Multivariate analysis with the FactoMineR package

Julie Josse, Sébastien Lê & François Husson

AGROCAMPUS RENNES

FACTOMINER

Outline

- Why a new package in multivariate data analysis?
- The classical methods with a lot of helps to interpret
- Advanced methods
- The Graphical User Interface

Why a new package in multivariate data analysis?

- Possibility to add supplementary information in classical methods of data analyses
- The use of a more geometrical point of view allowing to draw graphs
- The possibility to propose new methods (taking into account different structure on the data)
- To have a package user friendly and oriented to practitioner (a very easy GUI)

1 – The classical methods

- Methods implemented are similar in their main objective: to sum up and simplify the data by reducing the dimensionality of the dataset
- Methods are used according to the type of data (quantitative or qualitative)

Quantitative variables	Principal Components Analysis
Contingency table	Correspondence Analysis
Qualitative variables	Multiple Correspondence Analysis

1 – The classical methods

Transition formulas in PCA:

Coordinate of individual *i* on axis *s* $$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_k x_{ik} G_s(k),$$

Coordinate of variable *k* on axis *s* $$G_s(k) = \frac{1}{\sqrt{\lambda_s}} \sum_k x_{ik} F_s(i),$$

Eigenvalue associated with axis *s*

General term of the data table (row *i*, column *k*)

What does it mean?

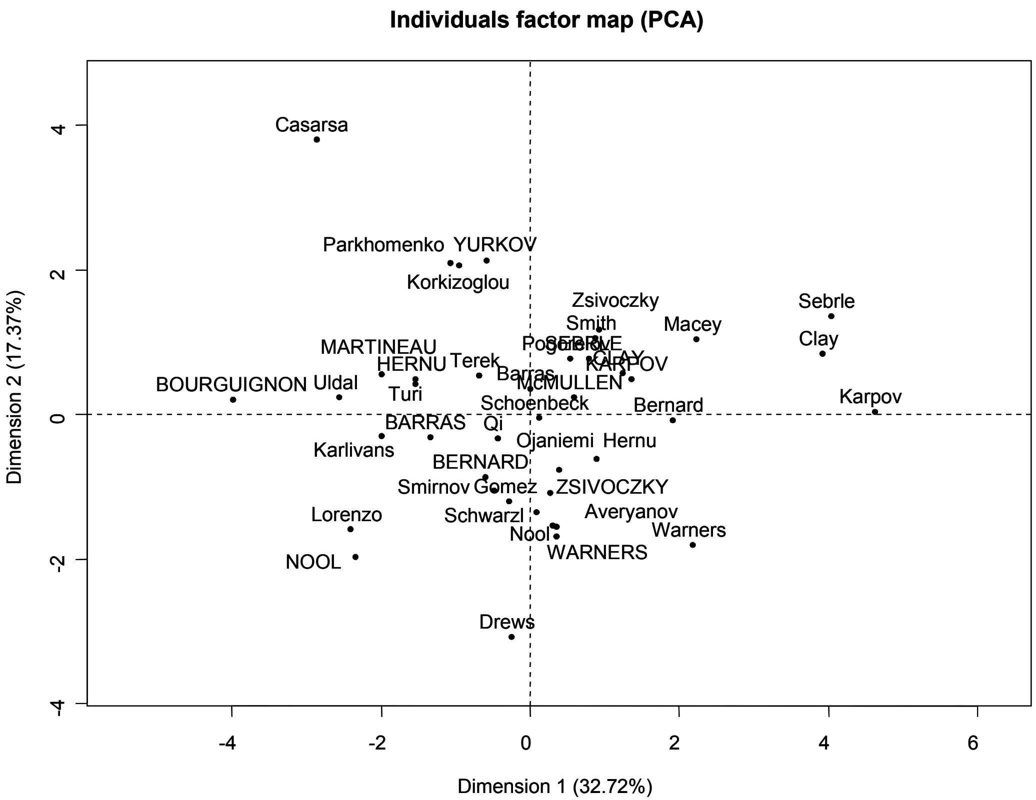
An individual is at the same side as the variables for which it takes high values

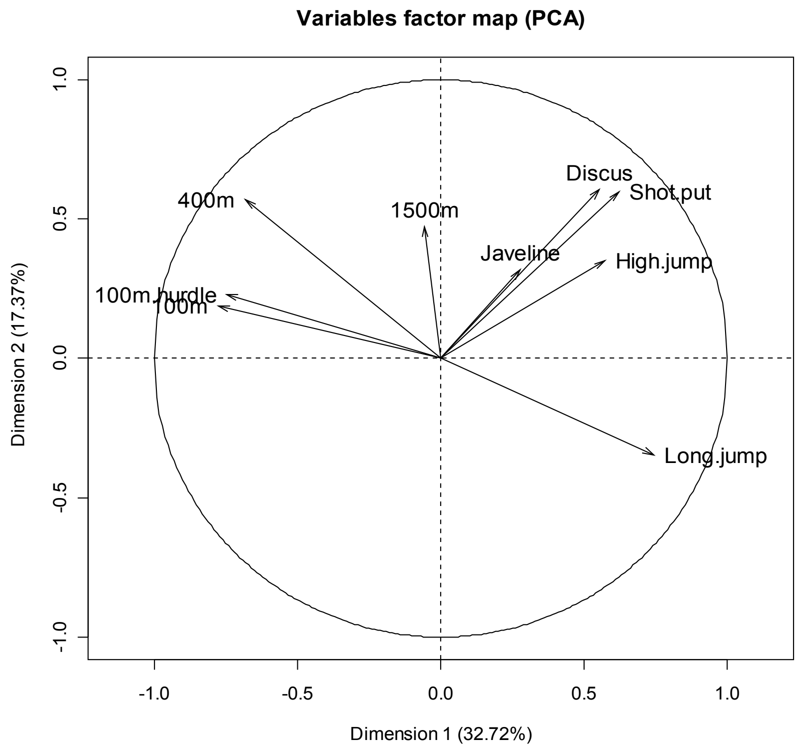
Example of PCA

The data: the performances of 41 athletes during two meetings of decathlon

	100m	Long.jump	Shot.put	High.jump	400m	110m.hurdle	Discus	Pole.vault	Javeline	1500m	Rank	Points	Competition
SEBRLE	11.04	7.58	14.83	2.07	49.81	14.69	43.75	5.02	63.19	291.70	1	8217	Decastar
CLAY	10.76	7.40	14.26	1.86	49.37	14.05	50.72	4.92	60.15	301.50	2	8122	Decastar
KARPOV	11.02	7.30	14.77	2.04	48.37	14.09	48.95	4.92	50.31	300.20	3	8099	Decastar
BERNARD	11.02	7.23	14.25	1.92	48.93	14.99	40.87	5.32	62.77	280.10	4	8067	Decastar
YURKOV	11.34	7.09	15.19	2.10	50.42	15.31	46.26	4.72	63.44	276.40	5	8036	Decastar
Sebrle	10.85	7.84	16.36	2.12	48.36	14.05	48.72	5.00	70.52	280.01	1	8893	OlympicG
Clay	10.44	7.96	15.23	2.06	49.19	14.13	50.11	4.90	69.71	282.00	2	8820	OlympicG
Karpov	10.50	7.81	15.93	2.09	46.81	13.97	51.65	4.60	55.54	278.11	3	8725	OlympicG
Macey	10.89	7.47	15.73	2.15	48.97	14.56	48.34	4.40	58.46	265.42	4	8414	OlympicG
Warners	10.62	7.74	14.48	1.97	47.97	14.01	43.73	4.90	55.39	278.05	5	8343	OlympicG

Example of PCA





AGROCAMPUS RENNES

FACTOMINER

1.1 – Introduction of supplementary information

Objective: enrich the interpretation of the dimensions

Based on: the transition formulas

What type of information?

- supplementary individuals
- quantitative supplementary variables
- qualitative supplementary variables

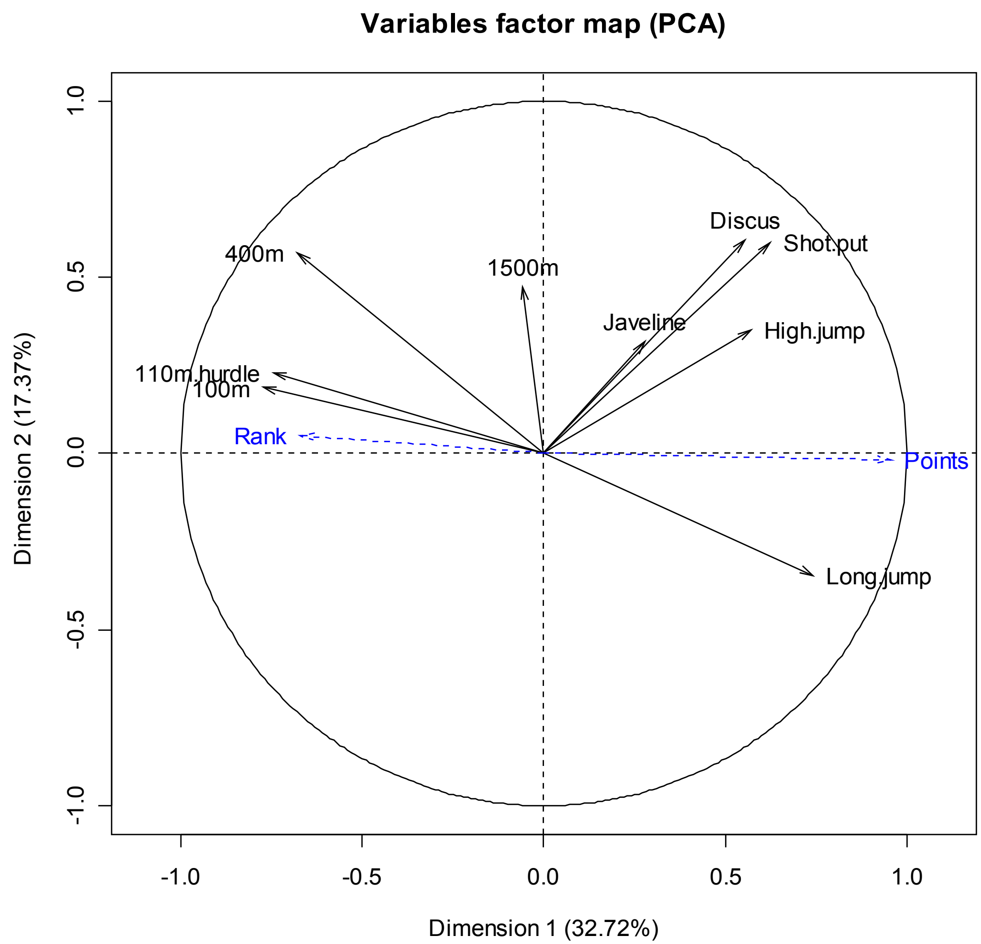
1.1 – Introduction of supplementary information

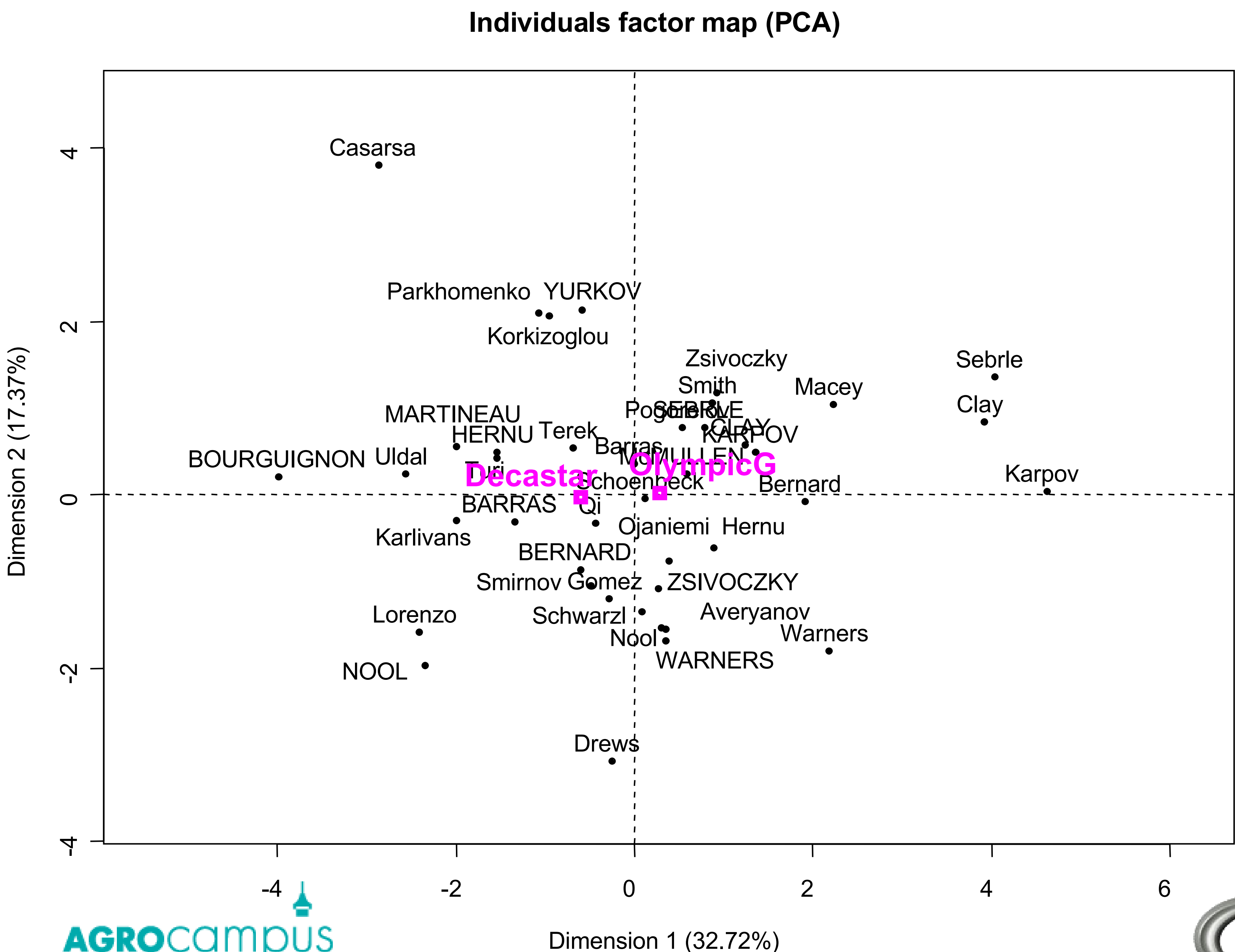
Transition formulas can be applied to supplementary information:

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

$$G_s(k') = \frac{1}{\sqrt{\lambda_s}} \sum_l x_{ik'} F_s(i)$$

Note that supplementary information do not intervene in the construction of the axes





Individuals factor map (PCA)
Casarsa
Parkhomenko
YURKOV
Korkizoglou
Zsivoczky
Sebrle
Smith
Macey
Clay
MARTINEAU
HERNU
Terek
KARPOV
BOURGUIGNON
Uldal
Barras
Decastar
OlympicG
Schoenbeck
Bernard
Karpov
BARRAS
Qi
Ojaniemi
Hernu
Karlivans
BERNARD
Smirnov
Gomez
ZSIVOCZKY
Lorenzo
Schwarzl
Averyanov
Warners
Nool
WARNERS
NOOL
Drews
Dimension 2 (17.37%)
Dimension 1 (32.72%)
-4
-2
0
2
4
6

1.2 – Other helps for the interpretation

- Graphs can be enriched:
  - with colors when adding supplementary information
  - variables can be represented according to their quality of representation

- Indicators are available:
  - Contribution of the individuals and the variables to the construction of the axes
  - Quality of representation of the individuals and of the variables

"Enable effective and rapid exploration of data"
John Chambers' mission

1.3 – Description of the dimensions

➢ **By the quantitative variables:**

• The correlation between each variable and the coordinate of the individuals on the axis *s* is calculated

• The correlation coefficients are sorted

• Only the significant correlations are given

Describes the best the 1st dimension →

```
$Dim.1
$Dim.1$quanti
              Dim.1
Points         0.96
Long.jump      0.74
Shot.put       0.62
Rank          -0.67
400m          -0.68
110m.hurdle   -0.75
100m          -0.77
```

```
$Dim.2
$Dim.2$quanti
              Dim.2
Discus         0.61
Shot.put       0.60
```

Significant level = 0.05

1.3 – Description of the dimensions

➢ By the qualitative variables:

• Perform a one-way analysis of variance with the coordinates of the individuals on the axis explained by the qualitative variable

• For each category, a student *T*-test is used to compare the average of the category with the general mean

• The p-value is transformed to a normal quantile to know if the mean of the category is significantly less or greater than 0

```
$Dim.1$quali
            Dim.1
OlympicG   1.43
Decastar  -1.43
```

```
$Dim.2$quali
[1] Dim.2
<0 rows> (or 0-length row.names)
```

Significant level = 0.2

2 – Structure on the data

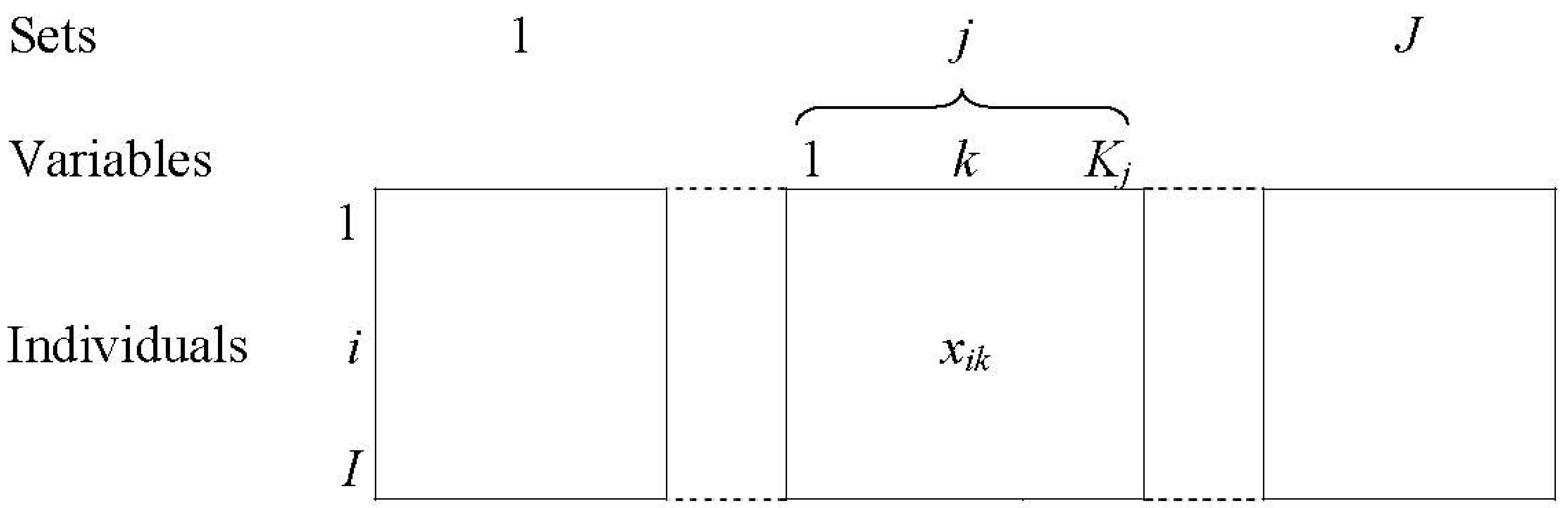
Different structure on the data are proposed:

- partition on the variables: several sets of variables are simultaneously studied: Multiple Factor Analysis, Generalized Procrustes Analysis
- a hierarchy on the variables: variables are grouped and subgrouped (like in questionnaires structured in topics and subtopics): Hierarchical Multiple Factor Analysis
- a partition on the individuals: several sets of individuals described by the same variables: Dual Multiple Factor Analysis

AGROCAMPUS RENNES

FACTOMINER

Groups of variables (MFA)



The groups of variables can be quantitative and/or qualitative

Groups of variables (MFA)

Examples :

- genomic: protein, DNA
- sensory analysis: sensorial, physico-chemical
- questionnaires: student health (addicted consumptions, psychological conditions, sleep, identification, …)

- comparison of coding (quantitative, qualitative)

Groups of variables (MFA)

Taking into account the structure on the data allows to:

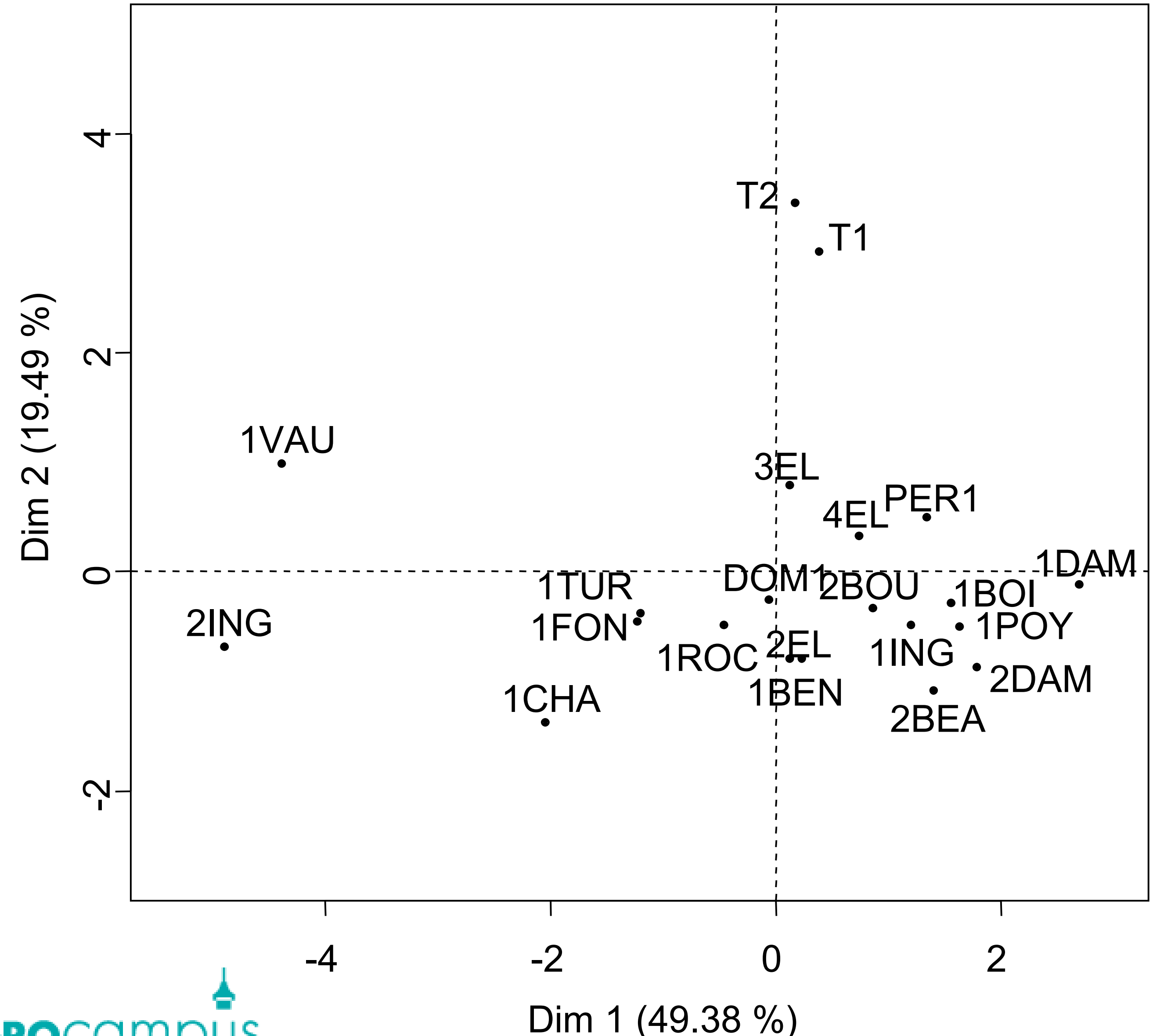
- Balance the influence of each group of variables
- Study the link between the sets of variables
- Give the classical graphs but also specific graphs:
  - Partial representation (individuals seen by one group of variables)
  - Groups of variables

Example of MFA: the data

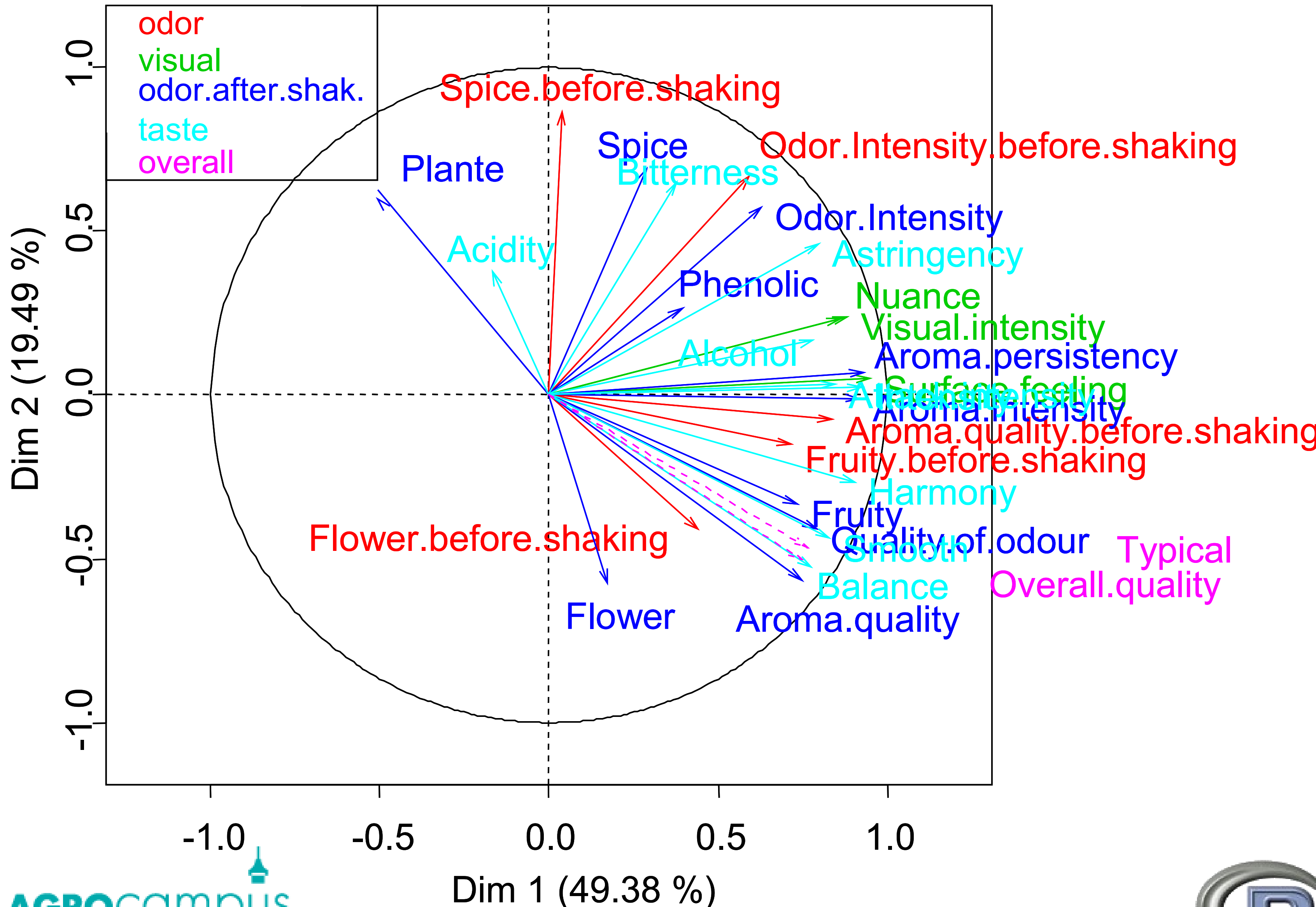
The data: a panel of experts described 21 wines with 29 sensory descriptors + 2 qualitative variables (the origin and the soil)

	Qlt	$\underbrace{\hphantom{xxxxxxxxxxxxxxxxxxxx}}$ Quantitative groups of variables				
	Origin (2)	Odor (5)	Visual (3)	Taste (10)	Odor after shaking (9)	Assessment (2)
Wine 1						
Wine 2						
…						
Wine 21						

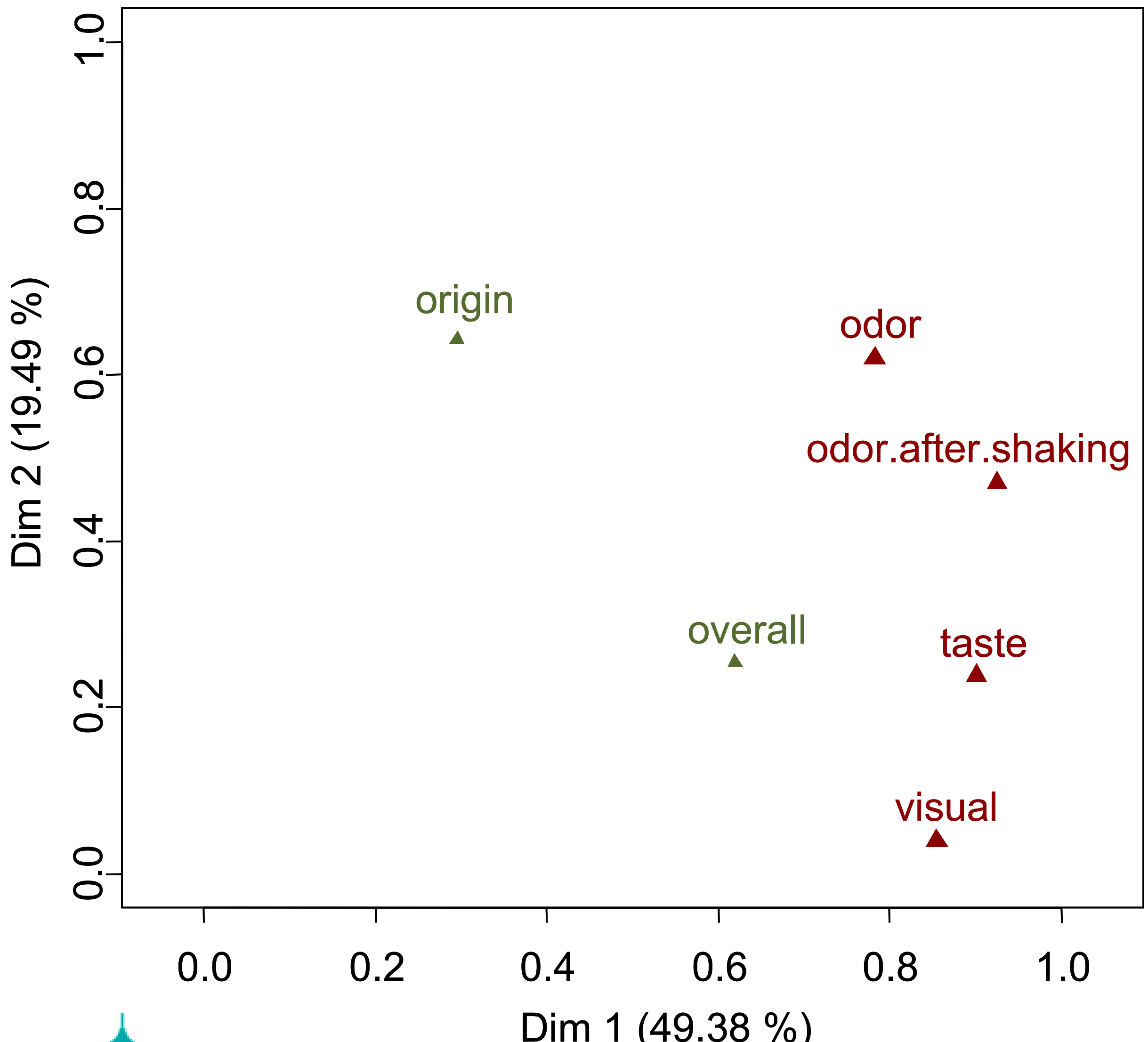
Example of MFA: representation of the individuals



Example of MFA: representation of the variables



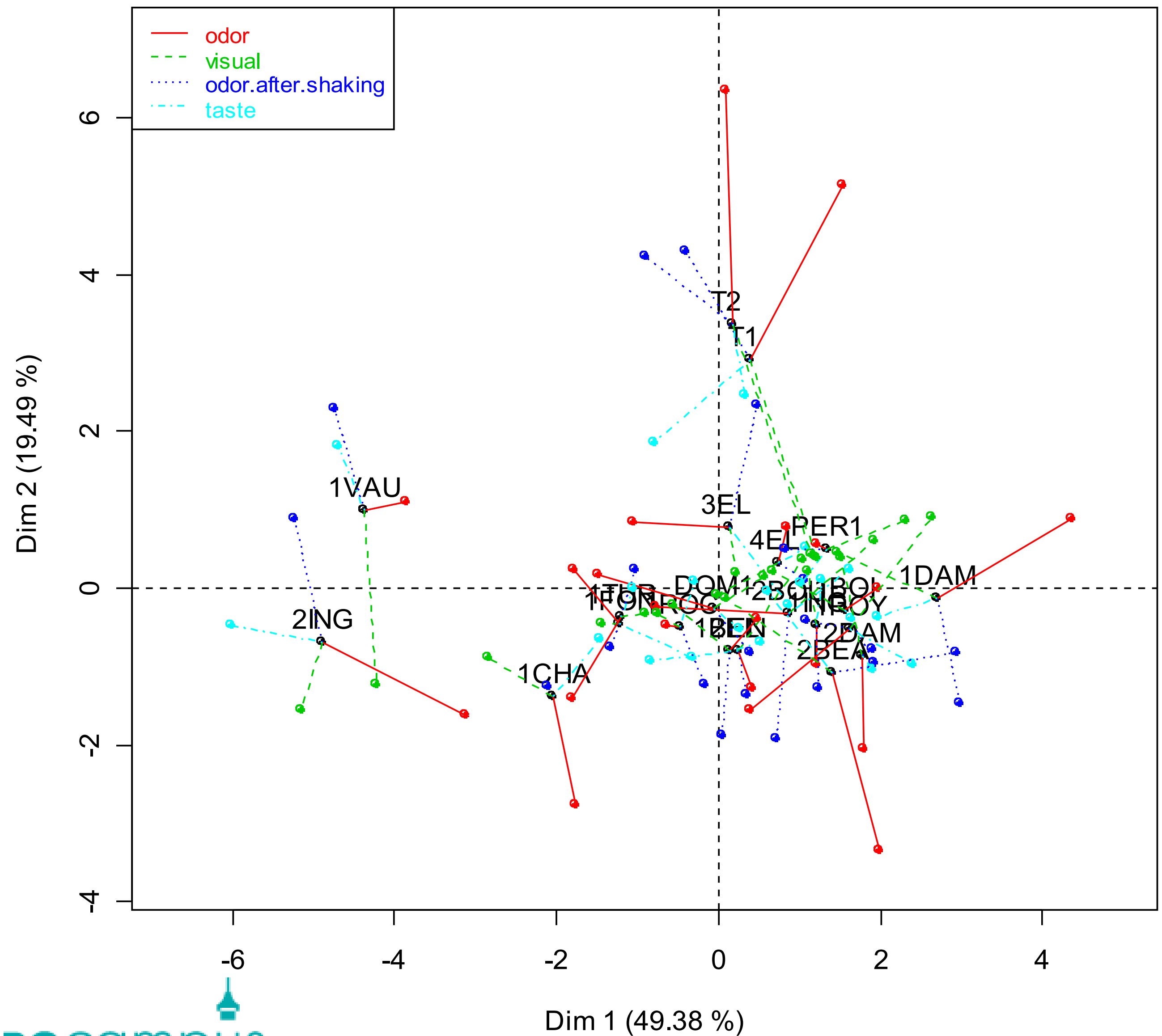
Example of MFA: representation of the groups



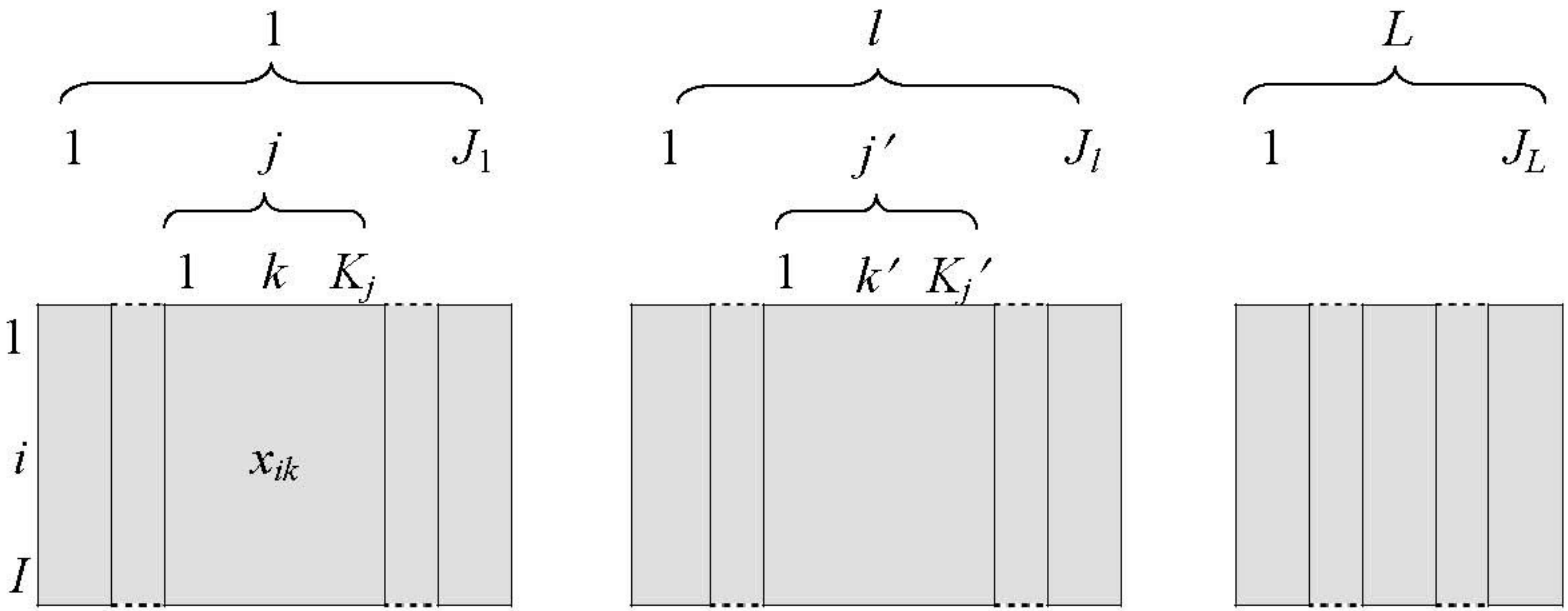
AGROcampus RENNES

FACTOMINER

Example of MFA: representation of the partial points



Hierarchy on the variables (HMFA)



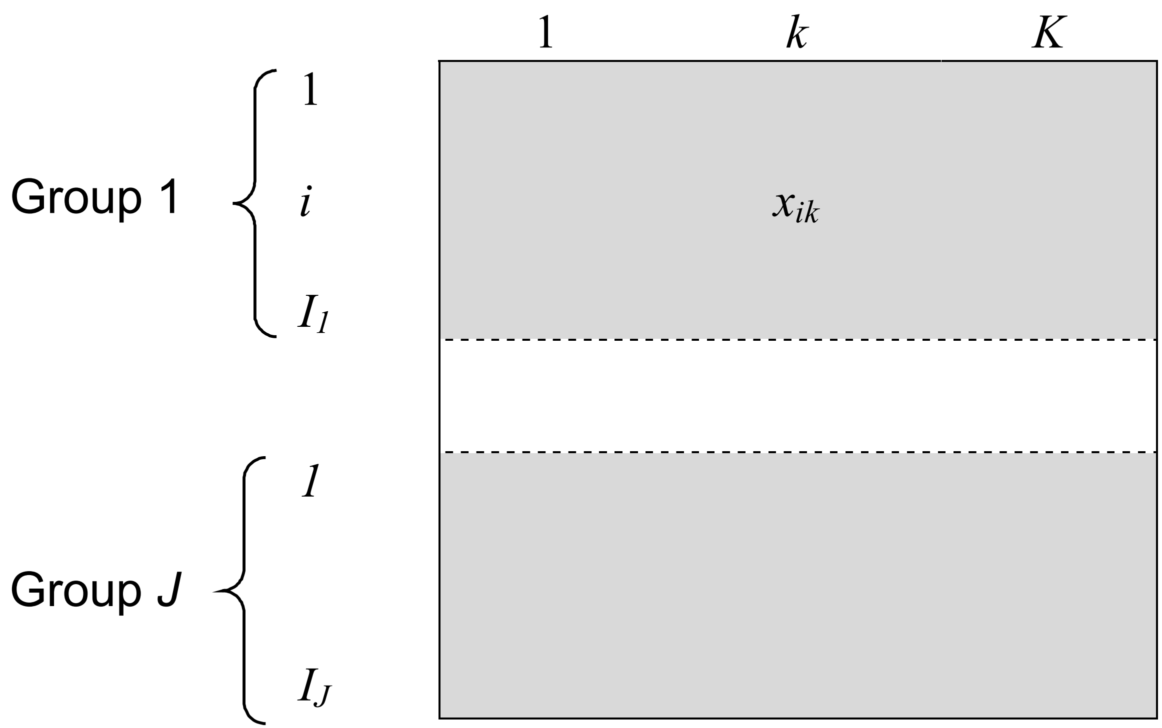
Two levels for the hierarchy: the first one contains L groups,
each l group contains J_l subgroups, and each subgroup have K_j variables

Objective: to balance the groups and the subgroups of variables

AGROcampus RENNES

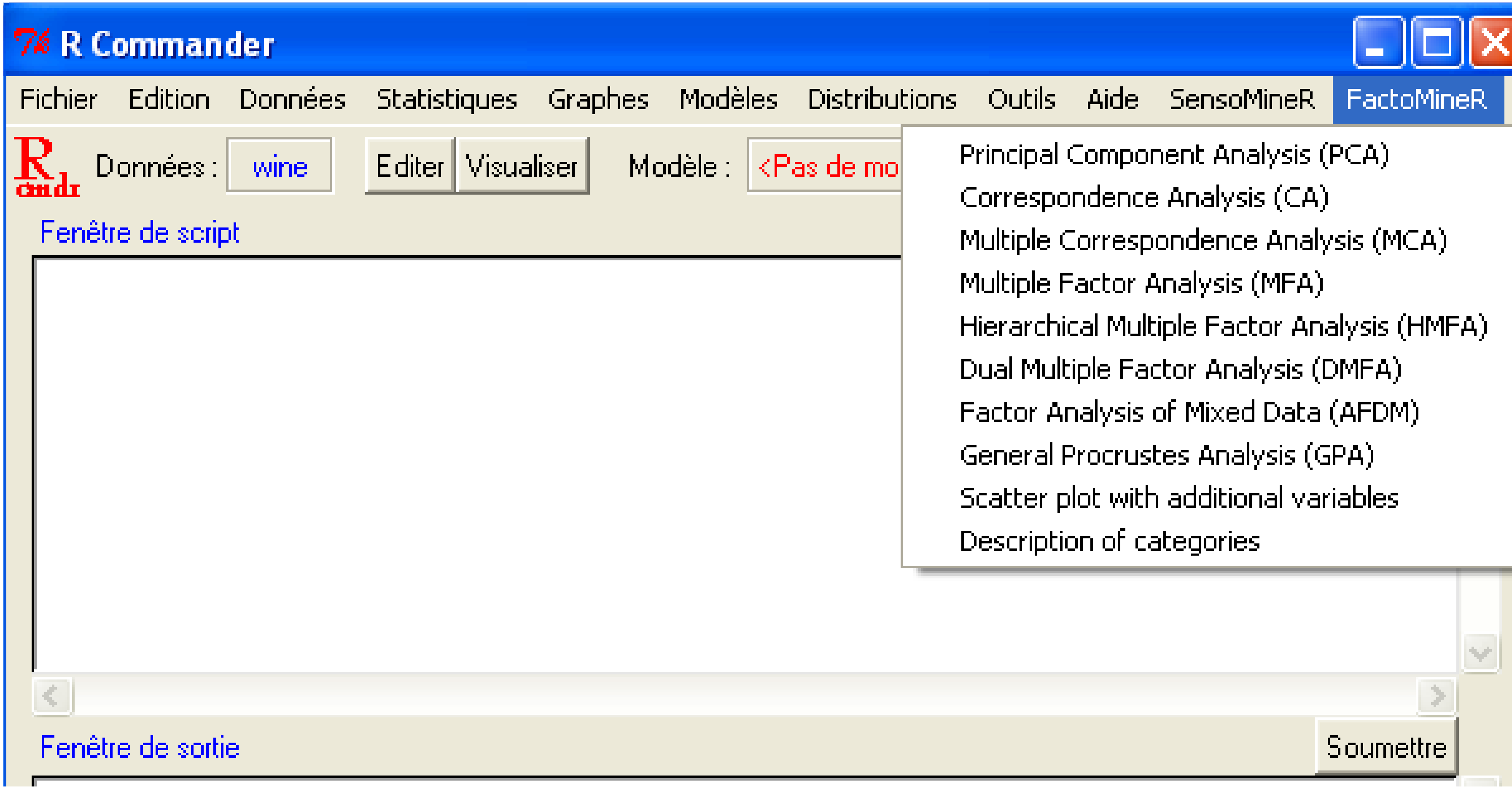
FACTOMINER

Partition on the individuals (DMFA)



Objective: to compare the covariance matrices

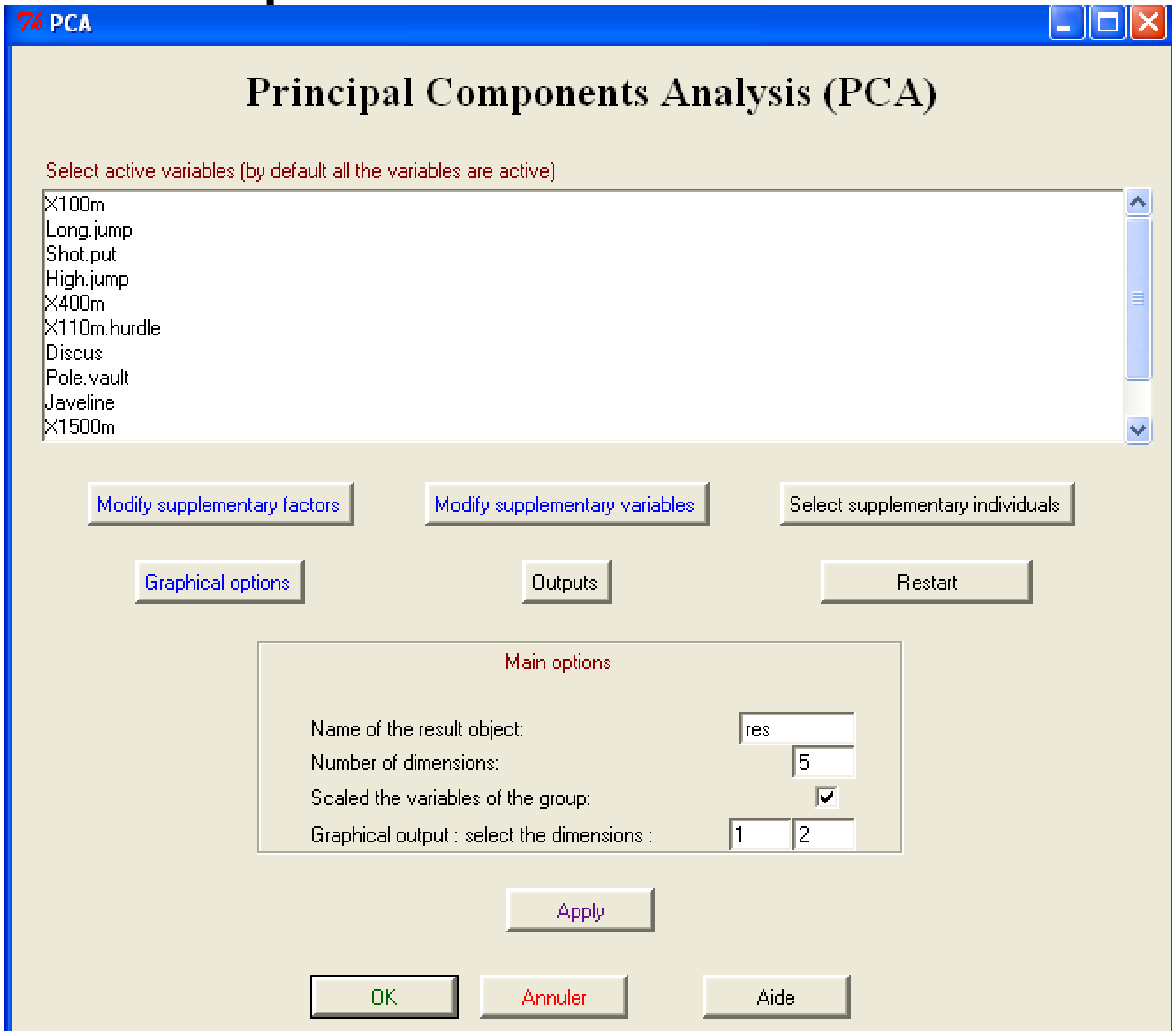
3 – Graphical User Interface



Menu of the FactoMineR GUI

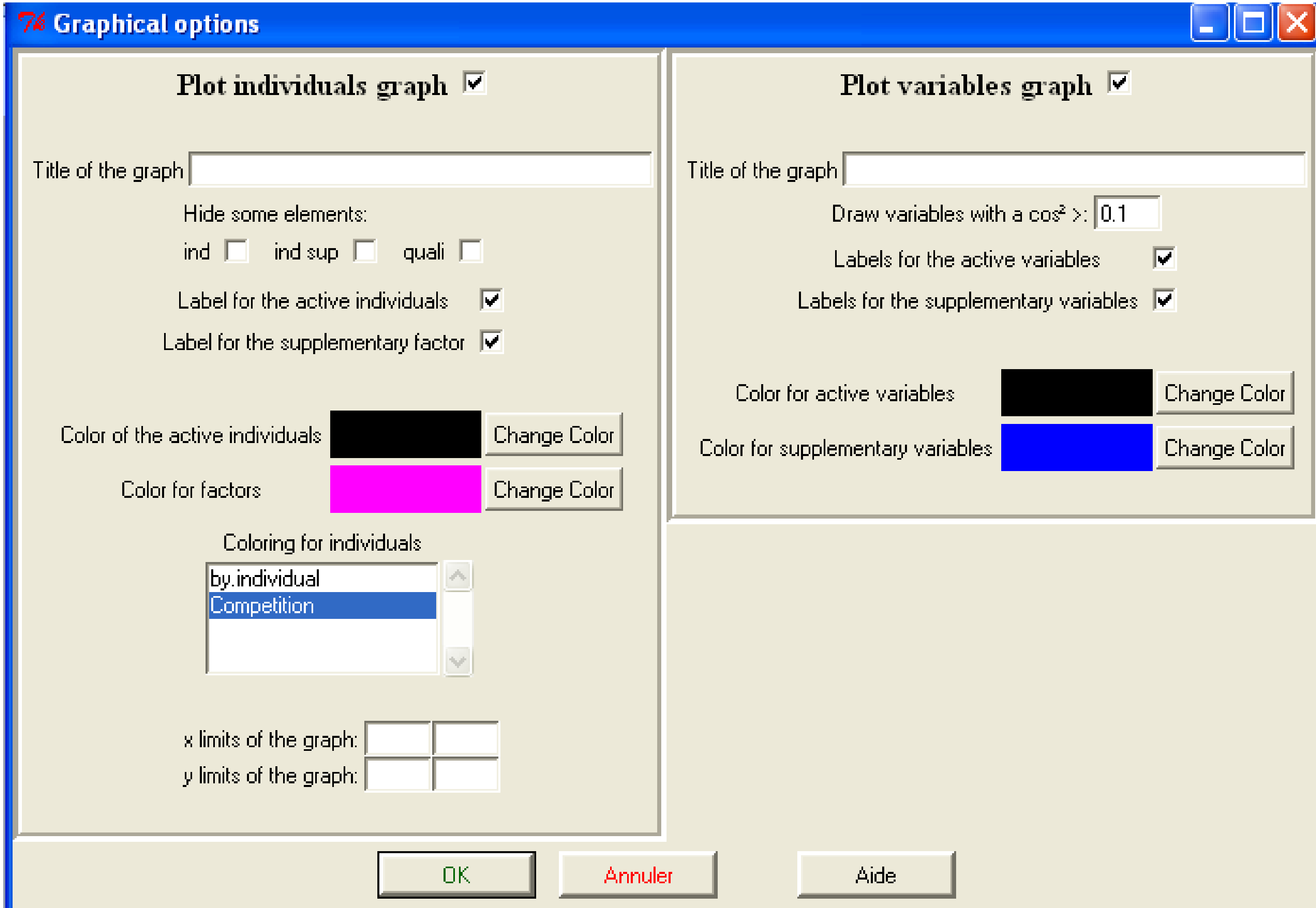
3 – Graphical User Interface

Main window of the PCA



3 – Graphical User Interface

Graphical options



4 – Conclusion

For researchers, practitioners and students: with classical and advanced methods

The FactoMineR package is available on the CRAN

The GUI can be simply loaded:

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

A website is dedicated to this package: http://factominer.free.fr

Future: dynamical graphs

The team of FactoMineR is nearly all the time ready to improve the package