From the package FactoMineR to a project on exploratory multivariate analysis

or how to improve the visibility of its R package

François Husson

Department Statistics & Computer science, Agrocampus Ouest
husson@agrocampus-ouest.fr

Murcia, 22 de noviembre de 2018 – X Jornadas de usuarios de R

Plan

Introduction

The construction of a package allows to:

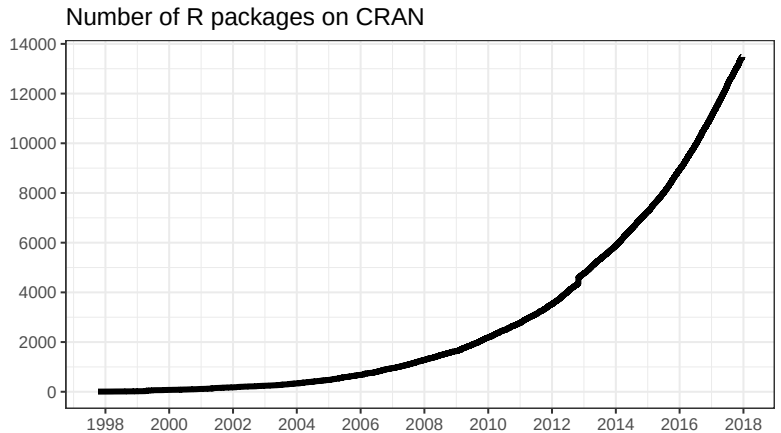
- propose new statistical methods or methodological approaches
- share its work with the entire scientific community
- facilitate the comparison of methods
- make data sets available

The creation of a package is time-consuming: it MUST be beneficial to the package author AND the scientific community

More and more packages

November 16, 2018:

- CRAN : 13 403 packages
- Bioconductor : 2 955 packages
- R-Forge : 2 086 projects
- GitHub: ??? projects



$\Longrightarrow$ the visibility of a package is increasingly limited

Many packages are unused... and therefore useless!

Plan

Before the package submission

- A package to do what?
  - what contribution compared to existing packages?
  - is it possible to propose a function to the authors of another package?

- A package for who?
  - for a few researchers in the field $\Longrightarrow$ GitHub or web page
  - for a large audience $\Longrightarrow$ CRAN or bioconductor (GitHub)

Before the package submission

The first version of the package may be limited, but what is done must be done well

- the package will evolve and some choices are difficult to modify
  - the name of the package
  - the names of the main functions
  - the default arguments
- users won't use the package if they don't understand how it works
  - properly document his functions
  - choose your examples carefully
  - make a vignette ("first steps guide")

After the package submission

"Make alive" and maintain the package:

- fix the package if there are some errors
- answer users' questions
- include new developments, new options
- improve programming (Rcpp, parallelization, etc.)

Build additional packages

Make the package known

Plan

FACTOMINER in a few words

The package

- allows to explore and visualize data sets
- offers principal component methods and clustering methods
- gives many indicators (quality of representation, contribution, automatic description of dimensions,...)
- possibility to add additional elements
- graphical interface (in French and English)
- missing data management (with the missMDA package)
- user assistance (website, videos)
- course on the methods (books, MOOC)

FACTOMINER in a few words

Different methods for different data formats:

Data	Methods	Function
Quantitative variables	Principal Component Analysis	PCA
Contingency table	Correspondence Analysis	CA
Qualitative variables	Multiple Correspondence Analysis	MCA
Mixed data	Factor Analysis for Mixed Data	FAMD
Variable groups	Multiple factorial Analysis	MFA
Hierarchy on variables	Hierarchical Multiple Factor An.	HMFA
Groups of individuals	Dual Multiple Factor Analysis	DMFA
Contingency Table and Contextual Variables	Generalized Correspondence Analysis On Generalised Aggregated Lexical Table	CaGalt

Clustering methods and complementary tool methods:

Methods	Function
Hierarchical Ascendant Clustering	HCPC
Description of a qualitative variable (e.g. cluster var.)	catdes
Description of a quantitative variable (e.g. a dimension)	condes, dimdesc

Example on a sensory description of wines

- 10 white wines from Val de Loire: 5 Vouvray - 5 Sauvignon
- sensory descriptor: acidity, bitterness, aroma intensity, etc.

Wine data set

- 10 individuals (rows): white wines from Val de Loire
- 30 variables (columns):
  - 27 continuous variables: sensory descriptors
  - 2 continuous variables: odour and overall preferences
  - 1 categorical variable: label of the wines (Vouvray - Sauvignon)

	O.fruity	O.passion	O.citrus	...	Sweetness	Acidity	Bitterness	Astringency	Aroma.intensity	Aroma.persistency	Visual.intensity	Odor.preferene	Overall.preference	Label
S Michaud	4.3	2.4	5.7	...	3.5	5.9	4.1	1.4	7.1	6.7	5.0	6.0	5.0	Sauvignon
S Renaudie	4.4	3.1	5.3	...	3.3	6.8	3.8	2.3	7.2	6.6	3.4	5.4	5.5	Sauvignon
S Trotignon	5.1	4.0	5.3	...	3.0	6.1	4.1	2.4	6.1	6.1	3.0	5.0	5.5	Sauvignon
S Buisse Domaine	4.3	2.4	3.6	...	3.9	5.6	2.5	3.0	4.9	5.1	4.1	5.3	4.6	Sauvignon
S Buisse Cristal	5.6	3.1	3.5	...	3.4	6.6	5.0	3.1	6.1	5.1	3.6	6.1	5.0	Sauvignon
V Aub Silex	3.9	0.7	3.3	...	7.9	4.4	3.0	2.4	5.9	5.6	4.0	5.0	5.5	Vouvray
V Aub Marigny	2.1	0.7	1.0	...	3.5	6.4	5.0	4.0	6.3	6.7	6.0	5.1	4.1	Vouvray
V Font Domaine	5.1	0.5	2.5	...	3.0	5.7	4.0	2.5	6.7	6.3	6.4	4.4	5.1	Vouvray
V Font Brûlés	5.1	0.8	3.8	...	3.9	5.4	4.0	3.1	7.0	6.1	7.4	4.4	6.4	Vouvray
V Font Coteaux	4.1	0.9	2.7	...	3.8	5.1	4.3	4.3	7.3	6.6	6.3	6.0	5.7	Vouvray

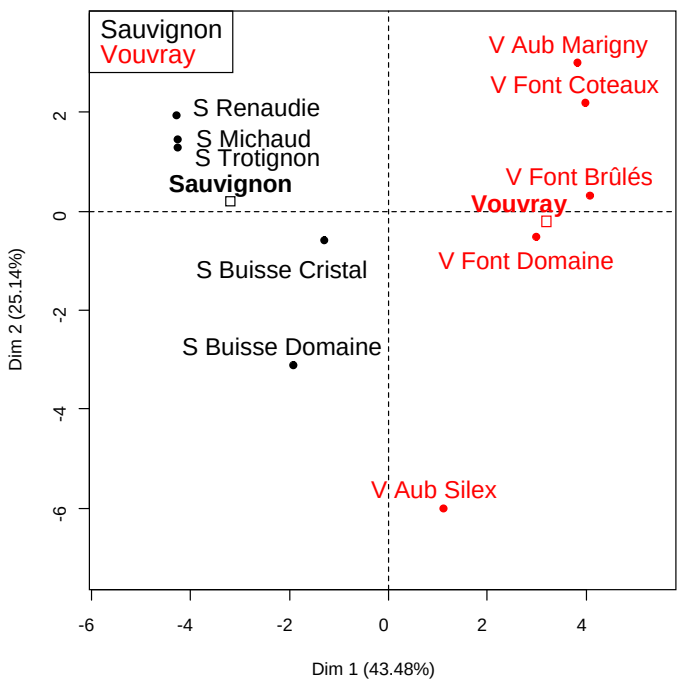
Description of the wines by the experts

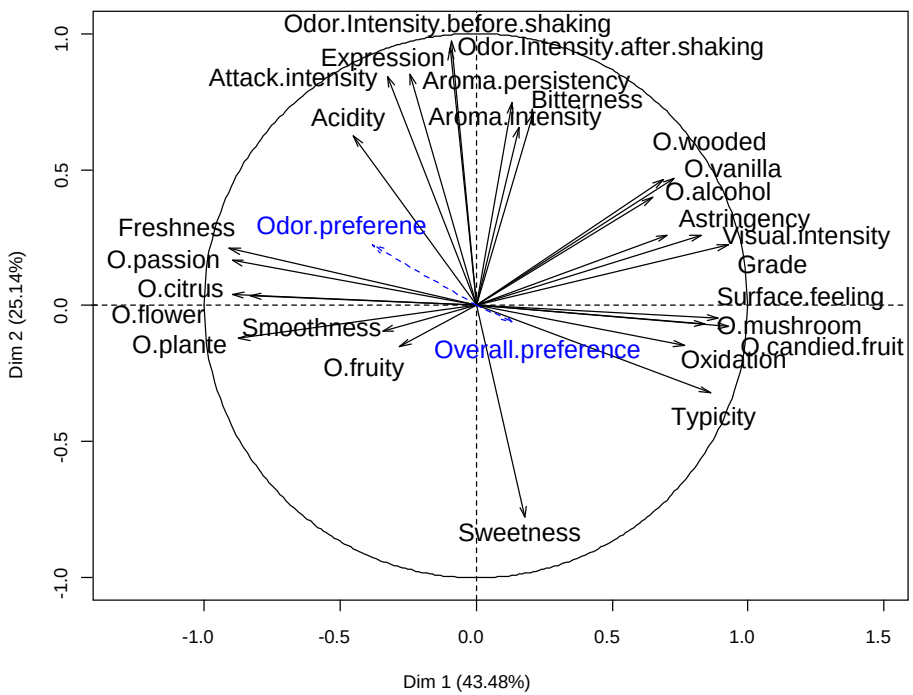
- PCA performed with supplementary information

```
res.pca <- PCA(Expert, quanti.sup=29:30, quali.sup=1)
```

- an algorithm optimize the label positions to limit overlap
- individuals can be coloured according to a categorical variable

```
plot(res.pca, habillage = 1)
```





Comparison of sensory panels with MFA

Continuous variables

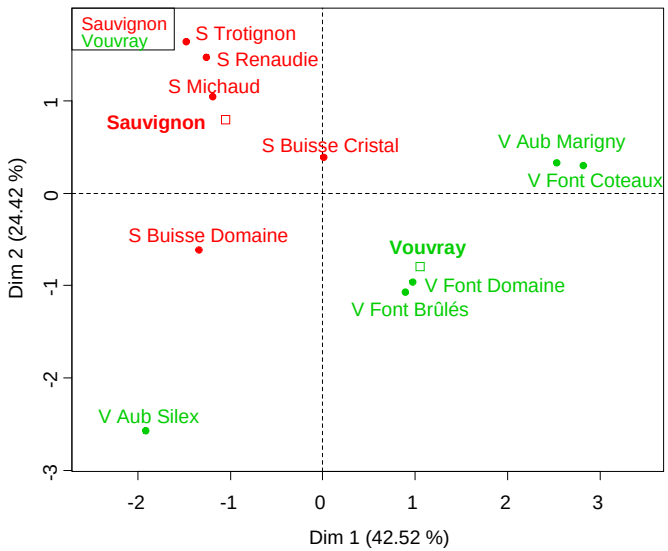
Categorical

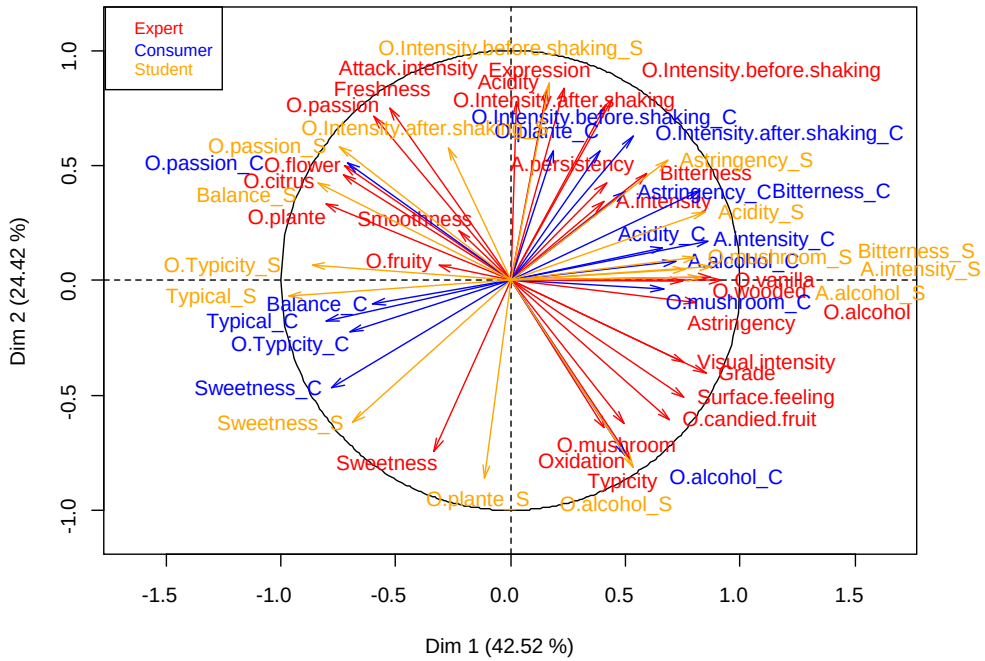
	Expert (27)	Consumer (15)	Student (15)	Preference (60)	Label (1)
wine 1					
wine 2					
...					
wine 10					

Objectives:

- How are the products described by the panels?
- Do the panels describe the products in a same way? Is there a specific description done by one panel?

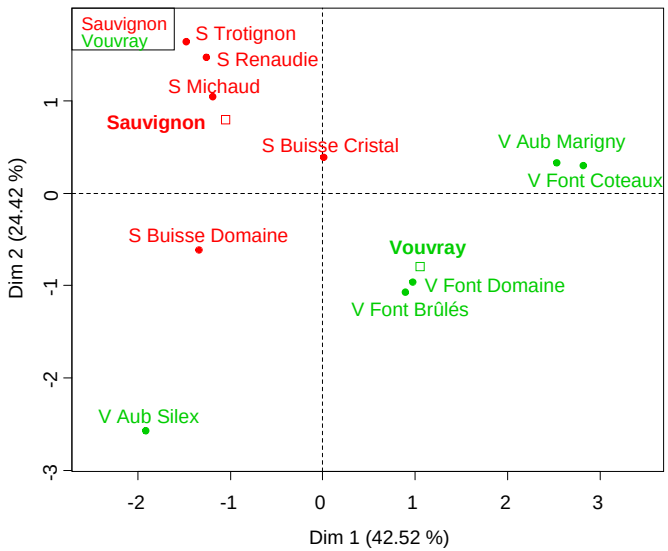
Representation of the individuals and of the variables

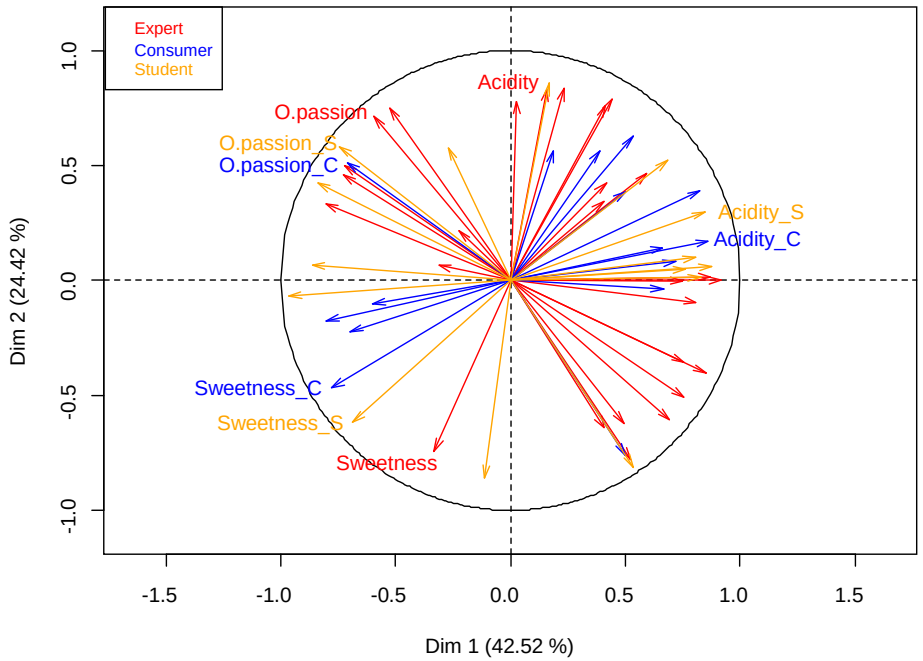




Same interpretation as in PCA

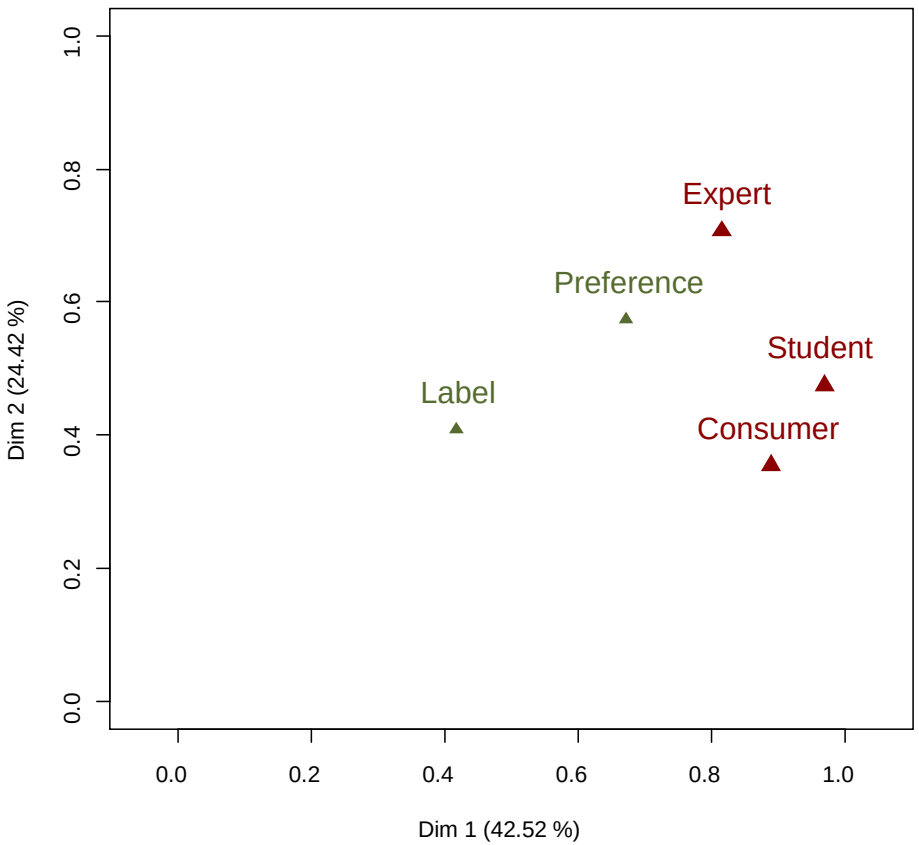
Representation of the individuals and of the variables





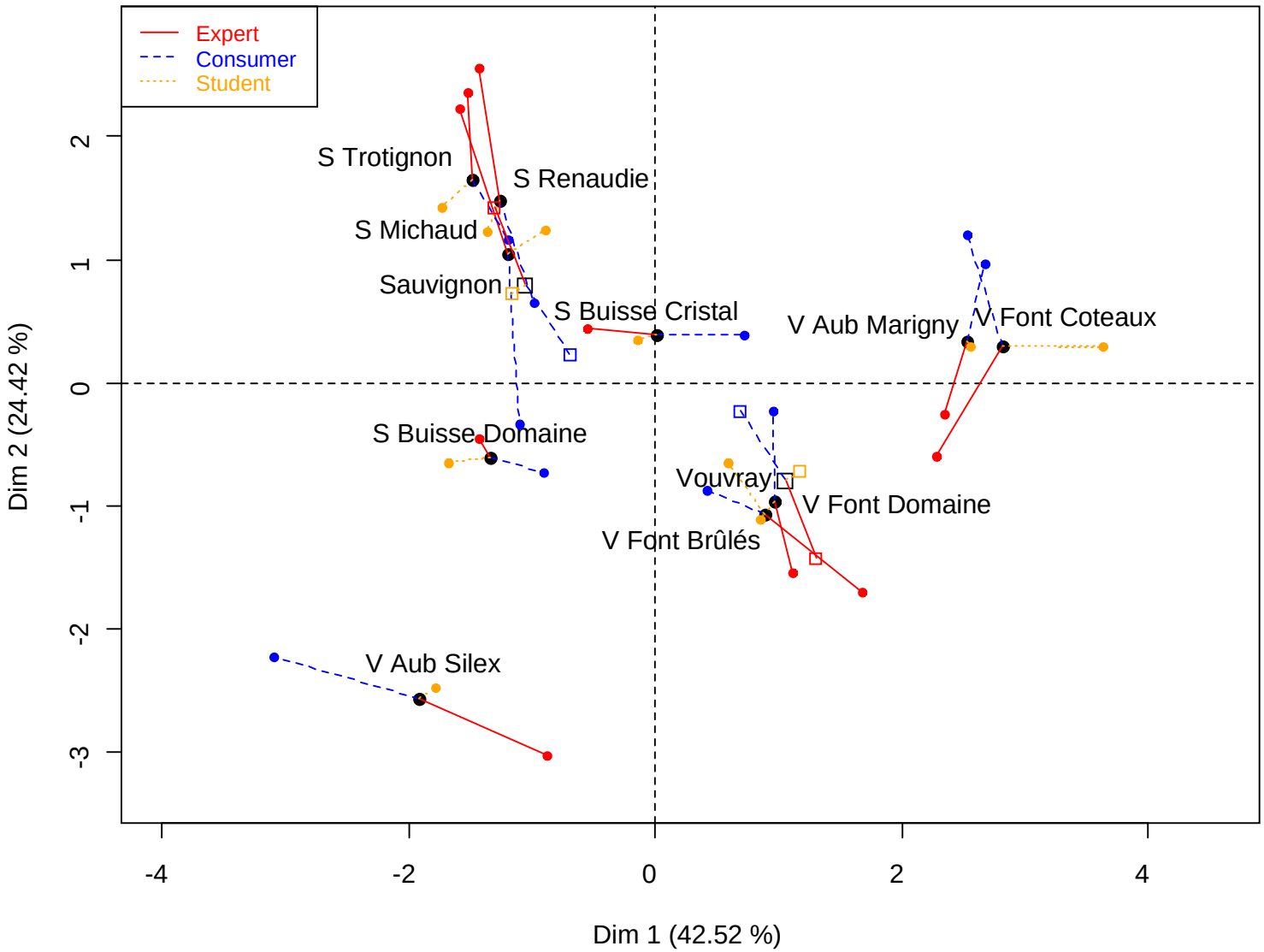
Same interpretation as in PCA

Representation of the groups



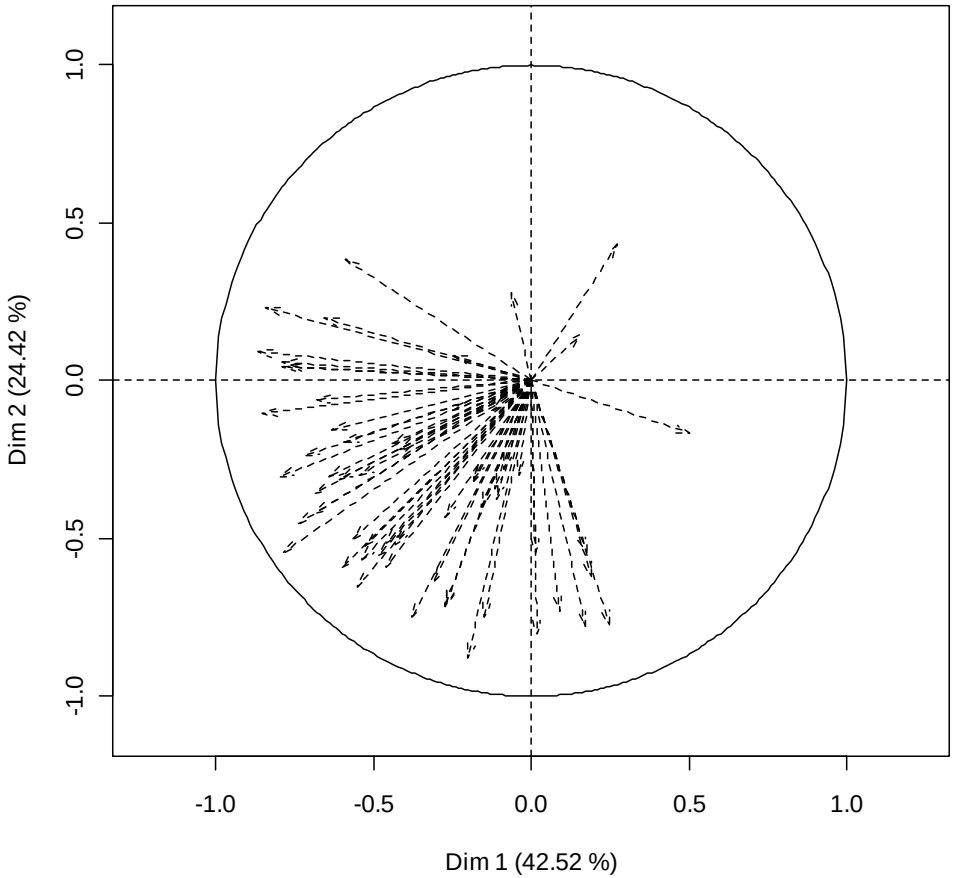
- 2 groups are all the more close that they induce the same structure
- The 1st dimension is common to all the panels
- 2nd dimension mainly due to the experts
- Preference linked to sensory description

Representation of the partial points



Representation of supplementary continuous variables

The favourite wine is
Vouvray Aubussière Silex



Preferences are linked to
sensory description

Plan

Supplementary packages

- `RcmdrPlugin.FactoMineR` : drop-down menu
- `Factoshiny` : GUI and interactive graphs

$\Longrightarrow$ make it easier for users to use the packages

- `FactoInvestigate`: automatic reporting

$\Longrightarrow$ propose an interpretation of the results

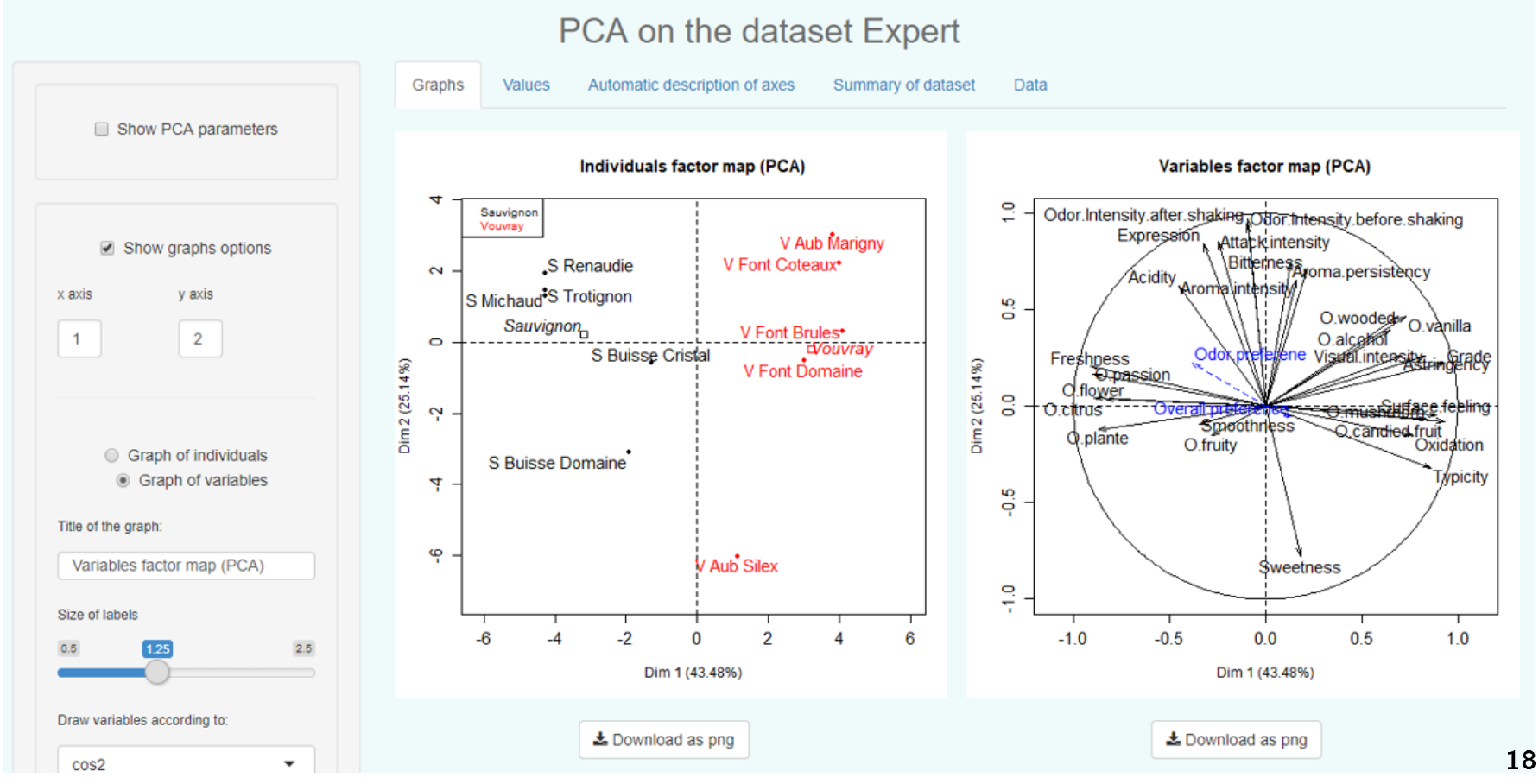
- `missMDA`: handling missing values

$\Longrightarrow$ go further than the standard methods of the package

GUI and interactive graphics with Factoshiny

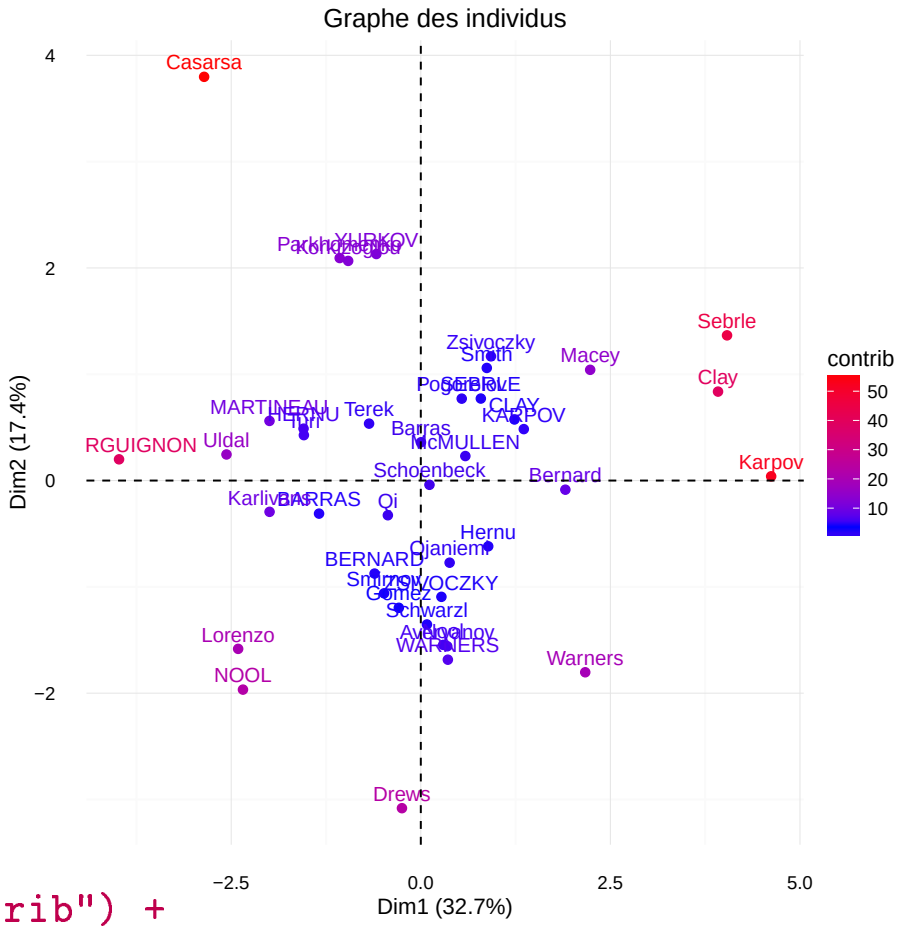
- Perform analyses without the need to master the code
- Real-time visualization of changes made

```
res <- PCAshiny(MyData)    ## principal component analysis on the data
res <- PCAshiny(res.pca)   ## graph on a result object of FactoMineR
res2 <- PCAshiny(res)      ## Factoshiny result object
```



Other graphical packages

- the package `explor`
  - interactive graphics
  - possibility to move the labels
- the package `factoextra`
  - based on `ggplot2`
  - sequential construction of graphs by adding layers




```
> library(factoextra)
> fviz_pca_ind(res.pca, col.ind="contrib") +
   labs(title="Graphe des individus") +
   scale_color_gradient2(mid="blue",high="red") +
   theme_minimal()
```

Automatic reporting with the FactoInvetigate package

Propose an interpretation of the results based on the result object

```
> res.pca <- PCA(MyData, ...)
> library(FactoInvetigate)
> Investigate(res.pca)
```

http://factominer.free.fr/reporting

Principal Component Analysis

Dataset Expert

This dataset contains 10 individuals and 30 variables, 2 quantitative variables are considered as illustrative, 1 qualitative variable is considered as illustrative.

1. Study of the outliers

The analysis of the graphs does not detect any outlier.

2. Inertia distribution

The inertia of the first dimensions shows if there are strong relationships between variables and suggests the number of dimensions that should be studied.

The first two dimensions of PCA express **68.62%** of the total dataset inertia ; that means that 68.62% of the individuals (or variables) cloud total variability is explained by the plane. This percentage is relatively high and thus the first plane well represents the data variability. This value is strongly greater than the reference value that equals **44.91%**, the variability explained by this plane is thus highly significant (the reference value is the 0.95-quantile of the inertia percentages distribution obtained by simulating 1983 data tables of equivalent size on the basis of a normal distribution).

From these observations, it should be better to also interpret the dimensions greater or equal to the third one.

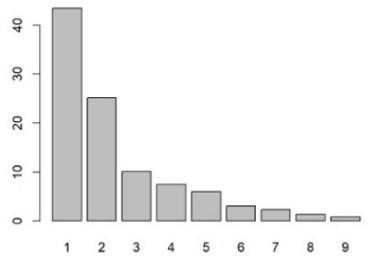
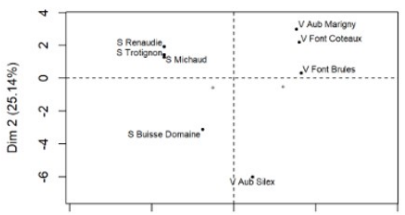


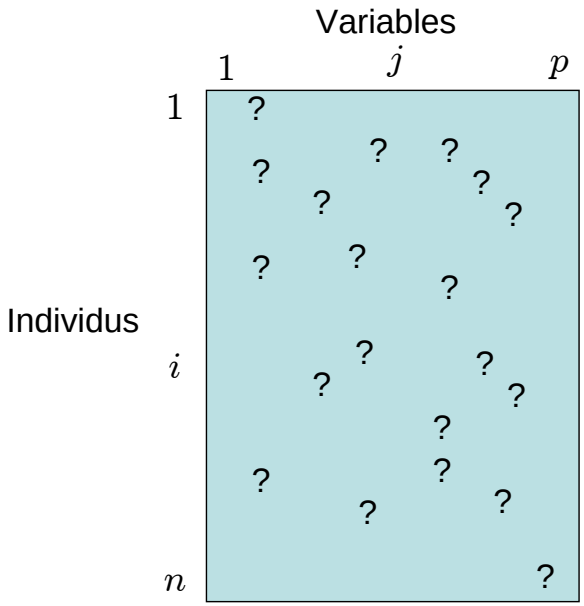
Figure 2 - Decomposition of the total inertia on the components of the PCA

An estimation of the right number of axis to interpret suggests to restrict the analysis to the description of the first 2 axis. These axis present an amount of inertia greater than those obtained by the 0.95-quantile of random distributions (68.62% against 44.91%). This observation suggests that only these axis are carrying a real information. As a consequence, the description will stand to these axis.

3. Description of the plane 1:2



missMDA: a package to handle the missing values



Study and implementation of PC methods in the presence of missing data: PCA, MCA, FAMD, MFA

1. Imputation by iterative principal component method
2. Analysis of the imputed dataset

Regularized iterative PCA

Principle: impute by values that do not influence the PCA results

1. initialization $\ell = 0$: X^0 (mean imputation)
2. iteration ℓ:
   (a) PCA on the completed data set $\rightarrow (F^\ell, U^\ell)$; S dimensions are kept
   (b) missing values imputed with $F^\ell U^{\ell\prime}$
   $\Longrightarrow X^\ell = W * X + (1 - W) * F^\ell U^{\ell\prime}$
3. steps of estimation and imputation are repeated

$\Longrightarrow$ gives the scores and loadings (better than Nipals)
$\Longrightarrow$ gives an imputed data set

Handling missing values: PCA example

```
> library(missMDA)
> data(orange)
> nb <- estim_ncpPCA(orange)          ## Estimate the number of components
> comp <- imputePCA(orange,ncp=nb)    ## Impute the data set
> res.pca <- PCA(comp$completeObs)    ## Perform PCA
```

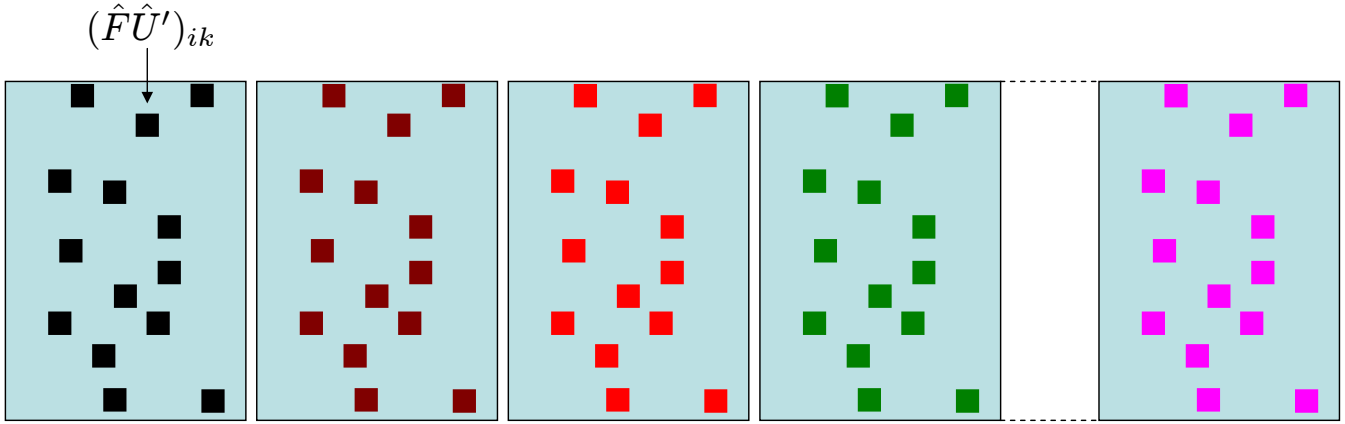
```
> orange
 Sweet Acid Bitter Pulp Typicity
  NA   NA  2.83   NA      5.21
5.46 4.13  3.54 4.62      4.46
  NA 4.29  3.17 6.25      5.17
4.17 6.75    NA 1.42      3.42
...
  NA   NA    NA 7.33      5.25
4.88 5.29  4.17 1.50      3.50
```

```
> comp$completeObs
Sweet Acid Bitter Pulp Typicity
 5.54 4.13   2.83 5.89     5.21
 5.46 4.13   3.54 4.62     4.46
 5.45 4.29   3.17 6.25     5.17
 4.17 6.75   4.73 1.42     3.42
...
 5.71 3.87   2.80 7.33     5.25
 4.88 5.29   4.17 1.50     3.50
```

$\Longrightarrow$ Use the completed data set as an input for PCA, clustering, ...

Multiple imputation in PCA

$\Rightarrow$ iterative PCA: a single imputation method

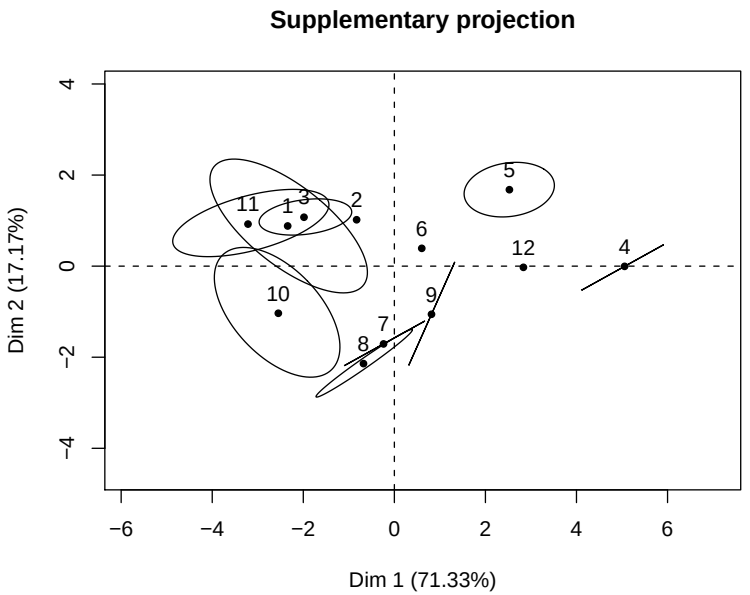


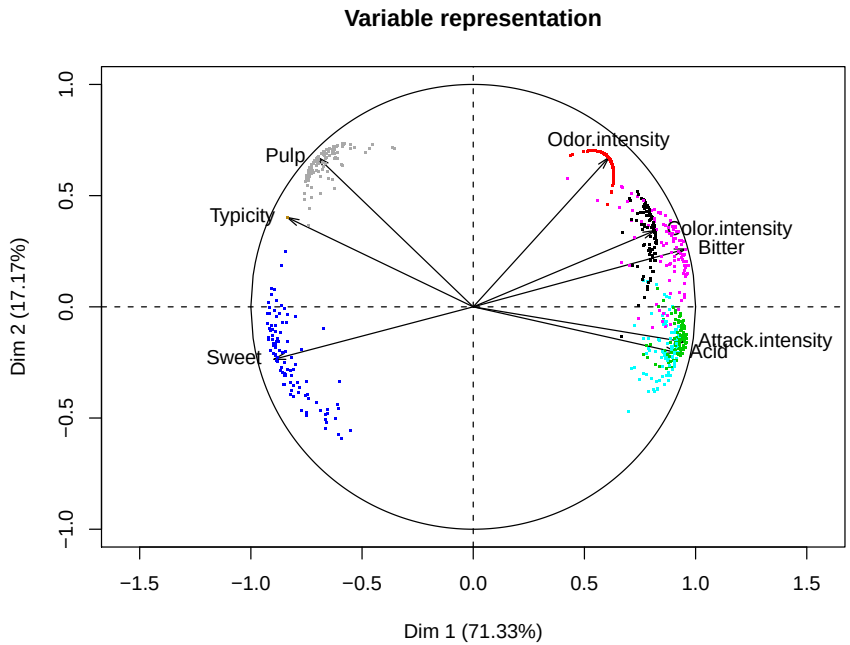
$\Rightarrow$ a single value can't reflect the uncertainty of prediction

$\Rightarrow$ Multiple imputation: generate several plausible values for each missing values

Visualization of the uncertainty due to missing values

```
> mi <- MIPCA(orange, scale = TRUE, ncp=2)
> mi$res.MI              ## output for the imputed data sets
> plot(mi)
```





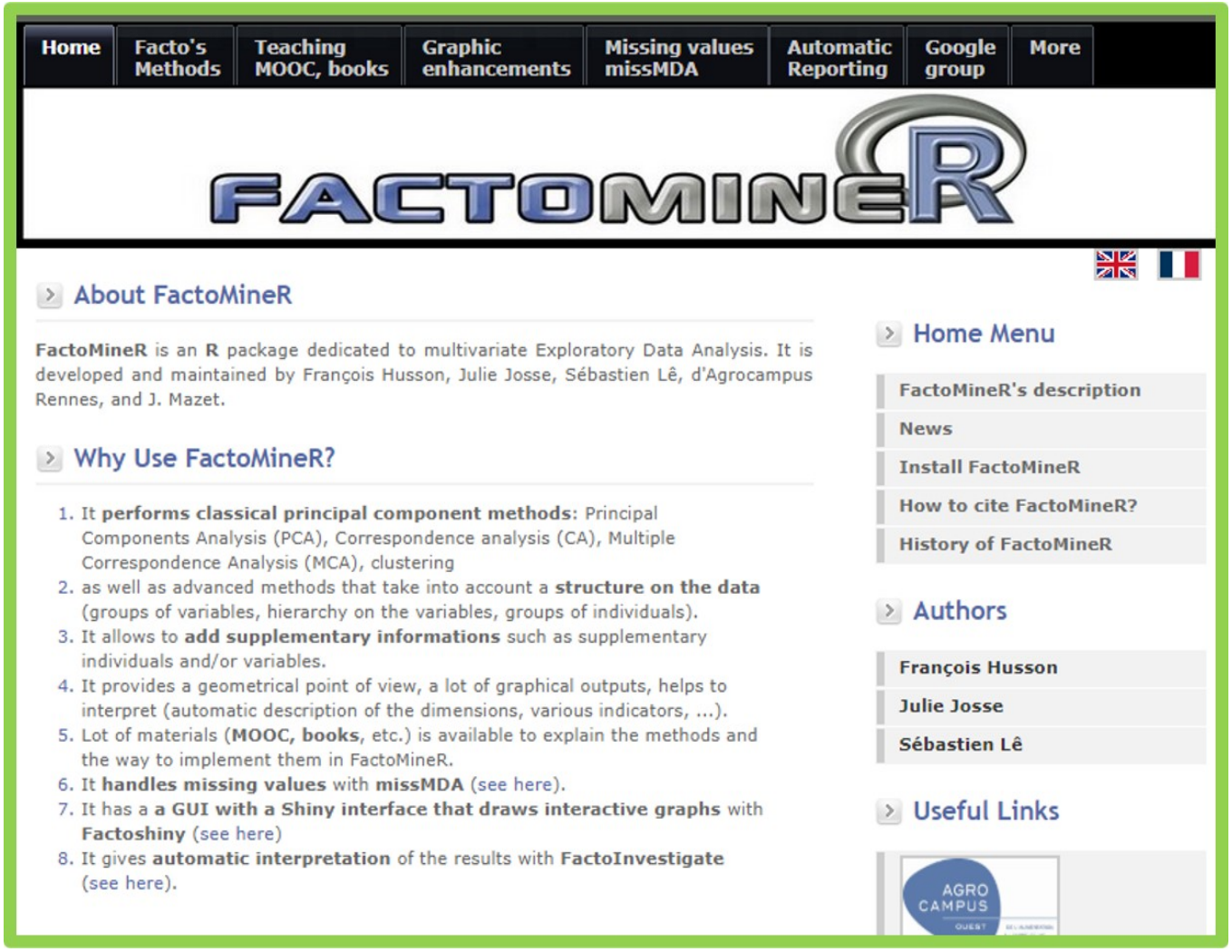
Avoid that you analyse data sets with too many missing data

For categorical data (MCA): estim_ncpMCA, imputeMCA, MIMCA
For mixed data (FAMD): estim_ncpFAMD, imputeFAMD, MIFAMD
For groups of variables (MFA): estim_ncpMFA, imputeMFA

Plan

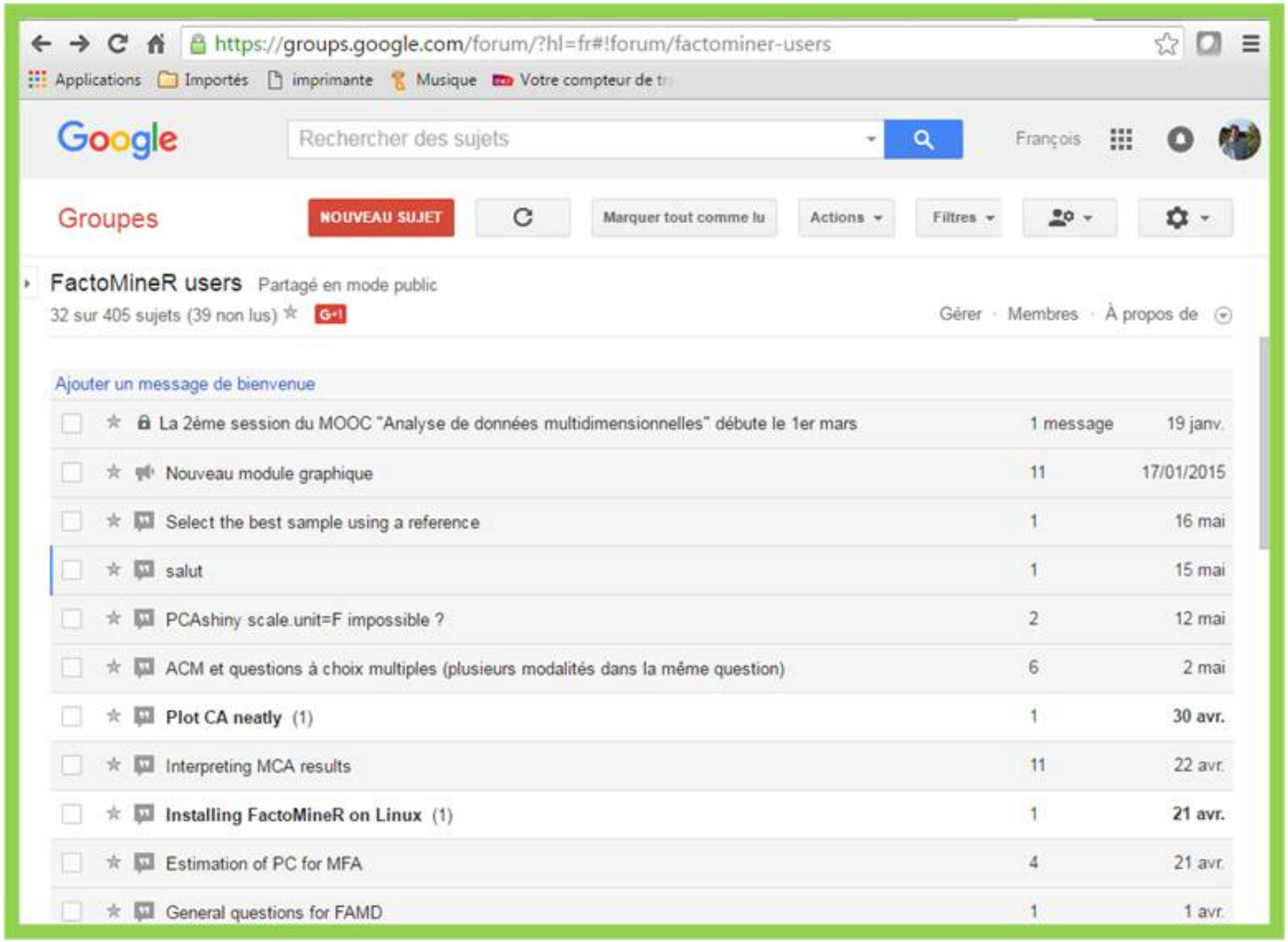
User aids: website

- http://factominer.free.fr in English and in French
- examples, aids on the functions, course, references, etc.



User aids: a Google group

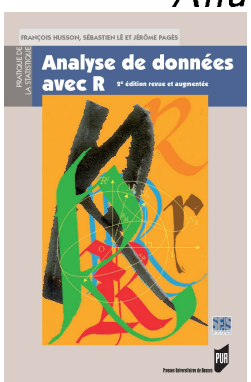
- https://groups.google.com/group/factominer-users/
- possibility to ask questions and answer in French or English

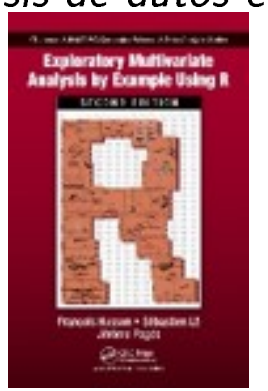


User aids: scientific dissemination

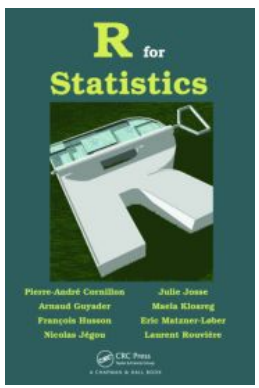
- Suggest the package in CRAN Task views
- Presentations to conferences (useR!, juR, SFdS, etc.)
- 2 papers in *R journal* (CA-galt, MFACT)
- 2 papers in *J. of stat. software* (FactoMineR, missMDA)
- Some books in French, English and Spanish

Análisis de datos con R





R pour la stat. et sc. des données



Plan

User aids: some videos

- available from my teaching webpage
- available from the helps of the functions of FactoMineR

Analyse de données avec FactoMineR

L'analyse de données avec R et FactoMineR. Comment faire une ACP, une ACM, une AFM ou encore une classification ? Quelles aides à l'interprétation ? Comment construire des graphiques lisibles ? commen...

Cours d'ACP : théorie et pratique (ancienne version)
de François Husson
38 635 vues • il y a 2 ans
CC

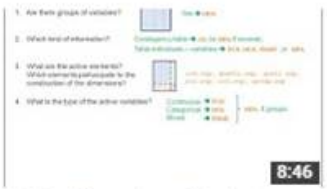
Démarche en analyse des données (ancienne version)
de François Husson
19 348 vues • il y a 3 ans

ACP avec le menu déroulant de FactoMineR
de François Husson
7 533 vues • il y a 3 ans

Analyse en Composantes Principales (ACP) avec...
de François Husson
24 748 vues • il y a 3 ans
CC

Exploratory multivariate analysis with R and FactoMineR

This video shows how to perform exploratory multivariate analyses in a French way using R and FactoMineR and how to handle missing values....



Methodology in multivariate exploratory data analysis
de François Husson
7 973 vues • il y a 3 ans

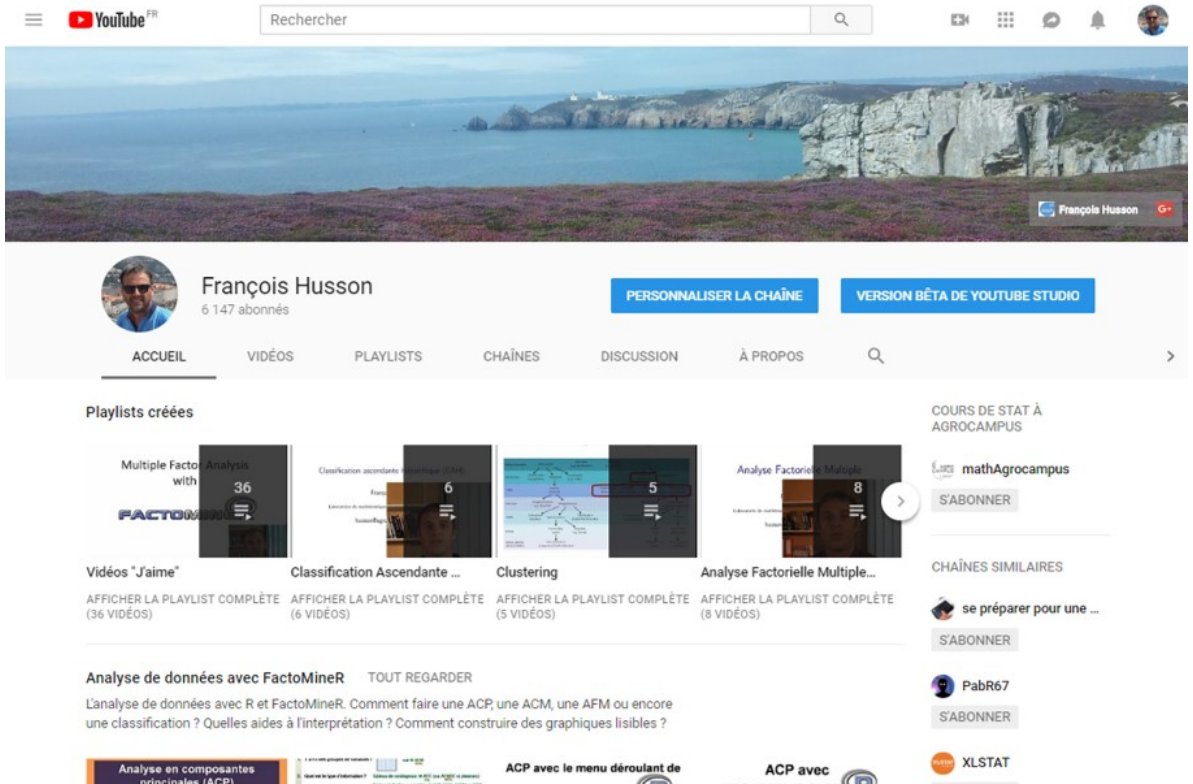
Course on PCA: theory and practice
de François Husson
20 801 vues • il y a 2 ans
CC

Principal component analysis (PCA) with R
de François Husson
25 377 vues • il y a 3 ans
CC

Correspondence analysis with FactoMineR
de François Husson
6 114 vues • il y a 2 ans

User aids: Youtube channel

- https://www.youtube.com/HussonFrancois
- Course videos (27 FR - 24 EN)
- Software tutorials (18 FR - 12 EN)

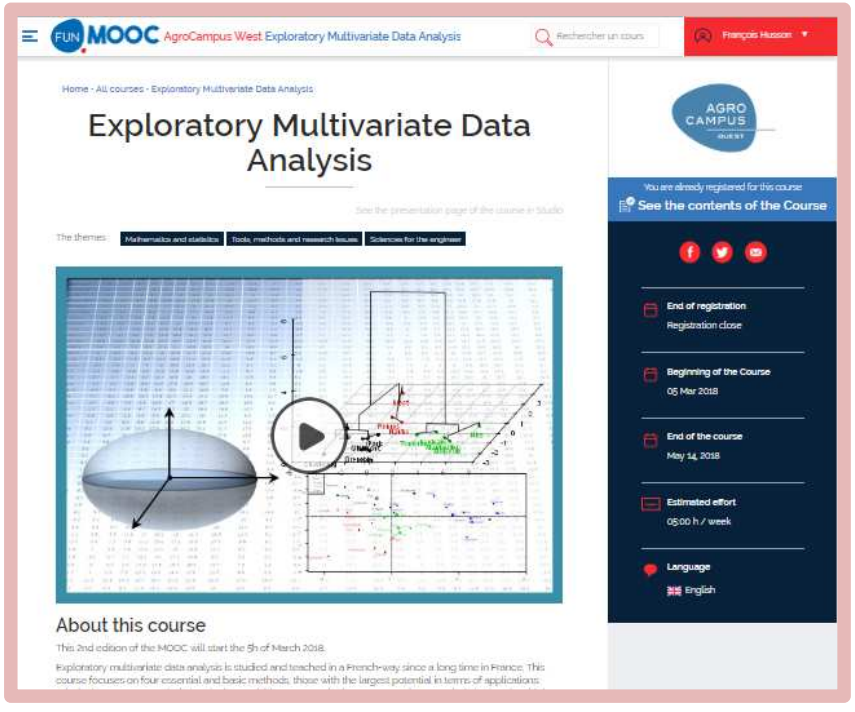


MOOC "Exploratory Multivariate Data Analysis"

RcmdrPlugin.FactoMineR: GUI
Factoshiny : Interactive graphs

missMDA : missing values

FactoInvestigate : automatic reporting



The R Journal

Journal of Statistical Software
April 2016, Volume 70, Issue 1. doi: 10.18637/jss.v070.i01

FACTOMINER

MOOC "Exploratory Multivariate Data Analysis"

MOOC = Massive Open Online Courses

Objective: to understand and know how to implement five exploratory multivariate methods

The MOOC has been designed for applications: many examples and software implementation (FactoMineR)

5 sessions in French, 2 in English:

- 2014 on a University platform (Moodle)
- Since 2015 in FR and 2017 in EN on the platform France Université Numérique

NEW SESSION in March 2019 in French and in English

MOOC "Exploratory Multivariate Data Analysis"

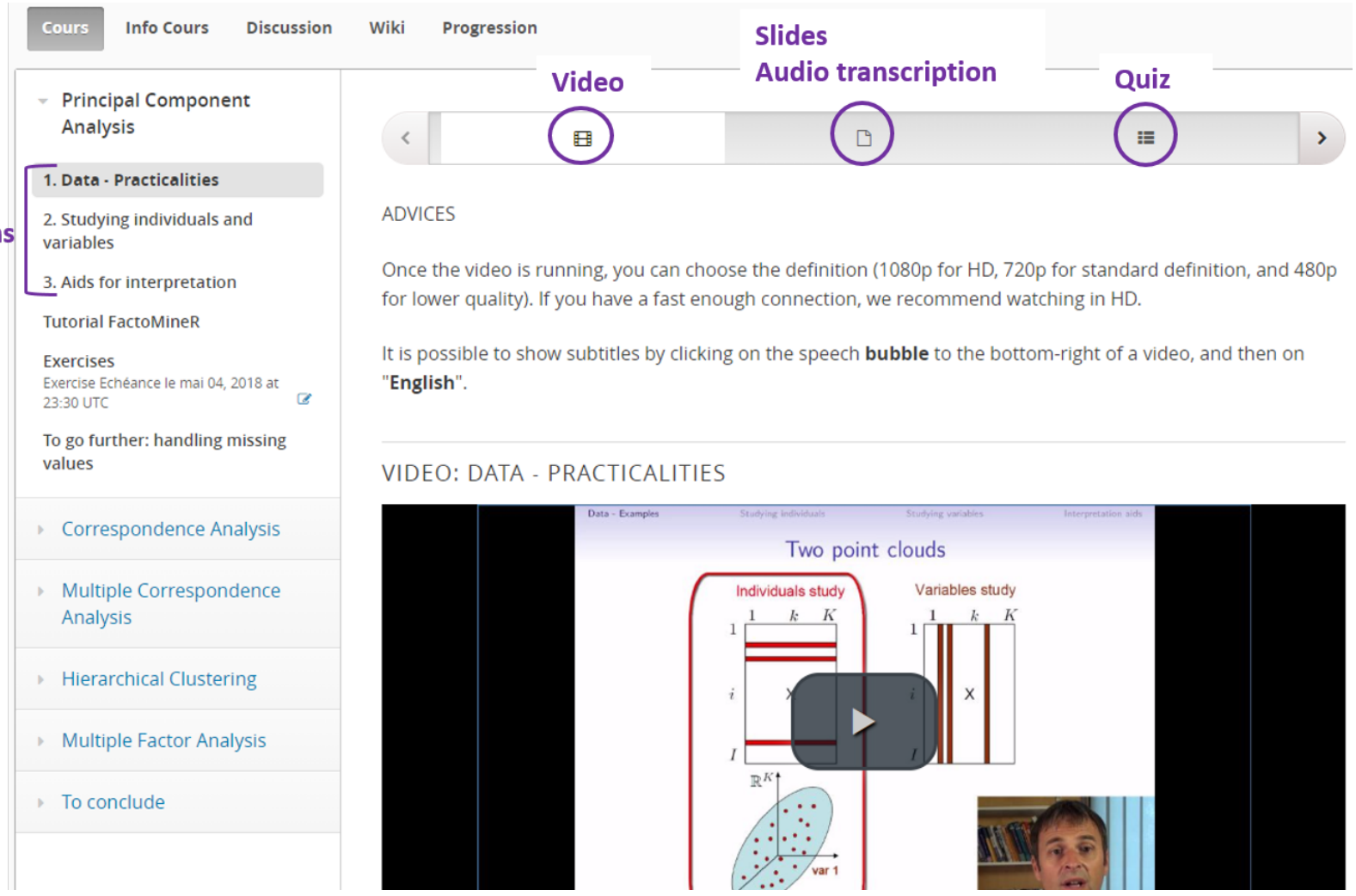
MOOC over 5 weeks: each week focuses on one analysis method (for example : Principal Component Analysis)

Each week consists of:

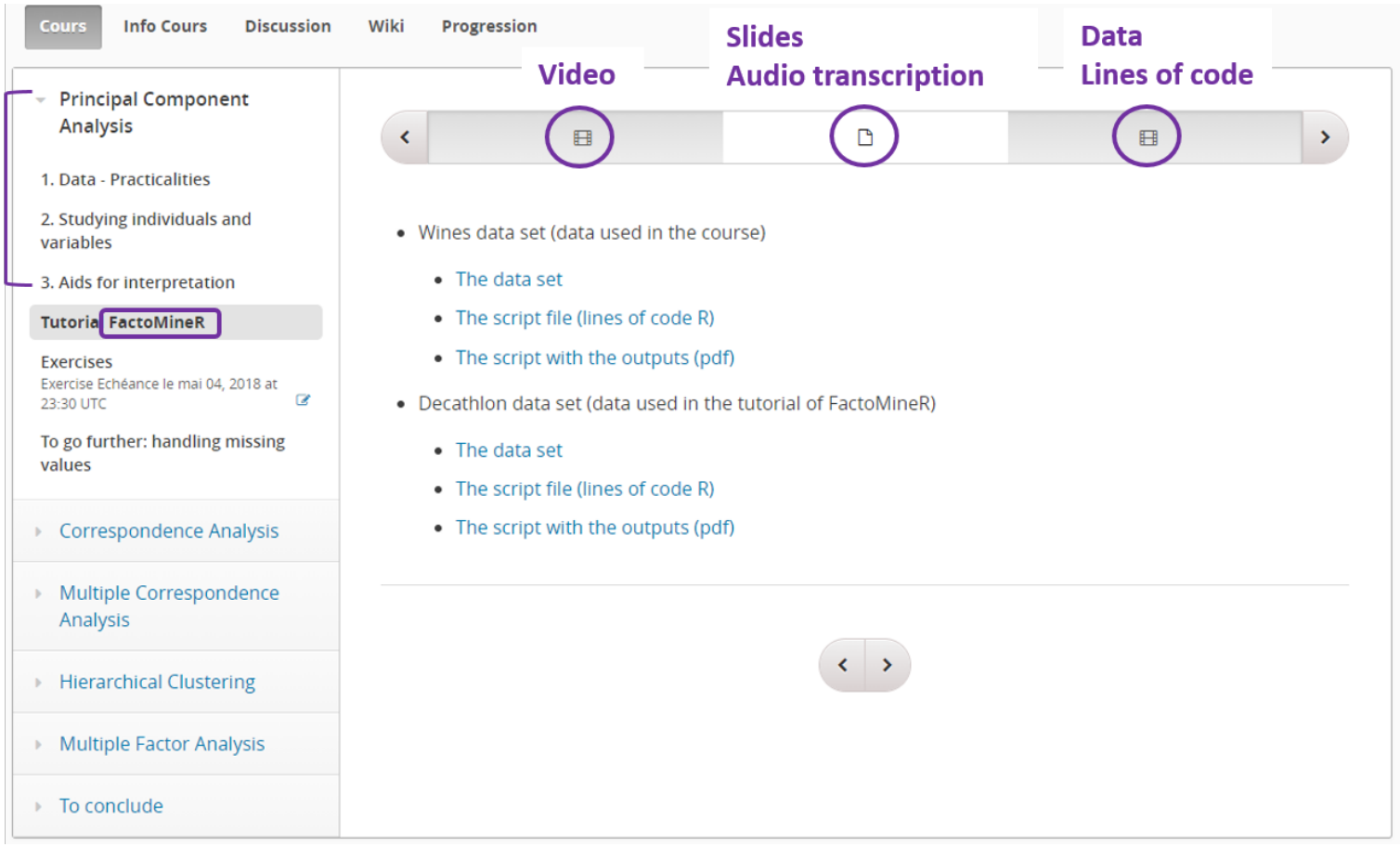
- video courses : description of the method
- quizzes
- software implementation
- exercises
- case studies

Moreover, the MOOC has a forum and a wiki
Possibility of obtaining a *successful completion certificate*

MOOC Exploratory Multivariate Data Analysis



MOOC Exploratory Multivariate Data Analysis



MOOC Exploratory Multivariate Data Analysis

Lessons

Software

Exercises

Cours | Info Cours | Discussion | Wiki | Progression

Tabletop Exercises **Computer Exercises**

Principal Component Analysis

1. Data - Practicalities
2. Studying individuals and variables
3. Aids for interpretation

Tutorial FactoMineR

Exercises
Exercise Echéance le mai 04, 2018 at 23:30 UTC

To go further: handling missing values

Correspondence Analysis

Multiple Correspondence Analysis

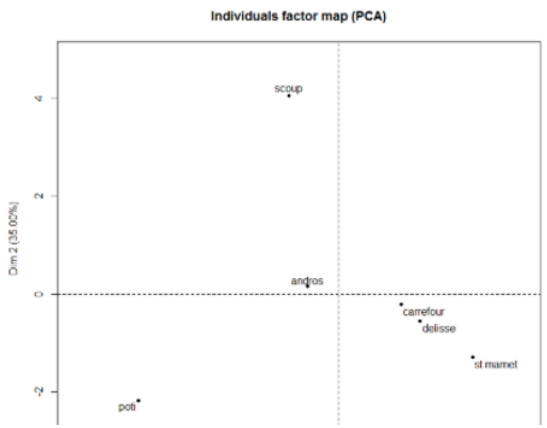
Hierarchical Clustering

Multiple Factor Analysis

To conclude

EXERCISE 1. INTERPRETING PCA PLOTS (6 points possibles)

In a sensory analysis, a jury assessed six apple sauces based on a list of 11 descriptors such as acidity, bitterness, raw apple flavor, and so on. The scores were gathered in a table with applesauces as rows, and descriptors as columns, with entries the scores for the given product and descriptor. The descriptors are of three kinds: flavor (F), texture (T), and odor (O). For example, F.sweet corresponds to sweetness. In addition, an appreciation score (hedonistic) was given to each applesauce. A PCA was carried out on the data table, with the appreciation score added as a supplementary variable. The plots of individuals and variables are provided below.



MOOC Exploratory Multivariate Data Analysis

Lessons

Software

Exercises

Cours | Info Cours | Discussion | Wiki | Progression

Principal Component Analysis

1. Data - Practicalities
2. Studying individuals and variables
3. Aids for interpretation

Tutorial FactoMineR

Exercises
Exercise Echéance le mai 04, 2018 at 23:30 UTC

To go further: handling missing values

Correspondence Analysis

Multiple Correspondence Analysis

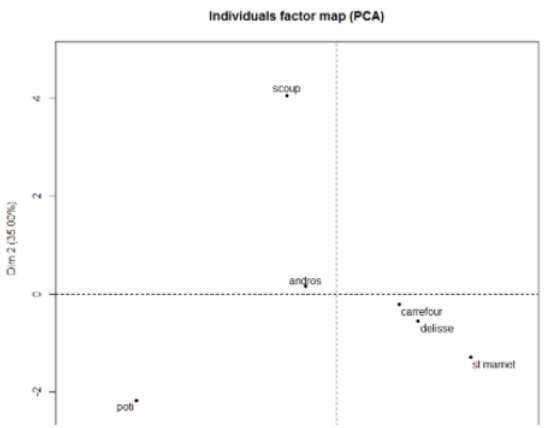
Hierarchical Clustering

Multiple Factor Analysis

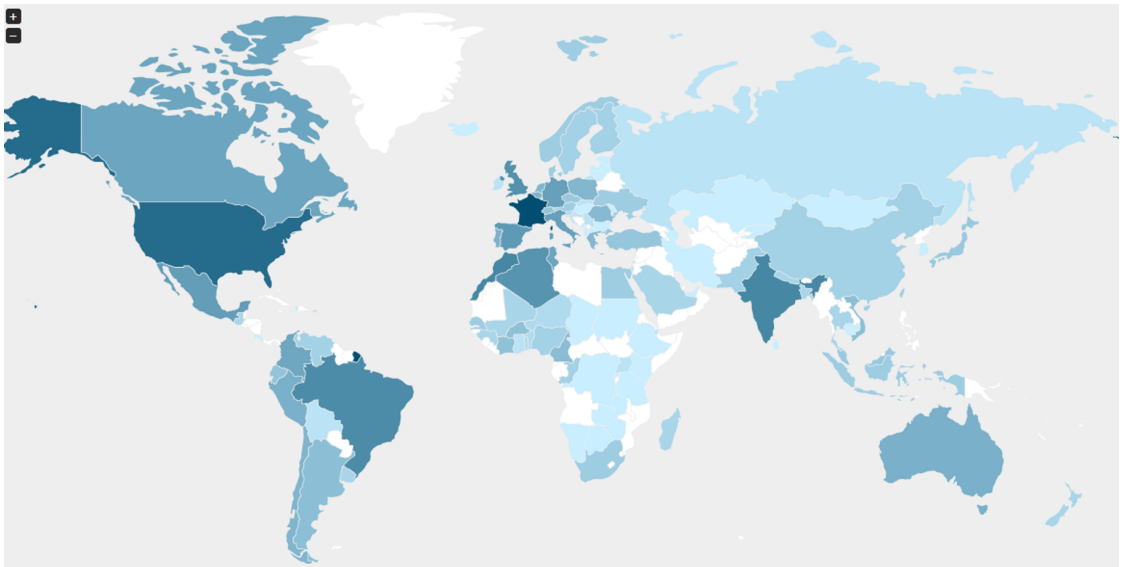
To conclude

EXERCISE 1. INTERPRETING PCA PLOTS (6 points possibles)

In a sensory analysis, a jury assessed six apple sauces based on a list of 11 descriptors such as acidity, bitterness, raw apple flavor, and so on. The scores were gathered in a table with applesauces as rows, and descriptors as columns, with entries the scores for the given product and descriptor. The descriptors are of three kinds: flavor (F), texture (T), and odor (O). For example, F.sweet corresponds to sweetness. In addition, an appreciation score (hedonistic) was given to each applesauce. A PCA was carried out on the data table, with the appreciation score added as a supplementary variable. The plots of individuals and variables are provided below.



What was the audience for this MOOC?



Each year:
$\approx$ 5 200 FR students
$\approx$ 1 800 EN students
112 countries
18 to 78 years ($\bar{x} = 35$)
60% Master, 18% pHD

- Learners from industrial and academic world: research institutes, universities, industrial and business companies
- Different fields: economy, biology, genomic, geography, linguistics...

Estimated working time if full reading: 5 hours / week
Diverse audience with different goals and different levels of investment

What interests to use R?

- Free of charge
- Very well known
- Available all over the world (with the same version)
- Same use regardless of the operating system
- Lines of code $\Rightarrow$ reproducibility of analyses
- Examples with Rmarkdown

MOOC *versus* course in live

During a MOOC learning, participants can

- manage their time
- learn some parts of the course more than others
- discuss with other participants: a more participative training
- diverse audience and many learners enriches the MOOC
- do not hesitate to ask question via the forum
- come back to the course when they want (they just need to remind that they have seen the method during the MOOC)

But a course in live has also some advantages!
$\Longrightarrow$ Using a MOOC in the classroom or during a long-life training session: a complementary approach for the future?

MOOC used with my students

Organisation :

- course on Exploratory multivariate data analysis (25h)
- 192 students in master – 3 teachers (for 10 hours)

Content :

- same videos, quizzes, exercises, forum, wiki
- permanences to discuss the course and their project
- evaluation by an additional quiz and a project on a case study

Positive aspects:

- students are more independent
- students questions-discussions are more interesting
- positive effect of learner discussions and comments
- more time to discuss and exchange on their project

Limitations:

- some students prefer the gentle tranquillity of the classroom
- those who are not working are difficult to motivate

The MOOC in my opinion

What it brings

- Everything I've already talked about but also...
- increased visibility (Youtube channel 800 views/day, MOOC $\approx$ 150 000 views)
- a very complete reference tutorial for a package
- a complete guide for a set of methods
- an opening to new audiences
- an incentive to rethink teaching and pedagogy

BUT

- It takes a lot of investment: 1 hour of class $\approx$ 20-30h!!!!
- Sequencing, quizzes and exercises are also time-consuming
- Are MOOCs getting old well?

From a package to a project on Exploratory Multivariate Analysis

- Apr 2006 : package FactoMineR (PCA, CA, MCA, MFA, etc.)
- Apr 2007 : paper on FactoMineR (Journal of Statistical Software)
- Apr 2007 : website
- Nov 2007 : package RcmdrPlugin.FactoMineR (GUI)
- Jun 2009 & Feb 2016: book *Analyse de données avec R*
- May 2010 : package missMDA (handling missing values)
- Jan 2011 : Book *Exploratory Multivariate Analysis by Example Using R*
- Apr 2011 : tutorials videos
- Sep 2011 : Google gr
- Nov 2012 : Book *Aná* *on R*
- Mar 2013 : Youtube
- Jun 2013 : paper on
- Feb 2014 : online course on Exploratory Data Analysis (in my University)
- Feb 2015: package Factoshiny (interactive graphs)
- Mar 2015 & 2016 & 2017 & 2018 : MOOC analyse de données (French)
- Apr 2015 : package FactoInvestigate (automatic reporting)
- Jun 2015 : paper on CA-Galt (R Journal)
- Apr 2016 : paper on missMDA (Journal of Statistical Software)
- Mar 2017 & 2018 : MOOC Exploratory Multivariate Data Analysis (English)
- Mar 2019 : New session MOOC Exploratory Multivariate Data Analysis (FR or EN)

A project on Exploratory Multivariate Analysis

RcmdrPlugin.FactoMineR: GUI
Factoshiny : Interactive graphs

missMDA : missing values

FactoInvestigate : automatic reporting

FACTOMINER

The R Journal

Journal of Statistical Software
April 2016, Volume 70, Issue 1. doi: 10.18637/jss.v070.i01