Multivariate Analysis with the FactoMineR package

Julie Josse, Sébastien Lê and François Husson

Agrocampus Rennes - 35 rue de Saint Brieuc - 35042 Rennes (France)
E-mail: Julie.Josse@agrocampus-rennes.fr

Abstract: The `FactoMineR` package allows to make 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 factorial 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 a 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

We present the FactoMineR package (Husson *et al.*, 2005), 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 and 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 (partition on the variables, partition on the individuals, 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 with or without the GUI.

We will first present the most commonly used factorial 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 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): Principal Component Analysis (PCA) when variables are quantitative, Correspondence Analysis (CA) for contingency tables, Multiple Correspondence Analysis (MCA) for categorical variables.

Let X be the data table of interest. 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:

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_k x_{ik} m_k G_s(k) \quad \text{and} \quad G_s(k) = \frac{1}{\sqrt{\lambda_s}} \sum_k x_{ik} p_i F_s(i).$$

where $F_s(i)$ (resp. $G_s(k)$) denotes the coordinate of the individual i (resp. 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:

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

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

The interpretation of the graphical outputs can 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 (*i.e.* such as the squared cosine and the contribution).

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.

2.4 Description of the dimensions

Each dimension of a multivariate analysis can be described by the variables (quantitative and/or qualitative; active or supplementary). For one quantitative variable, we calculate the correlation coefficient between the variable and the coordinates of the individuals on the axis s ($F_s(i)$); 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.

For one qualitative variable, we make a one-way analysis of variance with the coordinates of the individuals on the axis explained by the qualitative variable. Then, for each category, 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 Lebart *et al.* (1997).

2.5 An example in Principal Component Analysis

To illustrate the outputs and graphs of the package, we use an example of decathlon data (Husson and Pagès, 2005). 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).

By default, the PCA function gives two graphs, one for the variables (Fig. 1) and one for the individuals (Fig. 2). Variables are colored according to their status (black for active variables and blue for the supplementary ones). The individuals are colored according to a qualitative variable (in red when the athletes participated to the Olympic Game, in black when they participated to the Decastar).

Table 1 gives the description of the first dimension of the PCA. The variables are kept if the p-value is less than 0.20. 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 (an individual who has a great coordinate on the first axis has a low *X100m* time). The first dimension is also described by the qualitative 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 athletes for this competition made better performances.

3 Structure on the data

In `FactoMineR` 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, into a hierarchy on the variables. In this section we present the different structures and the methods associated with.

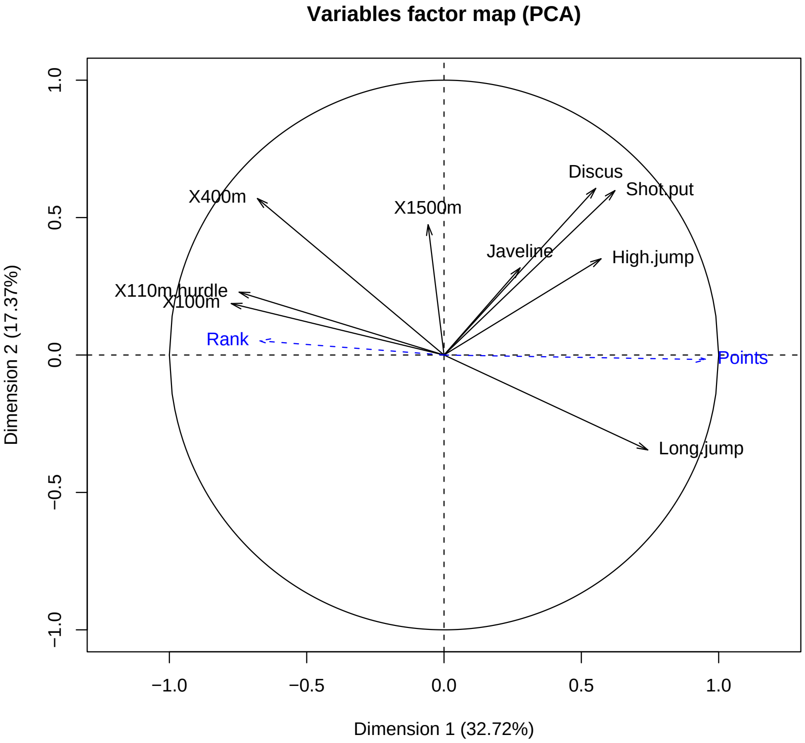


Figure 1: Variables graph (Decathlon data): supplementary variables are in blue

```
$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 1: Description of the first dimension for the Decathlon data

3.1 Groups of variables, the point of view of Multiple Factor Analysis

Multiple Factor Analysis (MFA; Escofier and Pagès, 1998) or Generalized Procrustes Analysis (GPA; Gower, 1975) allow to study the relations between several sets 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.

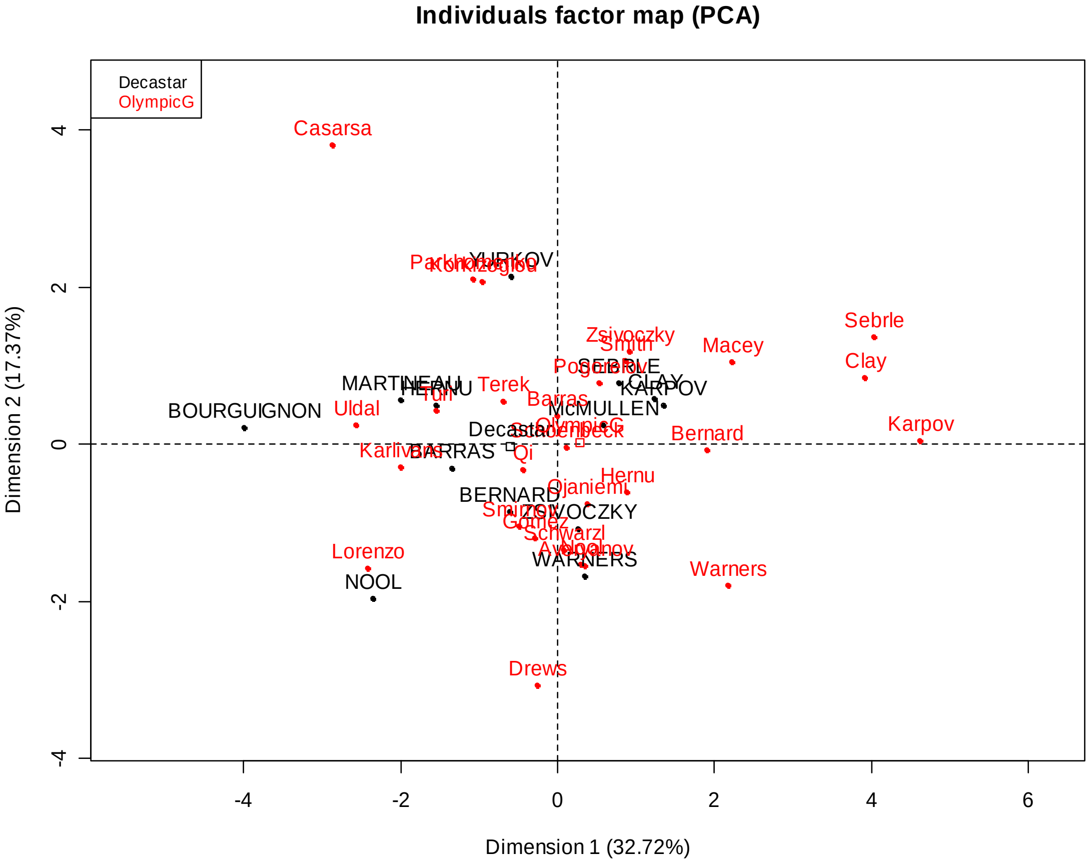


Figure 2: Individuals graph (Decathlon data): individuals are colored from the athletics meeting

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. For each group of variables a partial point is associated to each individual: it represents the projection of an individual seen by the variables of one group. It is also interesting to represent the groups of variables as points in a scatter plot to visualize their common structure.

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

3.2 Hierarchy on the variables

In many data sets, variables are structured according to a hierarchy leading to groups and subgroups of variables. 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. Hierarchical Multiple Factor Analysis (HMFA, Le Dien and Pagès, 2003) consider such a structure on the variables in a global analysis involves balancing the groups of variables within every node of the hierarchy: it is an extension of MFA to the case where variables are structured according to a hierarchy.

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 can be a priori defined or may be issued from a previous statistical analysis such as a classification.

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, Lebart *et al.* (1997) 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 is then compared to the appropriate quantile of the Normal distribution. If u is greater than the Normal quantile, 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 qualitative, then the frequency N_{qj} corresponding to the number of individuals of the group q who take the category j (for the qualitative 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 qualitative variable). The categories are sorted from the highest to the lowest p-value.

Dual Multiple Factor Analysis: Dual Multiple Factor Analysis (DMFA, Lê *et al.*, 2007), is an extension of Multiple Factor Analysis in the case where individuals are structured by group. 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 gives a 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

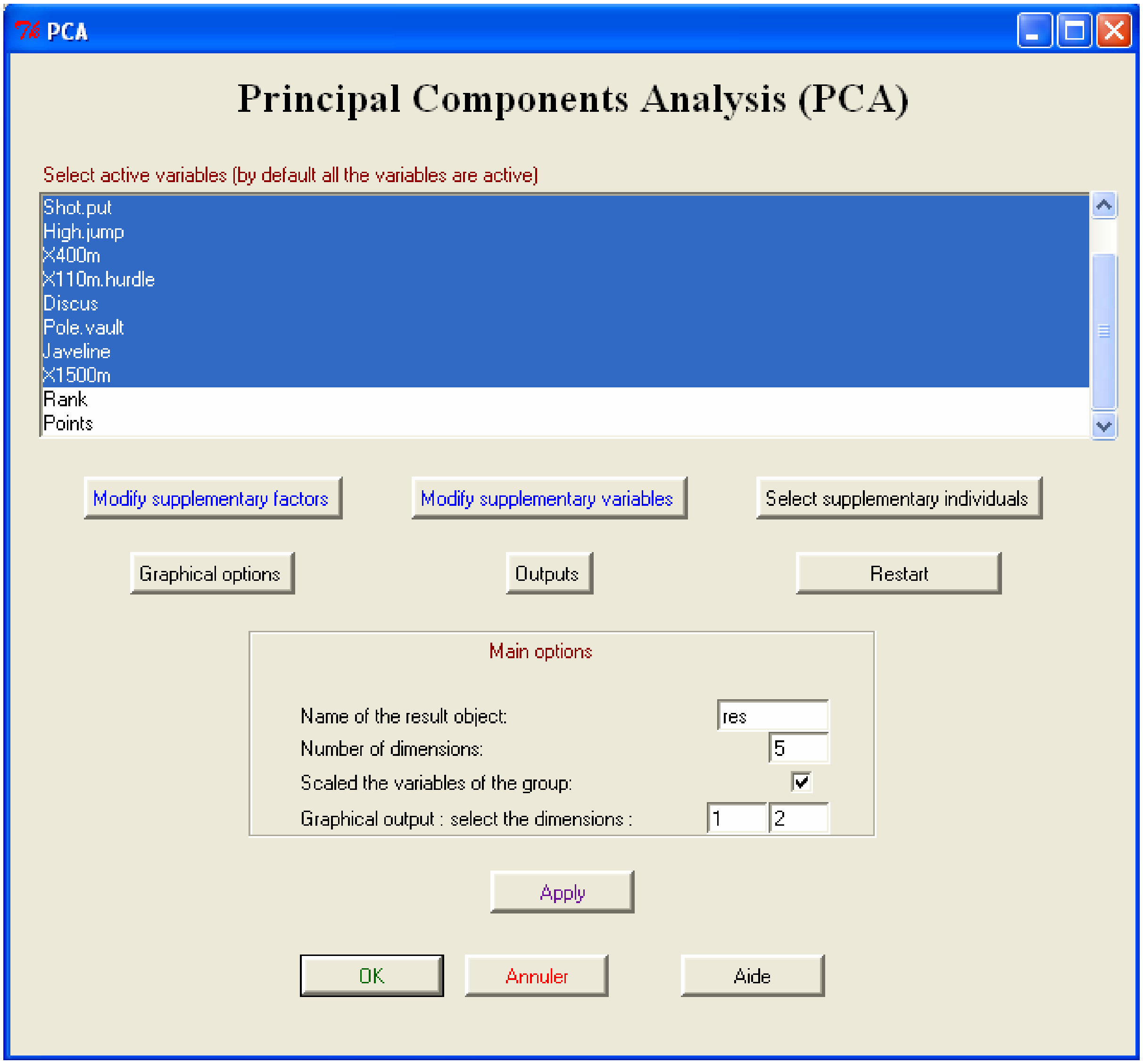
The user has the possibility to easily add an extra menu to the ones already proposed by the `Rcmdr` package: connected to the internet, he has to write the following line code in the `R` console:

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

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

As an example, we show the interface for the PCA function (Fig. 3). The main window allows to choose the active variables (by default all the variables are active and the PCA can be done). Several buttons allow to choose supplementary quantitative or qualitative variables, supplementary individuals, or to choose the outputs or the graphs to plot.

The graphical options concern the two graphs (individuals and variables, see Fig. 4). In the individuals graph, one may represent the active individuals, the supplementary individuals, the categories of supplementary categorical variables or to choose the elements that we want to draw.



Principal Components Analysis (PCA)

Select active variables (by default all the variables are active)

Shot.put
High.jump
X400m
X110m.hurdle
Discus
Pole.vault
Javeline
X1500m
Rank
Points

Modify supplementary factors
Modify supplementary variables
Select supplementary individuals
Graphical options
Outputs
Restart

Main options

Name of the result object: res
Number of dimensions: 5
Scaled the variables of the group: ☑
Graphical output : select the dimensions : 1 2

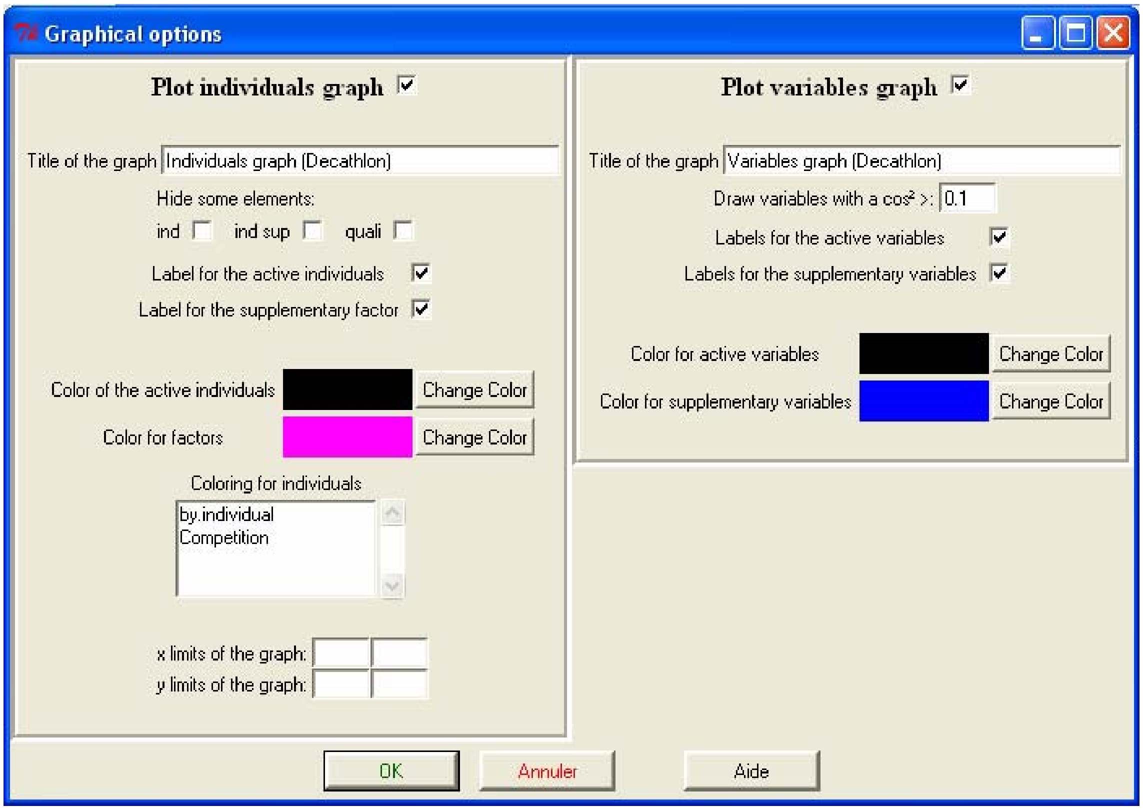
Figure 3: Main window for the PCA

The individuals can be colored according to one categorical variable (the categorical variables available are proposed in a list). For the variables graph, active and/or illustrative variables can be drawn. If there are a lot of variables, one can only draw the variables well projected on the plane (by default the variables are drawn if their quality of projection is greater than 10%).

The dialog box of the outputs allows to describe the dimensions and to write the results in a file (a $*.csv$ file which can be open with Excel).

5 Conclusion

The site `http://factominer.free.fr/` gives some examples for the different methods available in the package, and the site `http://agrocampus-rennes.fr/math/` gives the publications on the recent methods developed in our team.



Plot individuals graph ☑

Title of the graph: Individuals graph (Decathlon)

Hide some elements:

ind □ ind sup □ quali □

Label for the active individuals ☑

Label for the supplementary factor ☑

Color of the active individuals ■

Color for factors 🟪

Coloring for individuals: by.individual, Competition

x limits of the graph:

y limits of the graph:

Plot variables graph ☑

Title of the graph: Variables graph (Decathlon)

Draw variables with a cos² >: 0.1

Labels for the active variables ☑

Labels for the supplementary variables ☑

Color for active variables ■

Color for supplementary variables 🟦

Figure 4: Window with the graphical options available for the PCA

6 References

Escofier, B. and Pagès, J. (1998) *Analyses factorielles simples et multiples.* Dunod.

Gower, J. C. (1975) Generalized Procrustes Analysis. *Psychometrika*, **40**, 33-51.

Husson, F. and Lê, S. and Mazet, J. (2007) FactoMineR: Factor Analysis and Data Mining with R. `http://factominer.free.fr`.

Husson, F. and Pagès, J. (2005). *Statistiques générales pour utilisateurs.* Presses Universitaires de Rennes.

Le Dien, S. and Pagès, J. (2003) Hierarchical Multiple Factor Analysis: application to the comparison of sensory profiles. *Food Quality and Preference*, **14**, 397-403.

Lê, S. and Pagès, J. (2007) DMFA: Dual Multiple Factor Analysis. *12th International Conference on Applied Stochastic Models and Data Analysis.*

R Development Core Team (2006) R: A Language and Environment for Statistical Computing. Vienna, Austria. `http://www.R-project.org`.

Lebart, L., Morineau, A. and Piron, M. (1997). *Statistique exploratoire multidimensionnelle.* Dunod.