MFAg: a R package for carrying out the multiple factor analysis

Paulo César OSSANI[1]

Marcelo Ângelo CIRILLO[2]

[1] Doutorando em Estatística e Experimentação Agropecuária pela Universidade Federal de Lavras (UFLA). ossanipc@hotmail.com
[2] Prof. Adjunto IV do Depto. de Ciências Exatas da Universidade Federal de Lavras (UFLA) macufla@des.ufla

Recebido em: 22/08/2016 - Aprovado em: 10/11/2017 - Disponibilizado em: 30/12/2017

ABSTRACT:
In considering the study between groups of variables using a multivariate approach, the usual techniques are either limited or unviable to describe how distinct these groups are. The multiple factor analysis technique (MFA) enables a statistical analysis among groups, whether they are quantitative, categorical, frequency or even mixed. The core of the technique is a factor analysis applied to the set of variables in which each group of variables is balanced, leading to a representation of the observations and variables. Thus, this package carries out the multiple factor analysis technique, generating easy results of being surveyed, as in a PCA (Principal Components Analysis), in addition to statistical plots of easy understanding. It possesses also some other functions for multivariate analysis, which are used by the MFA technique.
Keywords: Multivariate analysis. Multiple factor analysis. Principal components analysis. Variable groups. Mixed data.

MFAg: um pacote R para executar a análise de múltiplos fatores

RESUMO:
Ao considerar o estudo entre grupos de variáveis, usando uma abordagem multivariada, às técnicas usuais são limitadas ou inviáveis para descrever quão distintos estes grupos são. A técnica de análise de múltiplos fatores (MFA), possibilita uma análise estatística entre grupos, sejam eles quantitativos, categóricos, frequência ou mesmo misto. O cerne da técnica é uma análise de fator aplicada ao conjunto de variáveis na qual cada grupo de variáveis é balanceado, conduzindo a uma representação das observações e variáveis. Assim este pacote realiza à técnica análise de múltiplos fatores, gerando resultados fáceis de serem analisados, como num PCA (Principal Components Analysis), alem de gráficos estatísticos de fácil entendimento. Possui também algumas outras funções para análise multivariada, que são usadas pela técnica MFA.
Palavras Chave: Análise multivariada. Análise de múltiplos fatores. Análise de componentes principais. Grupo de variáveis. Dados mixtos.

1. Introduction

In several research works, the need to establish the comparison among groups of variables appears. We can cite agricultural research, where a number of traits in groups of animals of the same species are sought to compare or not; a sensorial study of foods, where the interest is to know if distinct groups of individuals agree with some attributes of a certain food according to Pagès (2004), among other studies. So, a statistical tool which can furnish us comparable responses among groups of variables is important in all the social sectors.

Out of the several techniques proposed to analyze data, the multiple factor analysis characterizes for allowing groups of variables (columns) with different sizes and of distinct nature, which can be quantitative, categorical or of frequency, defined in the some set of observations (rows), as seen in Escofier and Pagès (1982, 1990, 2008) and Bécue-Bertaut

and Pagès (2004). This method can be very useful for analyzing of studies in which several groups of variables can be identified or for the studies in which the same questions are asked in different intervals of time, according to Abdi, Williams and Valentin (2013).

The advantage provided by this method consists of generating results with analysis similar to the PCA, enabling also the viewing in a space of two or three dimensions, the groups of variables (each group being represented by a point), the variables, main axes and further the observations, according to Abdi; Williams (2010).

In virtue of what was mentioned, the MFAg package in Ossani and Cirillo (2016) was created to carry out the MFA technique for sets of quantitative, categorical, frequency data and further set of mixed data.

This paper is organized in the following way: at first, we make a brief remark about the MFA technique. Next, we report the parameters of the functions of the MFAg package. At last, a pictorial example of the package is presented.

2. MFA technique

The core of the MFA technique is a factor analysis applied to the whole set of variables in which each group of variables is balanced, leading to a representation of the observations and variables, like in any factor analysis. Due to the balancing, this factor analysis can be interpreted as a canonical analysis, according to Escofier e Pagès (1994).

It is worth stressing that in the multiple factor analysis the number of variables in each group can differ and the nature of the variables (quantitative, categorical and frequency) can range from a group to the other, but the variables must be in the same nature in the group given, according to Abdi and Valentin (2007), Escofier and Pagès (2008) and Abdi and Williams (2010).

Missing data are allowed in the case of categorical variables, so the observations where there is no k category, will have 0 (zero) for all the indicator variables associated to the k category according to Escofier e Pagès (1994).

3. Mixture of quantitative, categorical and frequency data

In a great deal of studies, the statistical observations are described simultaneously by variables belonging to at least two types: quantitative, categorical and frequency, which generates the problem of mixed data.

The mixed data problem was already studied in Gower (1971), the first to propose a solution to balance quantitative and categorical variables, independently of its type. Specific distances are used for categorical and quantitative variables; the range of variation of each distance is

standardized to 1 before the clustering in the global distance, according to Bécue-Bertaut and Pagès (2008).

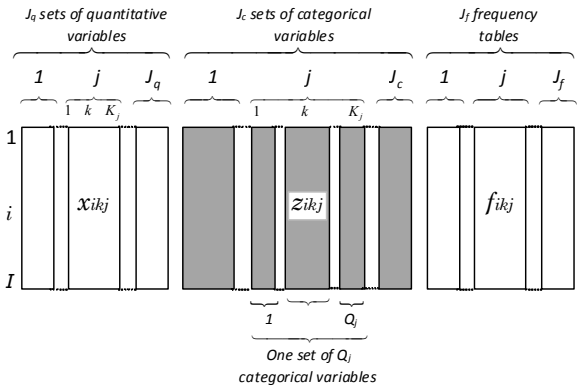
Also one can seed in Greenacre and Blasius (2006) a study combining the MFA method to deal with several tables in which the quantitative and categorical variables are juxtaposed.

According to Bécue-Bertaut and Pagès (2008) to analyze the groups described for these types of data, the starting point is the definition of a global distance combining the distances (named separate distances) of PCA emission (in the case of quantitative sets), MCA (in the case of categorical sets) and CA (in the case of frequency sets).

4. Notation: A single table

The I observations are described by J sets of variables: J_q sets of quantitative variables, J_c sets of categorical variables and J_f sets of frequency variables, in which $J=J_q+J_c+J_f$, Figure 1.

Figure 1: Global table with sets of juxtaposed quantitative, categorical and frequency variables.



Fonte: Bécue-Bertaut e Pagès (2008)

The symbols I, J, J_q, J_c, J_f, K ou K_j refer to the set and its cardinality.

Whatever may be the type of set, the letter j refers to a set, the letter k refers to a column and K_j is the number of columns in the set j. $K = \sum_{j \in J} K_j$ is the number of columns in all the sets. The table I x K_j is associated with each set j. The tables J together compose a multiple or global table I x K. For a quantitative or frequency set j, K_j is both the number of columns and the number of variables. For a categorical set j, with variables Q_j, K_j is both the number of columns and the number of categories summed through the variables Q_j. This type of set is represented by a table of indicator variables in which the column k is associated with the category k, according to Bécue-Bertaut and Pagès (2008).

At the crossing of the row i and column k (which belong to table j), we have:

- if j is quantitative set, the value x_{ikj} of the variable l k for observation i;
- if j is a categorical set, $z_{ikj} = 1$ if i belongs to the category k and 0 otherwise;
- if j is a set of frequencies, the proportion f_{ikj}, is calculated as the ration between the number of occurrence of events (belonging to the set j) for observation of i and the general total of the table which joins together all the frequency tables J_f; so: $\sum_{j \in J_f} \sum_{k \in K_j} \sum_{j \in I} f_{ikj} = 1$.

We used:

- $f_{i.j} = \sum_{k \in K_j} f_{ikj}$ and $f_{.jk} = \sum_{j \in I} f_{ikj}$ to denote the margins row and columns of the frequency table j as a subtable of the global table;
- $f_{i..} = \sum_{j \in J_f} \sum_{k \in K_j} f_{ikj}$ to denote the margin of row of the table collecting all the frequency tables J_f.

5. MFA: a geometrical approach to balance the influence of the different sets

According to Escofier and Pagès (1994, 2008) and cited by Bécue-Bertaut and Pagès (2008), the MFA method deals with different sets of quantitative variables and in order to balance the influence of those different sets, this adopts a geometric approach, considering the cloud associated with each set of variables and standardize the inertia of each cloud over the first main axis to 1. Technically, this property is obtained by dividing the weight of the columns belonging to the set j by λ_1^j, the first eigenvalue of the separate analysis of set j. so, the contribution of any set with the global distance depends on the real dimension of the cloud: a cloud with several important orthogonal inertias, will have a greater influence than an one-dimensional cloud. The basis of the
MFA method can be seen as a weighted PCA applied to multiple table. The weight p_i is denoted, generally uniform, ascribed to observation i.

For the categorical sets in the MFA method according to Escofier and Pagès (2008), Bécue-Bertaut and Pagès (2008) and Pagès (2002), the starting point is making use of the equivalence between MCA and the weighted PCA. The results of the MCA can be obtained through the carrying out of the PCA:

- Applied to the table with the general term $\left(z_{ikj} - w_{kj}\right)/ w_{kj}$, in which $z_{ikj} = 1$ if i belongs to the category k and 0 otherwise and $w_{kj} = \sum_{i \in I} p_i \cdot z_{ikj}$ (notice: $\sum_{k \in K_j} w_{kj} = Q_j$);
- Give the weight w_{kj} / Q_j to the column k of the set j;
- Giving p_i weight to the row i.

The distances between the observations and between the columns induced by the PCA are equal to the distances generally considered in MCA. In particular, according to Bécue-Bertaut and Pagès (2008), the square of the distance between the observations of i and i l is given by:

$$d^2(i,l) = \sum_{k \in K_j} \frac{Q_j}{w_{kj}} \left[\frac{p_i z_{ikj}}{p_i Q_j} - \frac{p_l z_{lkj}}{p_l Q_j} \right]^2 = \sum_{k \in K_j} \frac{1}{Q_j w_{kj}} \left[z_{ikj} - z_{lkj} \right]^2 .$$

According to Bécue-Bertaut and Pagès (2008), the introduction of frequency tables as sets of variables induces to a particular problem. The chief reference for such a table is CA, which imposes as weights of rows, the

coefficients of the margins of the rows. If the margins of the rows of the frequency tables are equal (or proportional), the MFA method can be applied to this type of data in compliance with Abdessemed and Escofier (1996). When the row margins are different, it becomes complicated for the fact that the weights of the rows range according to the set of variables. A solution found for this problem is called the MFACT (Multiple Factor Analysis for Contingency Tables) and can be seen in Bécue-Bertaut and Pagès (2004).

6. Mixed data

According to Bécue-Bertaut and Pagès (2008) the mixture of quantitative, categorical and frequency tables into a single analysis raises the problem of weighing units. Mixed data require that the value of the different variables in the global distance to be weighted. In the case of quantitative and categorical sets, the weights are generally uniform, fixed by the first eigenvalue, while in the case of frequency data; they are imposed by the margins of the tables. Fitting weights are necessary, since the weighing of the observations have to be identical in all the tables. A first solution consists in the adoption of the weights emitted from the frequency table, that is, $p_i = f_{i..}$. In this case, the MFA method is based upon a weighted non-normalized PCA performed in the multiple table presented in Table 1, using:

- $\{p_i = f_{i..}; i = 1, \ldots, I\}$ as row weights (and as metrics in the space of the columns);
- The initial weights of the columns (belonging to the set j), but divided by λ_1^j (first eigenvalue emitted from the PCA separated from table j) as weights of the columns (and as metrics in the space of the rows), this is, $\left(1/\lambda_1^j\right)$ in the case of a quantitative set, $\left(\left(w_{kj}/Q_j\right)/\lambda_1^j\right)$ in the case of a categorical set $\left(f_{.kj}/\lambda_1^j\right)$ in the case of a frequency set.

The choice of these weights of the observations has an effect on:

- quantitative sets through the calculation of the mean and standard deviation of each variable;
- the categorical sets through the calculation of the coefficients of w_{kj};
- the frequency sets through the use of $f_{i..}$ as weights of linhas.

Table1: Table with the appropriate transformations.

Observa-tion	Variable k in the quantita-tive set j	Variable k (indicator) In the categorical set j	Variable k in the frequency set j
1			
i	$\frac{x_{ikj} - \bar{x}_{kj}}{s_{kj}}$	$\frac{z_{ikj} - w_{kj}}{w_{kj}}$	$\frac{f_{ikj} - \left(f_{i.j}/f_{..j}\right)\cdot f}{p_i \cdot f_{kj}}$
I			
Weight of the columns	$\frac{1}{\lambda_1^j}$	$\frac{w_{kj}}{Q_j \lambda_1^j}$	$\frac{f_{.kj}}{\lambda_1^j}$

Here the weight of the observations i is given by: p_i

Another solution consists of adopting the weights emitted from the categorical and quantitative tables. Again the MFA method is equivalent to a non-normalized weighted PCA performed in the table presented in Table 1, now using $\{p_i = 1/I, i = 1,\ldots,I\}$ as weights of the rows (and as a metrics in the space of the columns) and the weights of the columns used in the same Table 1. In this case, the contribution of the quantitative and categorical sets to the global distance matches exactly, except by the weight $1/\lambda_1^j$ (for the distances emitted by the PCA or MCA). However, these unit weights modify the contribution of the frequency tables to the global distance, which no longer matches exactly to the distance emitted from the MFACT, according to Bécue-Bertaut and Pagès (2008).

According to Bécue-Bertaut and Pagès (2008) in the practice, the user will go to choose either one or another weighing unit, depending on the application characteristics. In the data sets with frequency tables emitted from open questions, the weights imposed by the frequency sets are adopted. This choice favorece the inquiridos with the longer answers, those generally use a richer and varying vocabulary. The MFA package opted for the weighing of the frequency tables when this one appears in a mixed set. So, the global matrix formed by all the concatenados and weighted data is given the name of matrix Z.

7. The R package MFAg

The package MFAg is a collection of functions written in R for the multiple factor analysis, Table 2. The package can be downloaded from any CRAN mirror site listed at http://CRAN.R-project.org/package=MFAg. The utilization of the functions and together with their entrance parameters and the values which return are described in help file.

Function	Description
MFA()	Main function. It performs the multiple factor analysis in group of variables. The groups can be of quantitative categorical and frequency or mixed variables.
GSVD()	It performs the decomposition of generalized singular value into a matrix of nxm order.
Plot.MFA()	It generates plots of the multiple factor analysis.
IM()	It transforms the categorical data into an indicator matrix.
NormData()	It normalizes the data per column or globally.

Table 2: Functions in MFAg.

In the package also is the following database: DataMix (set of mixed data), DataQuali (set of qualitative data) and DataQuan (set of quantitative data), useful in the examples inside the package.

As an illustration, the example follows: Let`s take simulated data among five coffee tasters of the cooperative A and five tasters of the cooperative B of a given region, having as a general purpose to obtain the topology of those cooperatives and to know if there are

differences among the groups of variables. The groups being formed by the variables:

Group 1 (Notes Coffee / Work): Means of the notes given to the coffees analyzed /Years of works as a taster.

Grupo 2 (Training / Dedication): Taster with technique background / Taster with exclusive commitment.

Grupo 3 (Coffee): Average frequency of the coffees classified as specialty / Average frequency of the coffees classified as commercial.

The analysis utilizing the function MFA(), which takes on the following parameters follows:

Data	Data to be analyzed.
Grupo	Number of columns for each group em ordem following the order of the datas of 'Data'.
TipoGrupo	"n" for numerical groups - default; "c" for categorical data; "f" for frequency data.
NomeGrupos	Names for each group.

And returns the following parameters:

MatrixG	Matrix with the sizes of each group;
MatrixNG	Matrix with the names of each group;
MatrixPLin	Matrix with the values used to balance the rows of the matrix Z;
MatrixPCol	Matrix with the values used to balance the columns of the matrix Z;
MatrixZ	Concatenated and balanced matrix;
MatrixA	Matrix with eigenvalues (variances);
MatrixU	Matrix U of the singular decomposition of the matrix Z;
MatrixV	Matrix V of the singular decomposition of the matrix Z;
MatrixF	Global matrix of the scores of the factors where the rows are the observations and the columns the components;
MatrixEFG	Matrix of the scores of the factors per group;
MatrixCCP	Correlation matrix of the principal components with the original variables;
MatrixEscVar	Matrix of the partial inertia /scores of the variables.

Utilizing the database "DataMix" of the package, we will have:

```
data(DataMix) # set of mixed data of the
cooperatives

Matriz = DataMix[,2:ncol(DataMix)]

rownames(Matriz) <- as.character(
t(DataMix[1:nrow(DataMix),1]))

GroupNames = c("Notes Coffee/Work",
"Training/Dedication", "Coffee")

MF <- MFA(Matriz, c(2,2,2),
c("n","c","f"), GroupNames) # perfoms MFA

print("Variances of the Principal
Component:"); round(MF$MatrixA,2)

        Eigenvalue % variance %
cumulative variance
Axis 1    2.25   56.02         56.02
Axis 2    1.03   25.77         81.79
Axis 3    0.46   11.53         93.32
Axis 4    0.21    5.16         98.48
Axis 5    0.06    1.52        100.00
```

As in a PCA, it is found that the two first axes are responsible for 81.79% of the total variation contained in the original variables.

```
print("Partial matrix of the inertia/
scores variables:")
```

```
round(MF$MatrixEscVar[,1:3],2)

                     Axis 1 Axis 2 Axis 3
Notes Coffee/Work      0.89   0.05   0.28
Training/Dedication    0.43   0.98   0.18
Coffee                 0.92   0.01   0.00
```

A high agreement between the discrimination of Group 1 (Notes Coffee/Work) and 3 (Coffee), with greater inertias on axis 1, in relation to Group 2 (Training/Dedication). The package also offers a graphical output through the function Plot.MFA(), which takes on the following parameters:

MFA Data of the function MFA

Titles Titles for the plots. If it is not defined, it takes on standard text.

PosLeg 1 for caption on the left upper corner;
2 for caption on the right upper corner - default;
3 for caption on the right lower corner;
4 for caption on the left lower corner.

BoxLeg "s" to place frame on the caption - default;
"n" does not place frame on the caption.

Color "s" for colored plots - default;
"n" for black and white plots.

NamArr "s" to put point names in the cloud around the centroid on the plot correspondent to the global analysis of the individuals and variables;
"n" Otherwise - default.

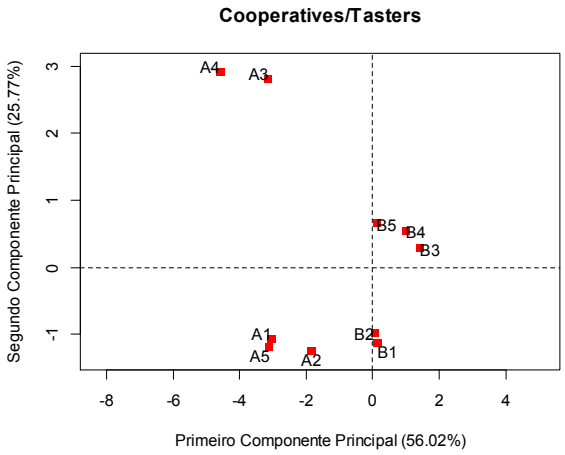
Let`s see the graphical output:

```
data(DataMix) # set of mixed data of the
cooperatives

Matriz = DataMix[,2:ncol(DataMix)]

rownames(Matriz) <- as.character(t(
DataMix[1:nrow(DataMix),1]))

GroupNames = c("Notes Coffee /
Work","Training / Dedication","Coffee")

MF <- MFA(Matriz, c(2,2,2), TipoGrupo =
c("n","c","f"), GroupNames) # performs
MFA

Titulos = c("Cooperatives/Tasters ",
"Observations/Variables", "Inertia Groups
Variables")

Plot.MFA(MF, Titulos, 2, "n", "s", "n") #
several screen graphics
```

The function Plot.MFA() returns automatically five plots, here only three were placed for exemplification.

From the obtained scores of MF$MatrixF, is generated in Figure 2.

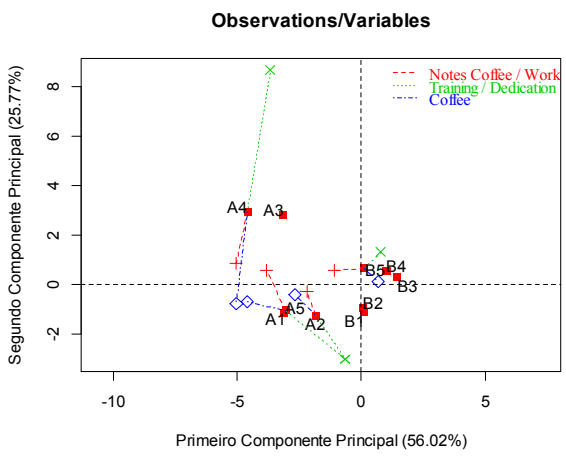
Figure 2: Plot of the analysis of the Cooperatives / Tasters.



One can interpret the first component as the opposition of the Cooperatives that have Tasters dedicated exclusively and those without. Those having located at the top of the component. On the second component there is clearly a separation of Cooperatives A and B.

The two first main components of the global analysis together with the projections of the Cooperatives/Tasters in the groups of variables, are shown in Figure 3. The scores of the groups of the variables are obtained from MF$MatrixEFG.

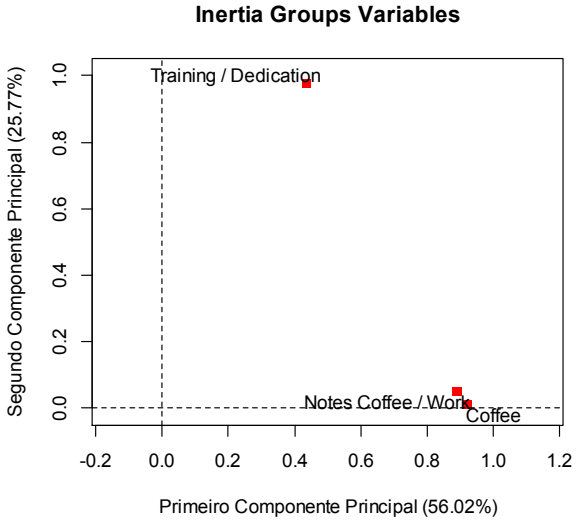
Figure 3: Plot of the groups of variables with the global analysis of the Cooperatives/Tasters



It is noticed that the position of each Cooperative/Taster in the global analysis is the barycenter (this is, the centroid) of their positions for the groups of variables. To make the interpretation easier, rows joining together the groups of variables with the global position of the cooperatives/Tasters were pulled. A priori, it is suggested that Group 2 (Training/Dedication) differs from other two variable groups.

From the inertias obtained from "Partial matrix of inertia/scores variables" in each group, aiming at a better interpretation, the plot of the inertias (Figure 4) is generated.

Figure 4: Plot of the inertias of the groups.



In short, an agreement between the discrimination of Groups 1 (Notes Coffee/Work) and 3 (Coffee), with greater inertias already mentioned in relation to Group 2 (Training/Dedication) is found.

8. **Summary**

The study of the correlations of the groups of variables proves important in all the segments of the society and the MFA technique which meets that need. So, the MFAg package adds functionality in the statistical analyses in the studies among the groups of variables of any nature, whether they be quantitative, categorical, frequency or even mixed data, generating several results of easy analysis, in addition to several plots which are easily interpreted, enabling the researcher or even the student a multivariate approach into a polyvalent problem.

9. Acknowledgments

We would like to thank CAPES for their financial support.

10. Bibliography

ABDESSEMED, L.; ESCOFIER, B.. Analyse factorielle multiple de tableaux de frequencies: comparaison avec l'analyse canonique des correspondences. **Journal de la Societe de Statistique de Paris**, Paris, v. 137, n. 2, p. 3-18, 1996.

ABDI, H.; VALENTIN, D.. Multiple factor analysis (MFA). In: SALKIND, N. J. (Ed.). **Encyclopedia of measurement and statistics**. Thousand Oaks: Sage, 2007. p. 657-663.

ABDI, H.; WILLIAMS, L.. Principal component analysis. **WIREs Computational Statatistics**, New York, v. 2, n. 4, p. 433-459, July/Aug. 2010.

ABDI, H.; WILLIAMS, L.; VALENTIN, D.. Multiple factor analysis: principal component analysis for multitable and multiblock data sets. **WIREs Computational Statatistics**, New York, v. 5, n. 2, p. 149-179, Feb. 2013.

BÉCUE-BERTAUT, M.; PAGÈS, J.. A principal axes method for comparing contingency tables: MFACT. **Computational Statistics & Data Analysis**, New York, v. 45, n. 3, p. 481-503, Feb. 2004

BÉCUE-BERTAUT, M.; PAGÈS, J.. Multiple factor analysis and clustering of a mixture of quantitative, categorical and frequency data. **Computational Statistics & Data Analysis**, New York, v. 52, n. 6, p. 3255-3268, Feb. 2008.

ESCOFIER, B.; PAGÈS, J.. **Analyse factorielles simples et multiples**. Paris: Dunod, 1990. 267 p.

ESCOFIER, B.; PAGÈS, J.. **Analyses factorielles simples et multiples: objectifs, methodes et interpretation**. 4th ed. Paris: Dunod, 2008. 318 p.

ESCOFIER, B.; PAGÈS, J.. Comparaison de groupes de variables definies sur le meme ensemble d'individus: un exemple d'applications. Le Chesnay: **Institut National de Recherche en Informatique et en Automatique**, 1982. 121 p.

ESCOFIER, B.; PAGÈS, J.. Multiple factor analysis (AFUMULT package). **Computational Statistics & Data Analysis**, New York, v. 18, n. 1, p. 121-140, Aug. 1994.

GOWER , J. C.. A general coefficient of similarity and some of its properties. **Biometrics** 27 (4), 857–871, 1971.

GREENACRE, M.; BLASIUS, J.. **Multiple correspondence analysis and related methods.** New York: Taylor and Francis, 2006. 607 p.

OSSANI, P. C.; CIRILLO, M. A.. **MFAg: Multiple Factor Analysis (MFA)**, 2016. URL http://CRAN.R-project.org/package=MFAg. R package version 1.4.

PAGÈS, J.. Analyse factorielle multiple appliquee aux variables qualitatives et aux donnees mixtes. **Revue de Statistique Appliquee**, Paris, v. 50, n. 4, p. 5-37, 2002.

PAGÈS, J.. Multiple factor analysis: main features and application to sensory data. **Revista Colombiana de Estadistica**, Bogota, v. 27, n. 1, p. 1-26, 2004.