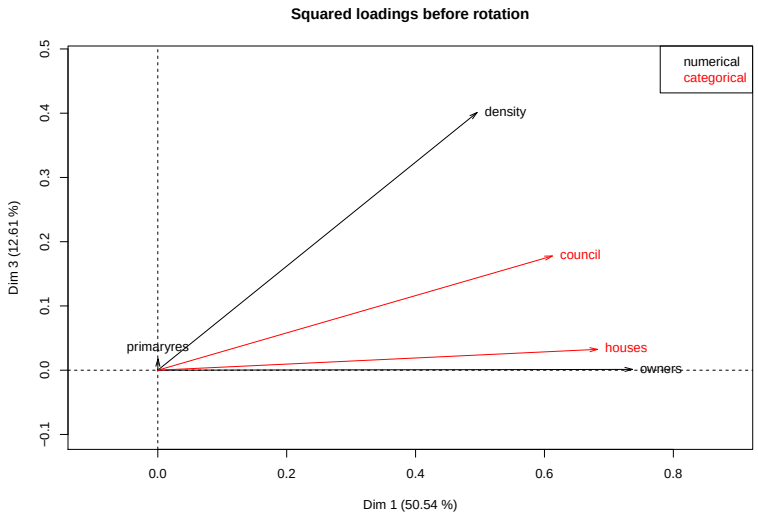
The R function

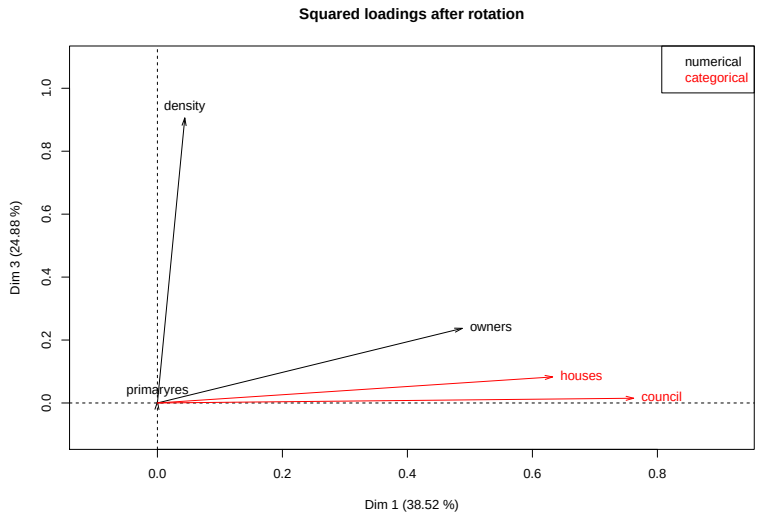
```
obj <- PCAmix(X.quanti=X1, X.quali=X2, rename.level=TRUE, graph=FALSE)
rot <- PCArot(obj,dim=3,graph=FALSE)
rot


##
## Call:
## PCArot(obj = obj, dim = 3, graph = FALSE)
##
## 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()

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

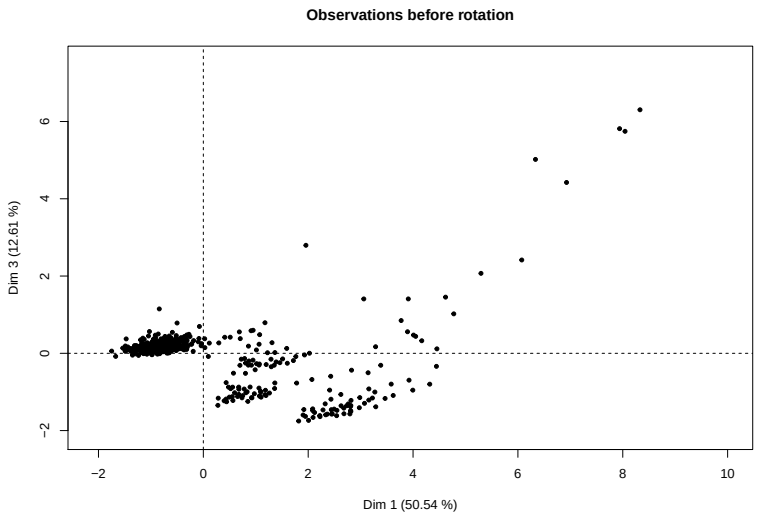
```
plot(rot,choice="sqload",coloring.var=TRUE,
  leg=TRUE,axes=c(1,3),posleg="topright",
  main="Squared loadings after rotation")
```



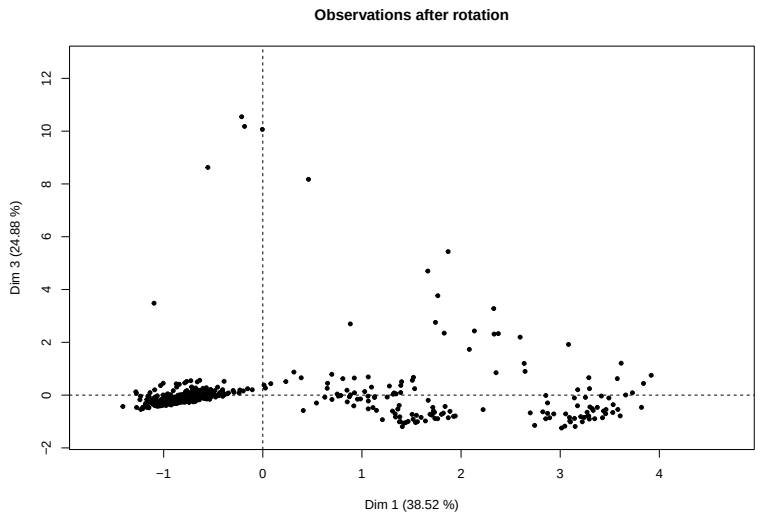


The method plot()

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




```
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)

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

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

Scores of the 5 first cities on the rotated principal components

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

```
##          dim1.rot dim2.rot dim3.rot
## ABZAC        3.28     0.36   -0.865
## AILLAS      -0.72     0.12   -0.223
## AMBARES      2.90     0.98   -0.037
## AMBES        1.73     1.15   -0.764
## ANDERNOS     0.33    -2.64    0.868
```

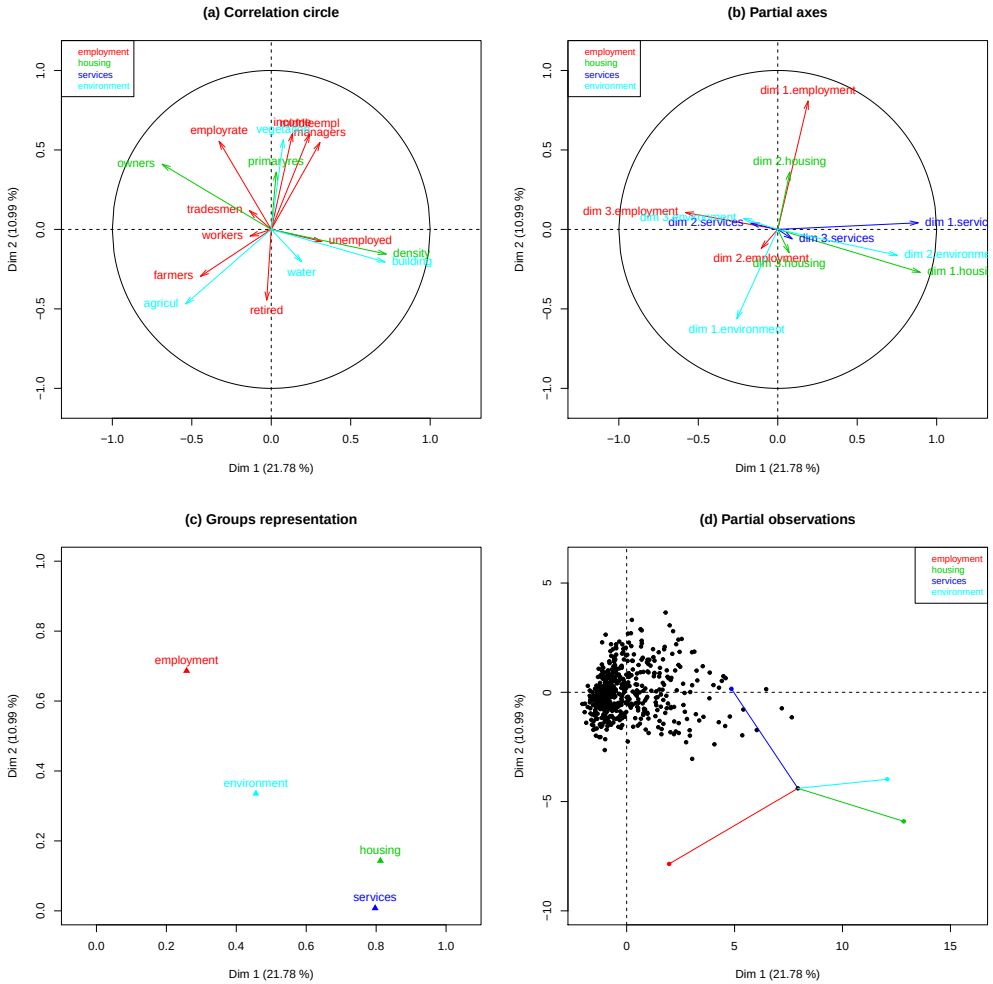
Multiple Factor Analysis for mixed data

- ⇒ Analyze a set of observations described by several groups of variables.
- ⇒ Make the importance of the groups 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

```
dat<-cbind(gironde$employment,gironde$housing, gironde$services,gironde$environment)
#definition of the groups of variables
groupind<-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=groupind,name.groups=names,ndim=3,rename.level=TRUE,graph=FALSE)
obj3


## **Results of the Multiple Factor Analysis for mixed data (MFAmix)**
## The analysis was performed on 542 individuals, described by 27 variables
## *Results are available in the following objects :
##
##       name                 description
##  [1,] "$eig"               "eigenvalues"
##  [2,] "$eig.separate"      "eigenvalues of the separate analyses"
##  [3,] "$separate.analyses" "separate analyses for each group of variables"
##  [4,] "$groups"            "results for all the groups"
##  [5,] "$partial.axes"      "results for the partial axes"
##  [6,] "$ind"               "results for the individuals"
##  [7,] "$ind.partial"       "results for the partial individuals"
##  [8,] "$quanti"            "results for the quantitative variables"
##  [9,] "$levels"            "results for the levels of the qualitative variables"
## [10,] "$quali"             "results for the qualitative variables"
## [11,] "$sqload"            "squared loadings"
## [12,] "$listvar.group"     "list of variables in each group"
## [13,] "$global.pca"        "results for the global PCA"
```

Some graphical output



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


Beaton, D., Chin Fatt, C. R., Abdi, H. (2014). An ExPosition of multivariate analysis with the singular value decomposition in R. *Computational Statistics & Data Analysis*, 72, 176-189.


Chavent, M., Kuentz, V., Liquet B., Saracco, J. (2012), ClustOfVar: An R Package for the Clustering of Variables. *Journal of Statistical Software* 50, 1-16.


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


Chavent, M., Kuentz, V., Saracco, J. (2012), Multivariate analysis of mixed type data: The PCAmixdata R package. *arXiv:1411.4911v1*.


Dray, S., Dufour, A., 2007. The ade4 package: implementing the duality diagram for ecologists. *Journal of Statistical Software* 22 (4), 120.


Lê, S., Josse, J., Husson, F., et al. (2008). Factominer: an R package for multivariate analysis. *Journal of Statistical Software* 25 (1), 118.

THANK YOU !