Journal of Statistical Software

March 2008, Volume 25, Issue 1. http://www.jstatsoft.org/

FactoMineR: an R package for multivariate analysis

Sébastien Lê
Agrocampus Rennes

Julie Josse
Agrocampus Rennes

François Husson
Agrocampus Rennes

Abstract

This vignette corresponds to the paper "FactoMineR: An R Package for Multivariate Analysis" published in *Journal of Statistical Software*, 25 (1), page 1–18 (**?**).

In this article, we present **FactoMineR** an R package dedicated to multivariate data analysis. The main features of this package is the possibility to take into account different types of variables (quantitative or categorical), different types of structure on the data (a partition on the variables, a hierarchy on the variables, a partition on the individuals) and finally supplementary information (supplementary individuals and variables). Moreover, the dimensions issued from the different exploratory data analyses can be automatically described by quantitative and/or categorical variables. Numerous graphics are also available with various options. Finally, a graphical user interface is implemented within the **Rcmdr** environment in order to propose an user friendly package.

Keywords: Multivariate data analysis, Groups of variables, Hierarchy on variables, Groups of individuals, Supplementary individuals, Supplementary variables, Graphical User Interface.

1. Introduction

This vignette corresponds to the paper "FactoMineR: An R Package for Multivariate Analysis" published in *Journal of Statistical Software*, 25 (1), page 1–18 (**?**).

In this paper we present the **FactoMineR** package (**?**, **?**), a package for multivariate data analysis with R (**?**, **?**). One of the main reasons for developing this package is that we felt a need for a multivariate approach closer to our practice via:

- the introduction of "supplementary" information;
- the use of a more geometrical point of view than the one usually adopted by most of the Anglo-American practitioners.

Another reason is that obviously it represents a convenient way to implement new methodologies (or methodologies dedicated to the advanced practitioner) as the ones we're presenting

thereafter that take into account different structure on the data such as:

- a partition on the variables;
- a partition on the individuals;
- a hierarchy structure on the variables.

Finally we wanted to provide a package user friendly and oriented towards the practitioner which is what led us to implement our package in the **Rcmdr** package. No need to mention that the practitioner has the possibility to use the package both ways, *i.e.* with or without the GUI.

We will first present the most commonly used exploratory data analysis implemented in the package, then some methodologies dedicated to data endowed with some structure, at the same time as we'll set out our practice and lastly, we will show an example of the GUI.

2. "Classic" multivariate data analyses

2.1. Description of the methods

Roughly the methods implemented in the package are conceptually similar with respect to their main objective, *i.e.* to sum up and to simplify the data by reducing the dimensionality of the data set. Those methods are used depending on the type of data at hand whether variables are quantitative (numerous) or qualitative (categorical or nominal):

- Principal Component Analysis (PCA) when individuals are described by quantitative variables;
- Correspondence Analysis (CA) when individuals are described by two categorical variables that leads to a contingency table;
- Multiple Correspondence Analysis (MCA) when individuals are described by categorical variables.

Let X be the data table of interest. In order to reduce the dimensionality, X is transformed to a new coordinate system by an orthogonal linear transformation. Let F_s (resp. G_s) denotes the vector of the coordinates of the rows (resp. columns) on the axis of rank s. Those two vectors are related by the so called "**transition formulae**". In the case of PCA, they can be written:

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_k x_{ik} m_k G_s(k), \tag{1}$$

$$G_s(k) = \frac{1}{\sqrt{\lambda_s}} \sum_k x_{ik} p_i F_s(i), \tag{2}$$

where $F_s(i)$ denotes the coordinate of the individual i on the axis s, $G_s(k)$ the coordinate of the variable k on the axis s, λ_s the eigenvalue associated with the axis s, m_k the weight

associated to the variable k, p_i the weight associated to the individual i, x_{ik} the general term of the data table (row i, column k).

The transition formulae lay the foundation of our point of view and consequently set the graphical outputs at the roots of our practice. From these formulae it is crucial to analyse the scatter plots of the individuals and of the variables conjointly: an individual is at the same side as the variables for which it takes high values, and at the opposite side of the variables for which it takes low values.

2.2. Supplementary elements

Another important feature of the transition formulae is that they can be applied to supplementary individuals and/or variables in order to add supplementary information on the scatter plots for a better understanding of the data. In the PCA framework, let i' be a new individual, its coordinate on the axis of rank s can be easily obtained as followed:

$$F_s(i') = \frac{1}{\sqrt{\lambda_s}} \sum_k x_{i'k} m_k G_s(k) \tag{3}$$

In the same manner, it is also easy to calculate the coordinate of a supplementary variable when the former is quantitative; in this case the supplementary variable lies in the scatter plot of the variables. When the variable is categorical, its modalities are represented by the way of a "mean individual" per modality. For each modality, the values associated with each "mean individual" are the means of each variable over the individuals endowed with this modality; in this case the supplementary variable lies in the scatter plot of the individuals.

Notice that the supplementary information don't intervene in any way in the calculus of the vectors F_s and G_s but represent a real support when interpreting the axis as illustrated further.

2.3. Helps for the interpretation

As mentioned above most significant is the importance attached to graphical outputs. That is why they are as user friendly as possible: as an example, the possibility to enrich them with colors when adding supplementary information, the possibility to represent variables according to their quality of representation, etc.

The interpretation of the graphical outputs can also be facilitated by the use of indicators that allow to detect among the individuals and the variables which ones are well projected and which ones contribute to the construction of the axes.

The quality of representation of an element (individual or variable) on the axis of rank s is measured by the squared cosine between the vector issued from the element and its projection on the axis. If this square cosine is close to one, it means that the element is well projected on the axis. Hence, if two individuals are well represented onto a plane the distance between them can be interpreted. Let's add that for the variables, the quality of representation of a variable on a plane can be visualized by the distance between the projected variable onto the plane and the correlation circle (circle of radius 1).

The contribution of each individual to the construction of one dimension allows to detect

among the individuals which ones are extreme and contribute to the construction of the dimension.

2.4. Description of the dimensions

Each dimension of a multivariate analysis can be described by the variables (quantitative and/or categorical). These variables can have participated to the construction of the factorial axes (they can be active or supplementary).

For one quantitative variable, we calculate the correlation coefficient between the variable and the coordinates of the individuals on the axis ($F_s(i)$); we only use the data concerning the active individuals. The correlation coefficients are calculated for all the variables, dimension by dimension. Then, we can test the significance of each correlation coefficient and sort the variables from the most correlated to the less correlated. Each dimension is then described by the variables (by default, we only keep significant variables). These helps are particularly useful for the interpretation of the dimensions when there is a lot of variables.

For one categorical variable, we make a one-way analysis of variance with the coordinates of the individuals on the axis explained by the categorical variable. Then, for each category of the categorical variable, a student T-test is used to compare the average of the category with the general average (using the constraint $\sum_i \alpha_i = 0$, we test $\alpha_i = 0$). Then the p-value associated to this test is transformed to a Normal quantile in order to take into account the information that the mean of the category is less or greater than 0 (we use the sign of the difference between the mean of the category and the overall mean). This transformation is named v-test by **?**.

2.5. Examples

An example in Principal Component Analysis

To illustrate the outputs and graphs of **FactoMineR**, we use an example of Decathlon data (**?**). The data refer to athletes' performance during two athletics meetings. The data set is made of 41 rows and 13 columns: the first ten columns corresponds to the performance of the athletes for the 10 events of the decathlon. The columns 11 and 12 correspond respectively to the rank and the points obtained. The last column is a categorical variable corresponding to the athletics meeting (2004 Olympic Game or 2004 Decastar). The code to perform the PCA is:

```
> data(decathlon)
> res.pca <- PCA(decathlon, quanti.sup=11:12, quali.sup = 13)
```

By default, the PCA function gives two graphs, one for the variables and one for the individuals. Figure **??** shows the variables graph: active variables (variables used to perform the PCA) are colored in black and supplementary quantitative variables are colored in blue.

The individuals can be colored according to a categorical variable in the individual graph. To do so, the following code is used:

```
> plot (res.pca, habillage = 13)
```

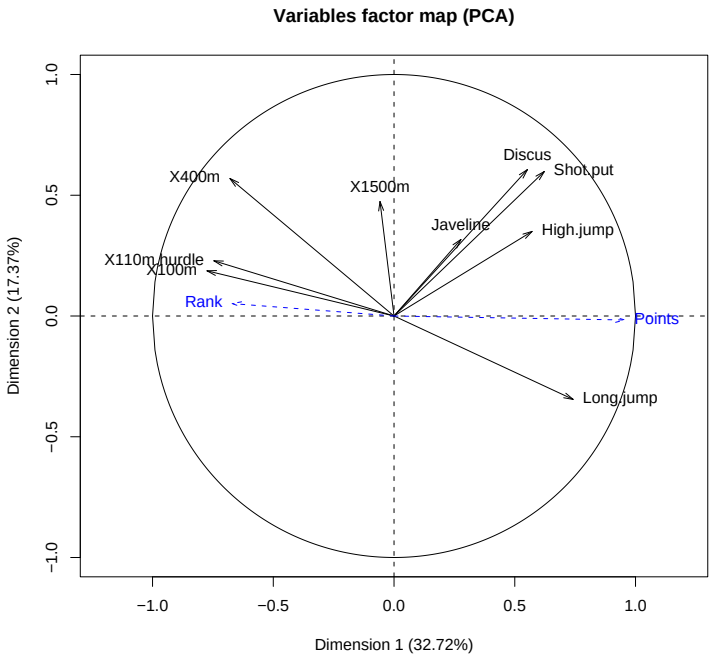

Figure 1: Variables graph (Decathlon data): supplementary variables are in blue

The `habillage = 13` indicates that individuals are colored according to the 13th variable.

Thus, the athletes are colored according to the athletics meeting (Fig. **??**). The athletes who participated to the Olympic Game are colored in red and the athletes who participated to the Decastar are colored in black.

The percentage of variability explained by each dimension is given: 32.72% for the first axis and 17.37% for the second one.

We can draw a bar plot with the eigenvalues (Fig. **??**) with the following code:

```
> barplot(res.pca$eig[,1], main = "Eigenvalues",
names.arg = paste("Dim",1:nrow(res.pca$eig), sep=""))
```

This graph allows to detect the number of dimensions interesting for the interpretation. The third and fourth dimension may be interesting, so we can plot the graph for these two dimensions. For the variables (Fig. **??**), we will use the code:

```
> plot(res.pca, choix = "var", axes = c(3,4), lim.cos2.var = 0)
```

The parameter `choix = "var"` indicates that we plot the graph of the variables, the parameter `axes = c(3,4)` indicates that the graph is done for the dimension 3 and 4, and the parameter `lim.cos2.var = 0` indicates that all the variables are drawn (more precisely, all the variables having a quality of projection greater than 0; this option is interesting to keep only the variables well projected).

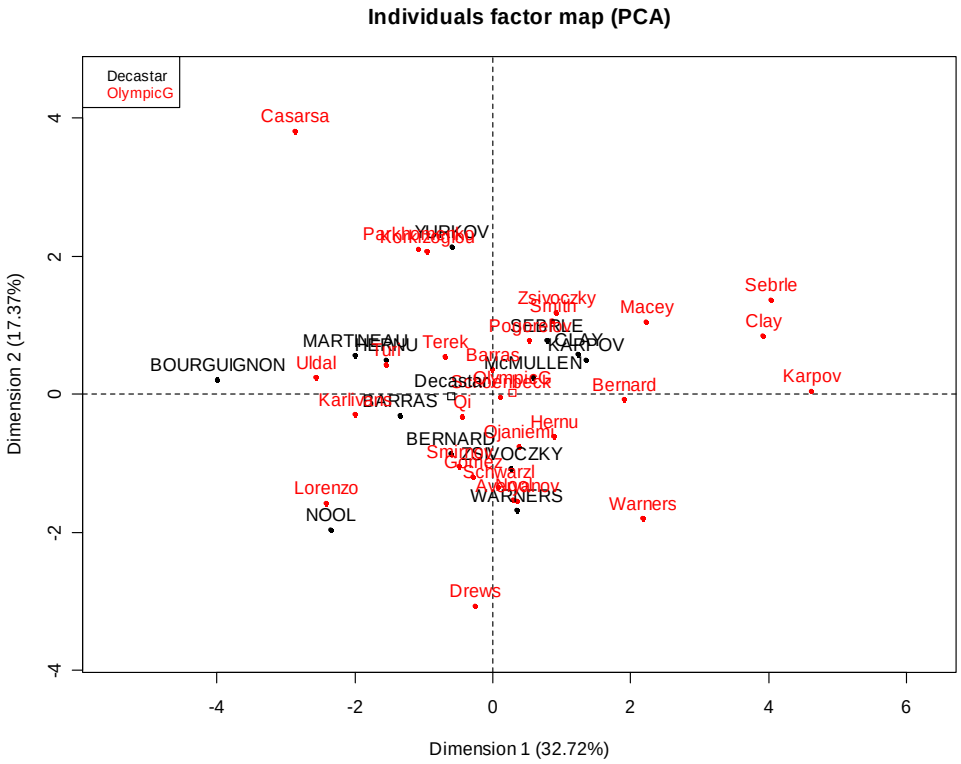

Figure 2: Individuals graph (Decathlon data): individuals are colored from the athletics meeting

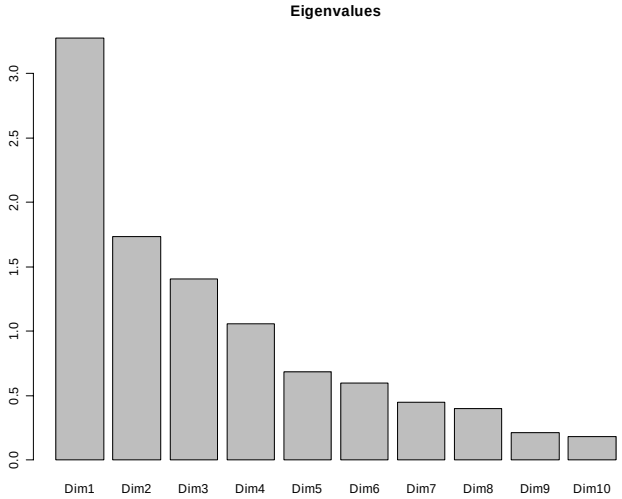

Figure 3: Barplot of the eigenvalues

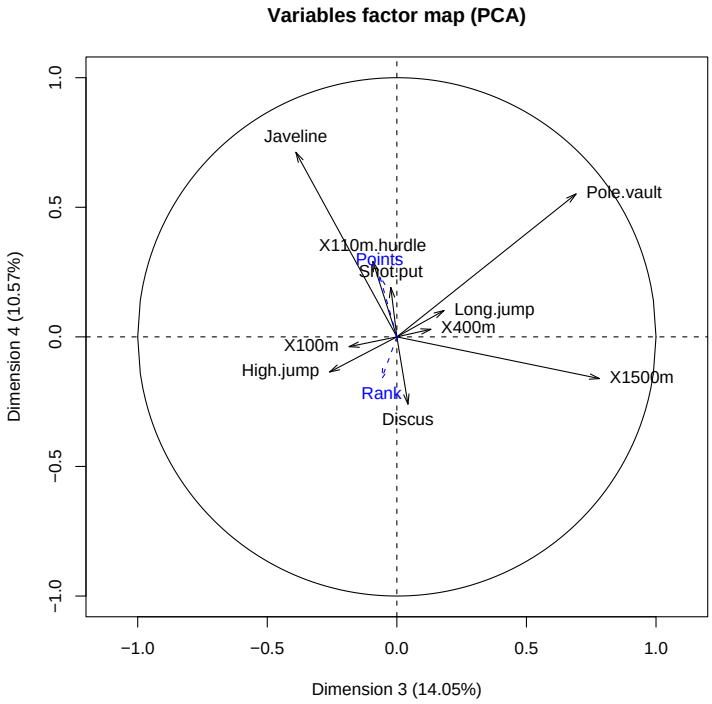

Figure 4: Variables graph (Decathlon data) for dimensions 3 and 4

The results are given in a list with several objects with the `print` function:

```
> print(res.pca)
```

Results (Table **??**) are given for the individuals, the active variables, the quantitative and categorical supplementary variables.

As mentioned above, we can describe each principal component using the `dimdesc` function:

```
> dimdesc(res.pca, proba = 0.2)
```

Table **??** gives the description of the first dimension of the PCA done on the Decathlon data. The variables are kept if the p-value is less than 0.20 (`proba = 0.2`). The variable which describe the best the first dimension is the *Points* variable (it was a supplementary variable), and then, it is the *X100m* variable which is negatively correlated with the dimension (the individuals who have a great coordinate on the first axis have a low *X100m* time). The first dimension is then described by the categorical variable *Competition*. The *Olympic Game* category has a coordinate significantly greater than 0 showing that the athletes of this competition have greater coordinates than 0 on the first axis. Since, the variable *Points* is highly correlated with this axis (the correlation is positive), the athletes for this competition made better performances.

An example in Correspondence Analysis

We present a Correspondence analysis done with **FactoMineR** on the data set presented in

```
**Results for the Principal Component Analysis (PCA))**

The analysis was done on  41 individuals, described by 13 variables

*The results are available in the following objects:

   nom                 description
1  "$eig"              "eigenvalues"
2  "$var"              "results for the variables"
3  "$var$coord"        "coordinates of the variables"
4  "$var$cor"          "correlations variables - dimensions"
5  "$var$cos2"         "cos2 for the variables"
6  "$var$contrib"      "contributions of the variables"
7  "$ind"              "results for the individuals"
8  "$ind$coord"        "coord. for the individuals"
9  "$ind$cos2"         "cos2 for the individuals"
10 "$ind$contrib"      "contributions of the individuals"
11 "$quanti.sup"       "results for the supplementary quantitative variables"
12 "$quanti.sup$coord" "coord. of the supplementary quantitative variables"
13 "$quanti.sup$cor"   "correlations supp. quantitative variables - dimensions"
14 "$quali.sup"        "results for the supplementary qualitative variables"
15 "$quali.sup$coord"  "coord. of the supplementary categories"
16 "$quali.sup$vtest"  "v-test of the supplementary categories"
17 "$call"             "summary statistics"
18 "$call$centre"      "mean for the variables"
19 "$call$ecart.type"  "standard error for the variables"
20 "$call$row.w"       "weights for the individuals"
21 "$call$col.w"       "weights for the variables"
```

Table 1: List with the results of the PCA

?. The data used here is a contingency table that summarizes the answers given by different categories of people to the following question: "according to you, what are the reasons that can make hesitate a woman or a couple to have children?" The data frame is made of 18 rows and 8 columns. Rows represent the different reasons mentioned, columns represent the different categories (education, age) people belong to.

```
> data(children)
> res.ca <- CA (children, col.sup = 6:8, row.sup = 15:18)
```

The columns from 6 to 8 are supplementaries (they concern the age groups of the people), and rows from 15 to 18 are either supplementaries. By default, the CA function gives one graphical output (Fig. **??**).

If we just want to visualize the active elements (Fig. **??**), we use the following code:

```
$Dim.1
$Dim.1$quanti
                 Dim.1
Points        0.9561543
Long.jump     0.7418997
Shot.put      0.6225026
High.jump     0.5719453
Discus        0.5524665
Rank         -0.6705104
X400m        -0.6796099
X110m.hurdle -0.7462453
X100m        -0.7747198

$Dim.1$quali
             Dim.1
OlympicG  1.429753
Decastar -1.429753
```

Table 2: Description of the first dimension for the Decathlon data

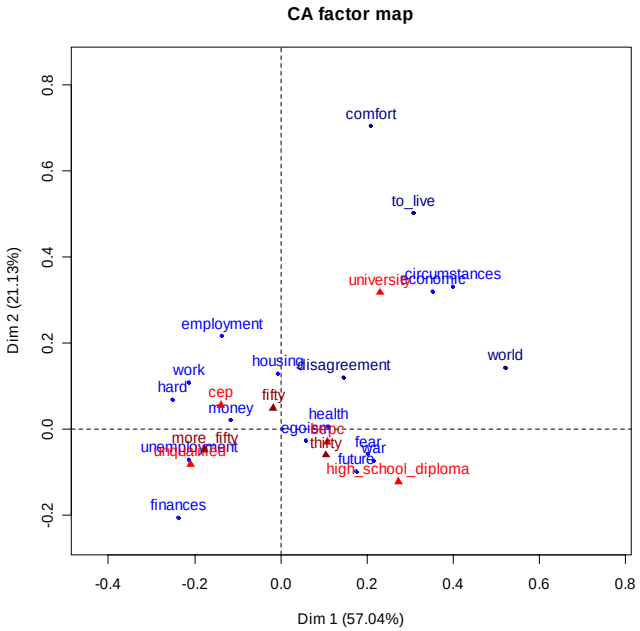

Figure 5: Correspondence Analysis: the active rows are colored in blue, the active columns are colored in red, the supplementary rows are colored in dark blue, the supplementary columns are colored in dark red

```
> plot (res.ca, invisible = c("row.sup", "col.sup"))
```

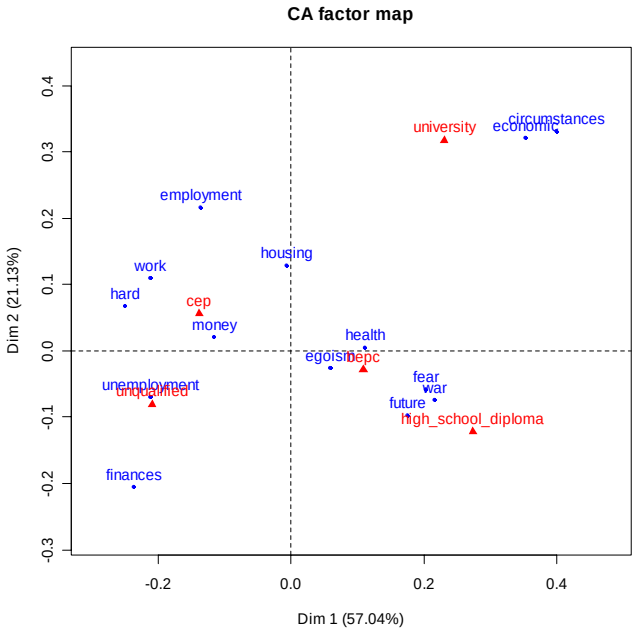

Figure 6: Correspondence Analysis with only the active elements

3. Structure on the data

In the **FactoMineR** package it is possible to take into account different types of structure on the data. Data may be organized into groups of individuals, groups of variables or into a hierarchy on the variables. In this section we present the different structures and the methods associated with.

3.1. Groups of variables, the point of view of Multiple Factor Analysis (in the sense of Escofier-Pagès)

One problem can be expressed when studying the relations between several sets of variables. This problem is very old and the first method suggested within this framework is the canonical analysis (**?**). This method has remarkable properties and plays a central theoretical part in data analysis particularly if we consider that a lot of traditional methods (linear regression, discriminant analyses, correspondence analyses, etc.) can be seen as a particular case. But, in practice, the canonical analysis does not hold its promises. The essential reason is that, in this method, each group of variables is just considered through the subspace that it generates. In other words, the repartition of the variables in these subspaces is not taken into account. Thus the analysis can highlight dimensions which are not closely related to any initial variables, which is poorly interesting. The taking into account of the variables repartition in different subspaces which they generate can be made by a Multiple Factor Analysis in the sense of Escofier-Pagès (MFA; **?** **?**) or Generalized Procrustes Analysis (GPA, (**?**)), two methods

implemented in the package.

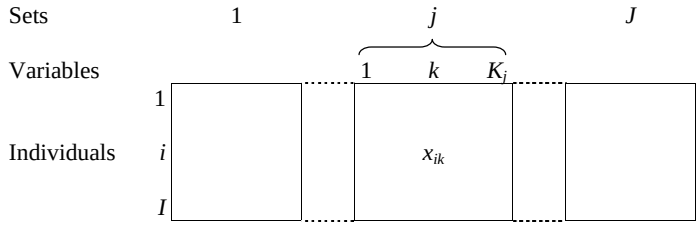

Figure 7: Data table subjected to MFA (I individuals, J groups of variables).

The heart of MFA is a PCA in which weights are assigned to the variables; in other words, a particular metric is assigned to the space of the individuals. More precisely, a same weight is associated to each variable of the group j ($j = 1, ..., J$). The weight is the first eigenvalue of the PCA on the group j. Thus, the maximum axial inertia of each group of variables is equal to 1. The influence of the groups of variables in the global analysis is balanced and the structure of each group is respected. This weighing presents a simple direct interpretation. It has also invaluable indirect properties; in particular it allows to consider MFA as a particular generalized canonical analysis within the meaning of **?** (**?**).

For each group of variables one can associate a cloud of individuals. This cloud is the one which is considered in the PCA for the only group j (after above mentioned standardization by the first eigenvalue). MFA provides a superimposed representation of these clouds, with the manner of a procrustes analysis. This representation can be presented in two ways: as a projection of a cloud of points and as a canonical variable. Here, a third way is chosen, based on a very useful property.

While taking into account the structure of variables in J groups and while using the weighting of MFA ($m_k = \frac{1}{\lambda_1^j}$ if the variable k is in the group j), this relation becomes:

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_{j=1}^{J} \frac{1}{\lambda_1^j} \sum_{k=1}^{K_j} x_{ik} G_s(k)$$

where K_j denotes the number of variables in the group j.

According to this relation, an individual is on the side of the variables for which it takes high values (and all the more far from the origin that these values are high) and on the opposite side of the variables for which it takes low values. The representation of the partial cloud is obtained by restricting the previous relation with only the variables of the group j. Thus, the coordinate ($F_s(i^j)$) on the axis s, of the individual i seen by the only group j (known as the partial individual i^j) can be written:

$$F_s(i^j) = \frac{1}{\sqrt{\lambda_s}} \frac{1}{\lambda_1^j} \sum_{k} x_{ik} G_s(k)$$

This equation is a general interpretation of the PCA but restricted to the only variables of the group j. The partial individual i^j is on the side of the variables of the group j for which it takes high values, and on the opposite side of the variables of the group j for which it takes

low values. This property expresses a direct relation between the positions of the partial individuals and the representation of the variables. It is so natural that many users of MFA use it ... without knowing it. It has no equivalent in the procrustes analyzes.

On the graphs it is pleasant to see the point i in the exact barycenter of the points $\{i^j, j = 1, ..., J\}$. In practice, the coordinates $F_s(i^j)$ are multiplied by J. Thus, without modifying the relative positions of the partial points, the required property is obtained:

$$F_s(i) = \frac{1}{J} \sum_{j=1}^{J} F_s(i^j)$$

It may be also interesting to represent the groups of variables as points in a scatter plot to visualize their common structure. To each group of variables j, one can associate the scalar product matrix between individuals. This matrix of dimension $I \times I$ (I is the number of individuals) is denoted W_j and can be regarded as a point in the Euclidean space of dimension I^2, denoted R^{I^2}. In this space, the cosine of the angle formed by the origin and the two points W_j and W_l is the RV coefficient between the two groups j and l. The representation of the groups provided by MFA is obtained by projection upon vectors of R^{I^2} induced by the MFA factors: one factor may be considered as a set consisting of a single variable; it is then possible to associate this set to a scalar product matrix and thus to a vector of R^{I^2}.

MFA allows to analyse several groups of variables which can be quantitative and/or categorical when GPA allows to analyse only groups of quantitative variables.

As in PCA, the practitioner has the possibility to add supplementary information (individuals, quantitative and categorical variables), and in the case of MFA, user can add supplementary groups of variables for instance.

3.2. Hierarchy on the variables

In many data sets, variables are structured according to a hierarchy leading to groups and subgroups of variables (Fig. **??**). This case is frequently encountered with questionnaires structured into topics and subtopics. Analyzing such data implies balancing the part of each group all together on the one hand, but also that of each subgroup among them on the other hand. To do so, it seems necessary to consider a hierarchy. The usual methods mentioned above do not suit this type of problem since they lead to outputs where a point of view of a group of variables may be preponderant in comparison to the point of view of other groups.

The approach to consider such a structure on the variables in a global analysis involves balancing the groups of variables within every node of the hierarchy.

Hierarchical Multiple Factor Analysis (HMFA, **? ?** and **? ?**) is an extension of MFA to the case where variables are structured according to a hierarchy.

In HMFA, a succession of MFA is applied to each node of the hierarchy in order to balance the groups of variables within every node, by going through the hierarchical tree from the bottom up. Not only HMFA provides a graphical display of the individuals according to the whole set of (weighted) variables, but it also displays the individuals as described by each group of variables: as mentioned above, an individual which is described by just one group of variables is called a "partial individual". An interesting feature of the analysis is that the partial representation of each individual at each node is at the center of gravity of the partial

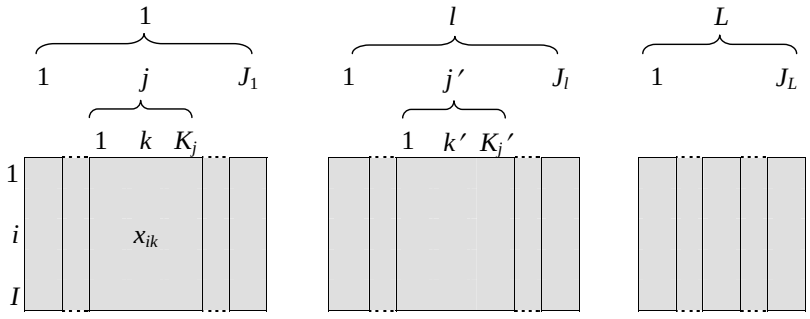

Figure 8: Example of hierarchy on the variables: there is two levels for the hierarchy. The first one contains L groups, each l group contains J_l subgroups, and each subgroup have K_j variables.

representation of this individual associated with the various subsets of variables nested within this node.

Moreover, HMFA provides a representation of the nodes involved in the hierarchy; the principle of this representation is similar to that of MFA.

3.3. Groups of individuals

The analysis of data comprising several sets of individuals described by a same set of variables is a problem frequently encountered. Those groups may be issued from a previous statistical analysis such as a classification; other examples are provided by international surveys where groups of individuals coming from different countries are questioned according to a same set of questions. In this section we present two methodologies implemented in the package to analyze data organized into groups of individuals.

Description of categories

For this first method we consider two cases depending on the type of the variable describing the groups, wether it is numerical or categorical.

If a variable is quantitative, the mean of one group for this variable is calculated and compared to the overall mean. More precisely, (**?**) proposed to calculate the following quantity:

$$u = \frac{\bar{x}_q - \bar{x}}{\sqrt{\frac{s^2}{n_q}\left(\frac{N-n_q}{N-1}\right)}}$$

where n_q denotes the number of individuals for the group q, N the total number of individuals, s the standard deviation for all the individuals.

The quantity u can then be compared to the appropriate quantile of the Normal distribution. If this quantity is more extreme than the quantile of the Normal distribution, then the variable is interesting to describe the group of individuals. The interesting variables are then sorted from the most to the less interesting variable.

If a variable is categorical, then the frequency N_{qj} corresponding to the number of individuals of the group q who take the category j (for the categorical variable) is distributed as an

hypergeometric distribution with the parameters N, n_j, n_q/N (where n_j denotes the number of individuals that have taken the category j). A p-value is then calculated by category (and by categorical variable). The categories are sorted from the highest to the lowest p-value.

Dual Multiple Factor Analysis

Dual Multiple Factor Analysis (DMFA, **? ?**), is an extension of Multiple Factor Analysis (in the sense of Escofier-Pagès) in the case where individuals are structured according to a partition. The heart of the method rests on a factorial analysis known as internal, in reference to the internal correspondence analysis, for which data are systematically centered by group. This analysis is an internal PCA when all the variables are quantitative. DMFA provides the classic results of a PCA as well as additional outputs induced by the consideration of a partition on individuals, such as the superimposed representation of the L scatter plots of variables associated with the L groups of individuals and the representation of the scatter plot of the correlations matrices associated each one with a group of individuals.

4. Rcmdr support for the FactoMineR package

The user has the possibility to easily add an extra menu to the ones already proposed by the **Rcmdr** package (Fig. **??** shows the menu of the **FactoMineR** interface). To do so, there are two possibilities. First possibility, the user can use the **RcmdrPlugin.FactoMineR** package which is available on the CRAN. Second possibility, once connected to the internet, all the user has to do is to write the following line code:

```
> source("http://factominer.free.fr/install-facto.r")
```

This interface is user-friendly and allows to make graphs and to save results in a file very easily as explained below.

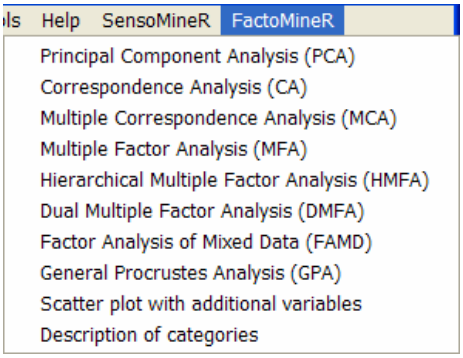

Figure 9: Menu of the **FactoMineR** package

As an example, we show the interface for the PCA function (Fig. **??**).

The main window allows to choose the active variables (by default all the variables are active and the PCA can be performed). Several buttons allow to choose the supplementary quantitative or categorical variables, the supplementary individuals, the outputs to be displayed or

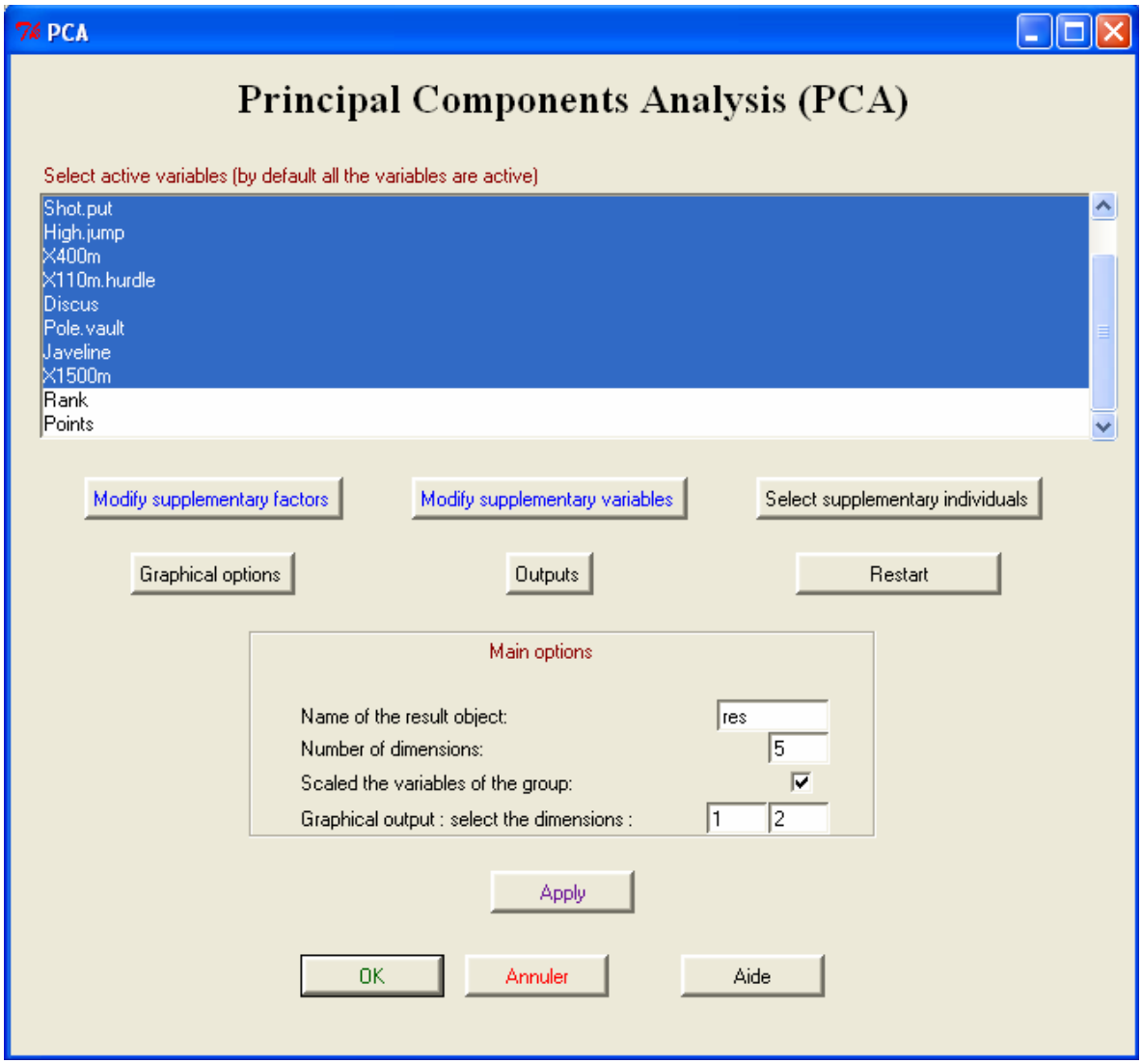

Figure 10: Main window for the PCA function

the graphs to be plotted.

The graphical options concern the two main graphs: the scatter plots of the individuals and of the variables. Relating to the individuals graph, it is possible to represent the active individuals, the supplementary individuals, the categories of the supplementary categorical variables; it is also possible to choose the elements that we want to draw. The individuals can be colored according to one categorical variable (the categorical variable available are proposed in a list).

Relating to the variables graph, active and/or illustrative variables can be drawn. If there are a lot of variables, one can represent only the variables that are well projected on the plane (by default the variables are drawn if their quality of representation is greater than 10%).

Several outputs are also available (Fig. **??**). The dialog box allows to give all the results

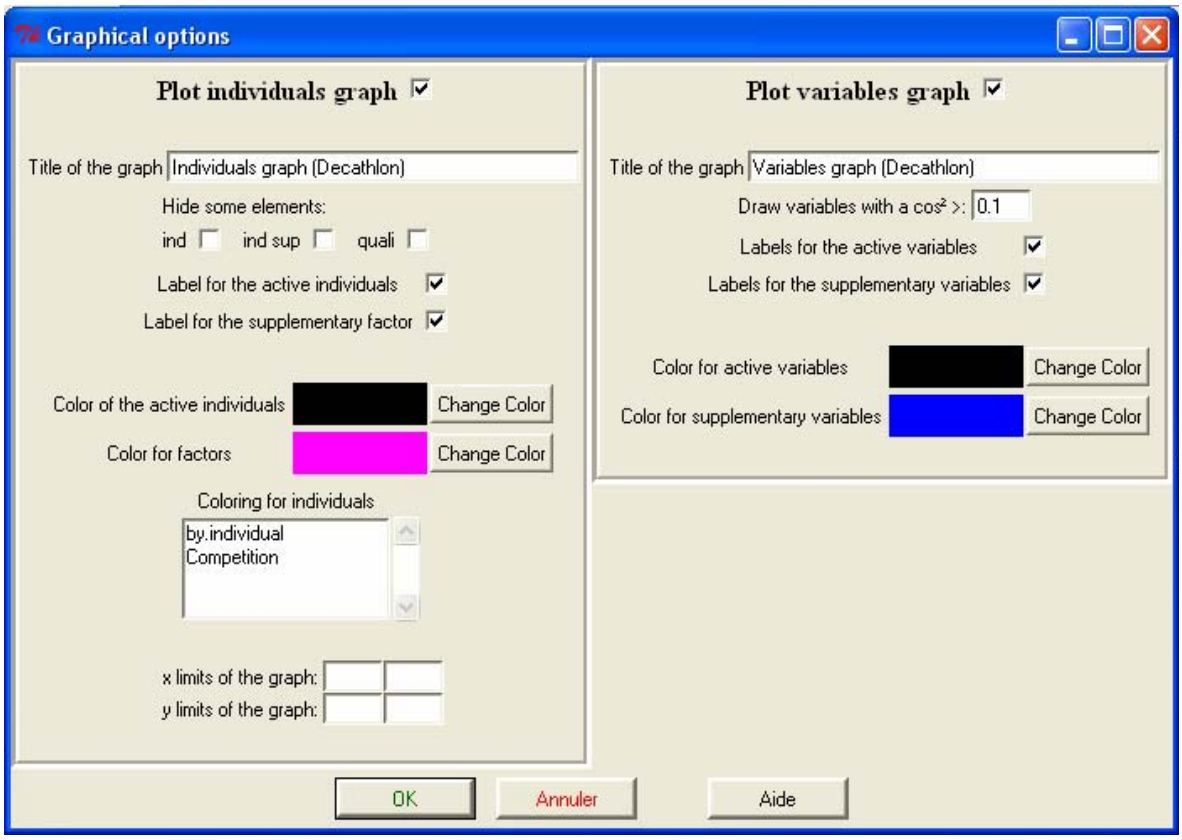

Figure 11: Window with the graphical options available for the PCA function

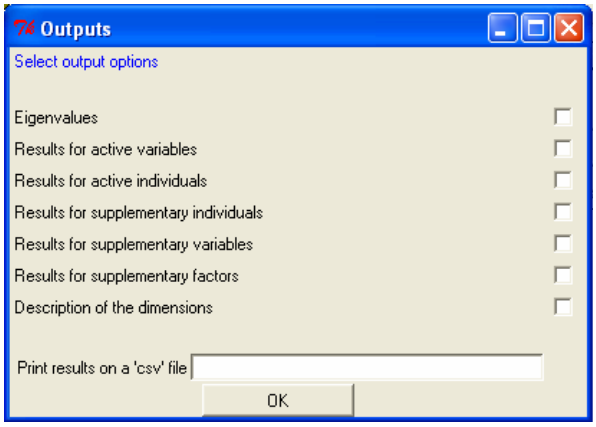

Figure 12: Window with the outputs available for the PCA function

from the PCA function, *e.g.* the eigenvalues, the results for the individuals and the variables (active or supplementary). One can also get an automatic description of the dimensions of the exploratory data analysis. All these results can be written in a file (a $*.csv$ file which can be open with Excel).

5. Conclusion

The main features of the R package **FactoMineR** have been explained and illustrated in this paper, using the data set `decathlon` that is available in the package.

The website http://factominer.free.fr/ gives some examples for the different methods available in the package; you can also find our latest references related to the methods developed in our team at the following address http://agrocampus-rennes.fr/math/.

Affiliation:

Sébastien Lê
Agrocampus Rennes
UMR CNRS 6625
65 rue de Saint-Brieuc
35042 Rennes Cedex
E-mail: Sebastien.Le@agrocampus-rennes.fr
URL: http://www.agrocampus-rennes.fr/math/le

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 25, Issue 1 *Submitted:* 2007-04-05
March 2008 *Accepted:* 2008-01-09