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

1. Introduction

2. Some Basic Tips (or Palisades)

3. Supplementary packages

4. Dissemination of information

5. Teaching
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

⇒ the visibility of a package is increasingly limited

Many packages are unused... and therefore useless!
Plan

1. Introduction
2. Some Basic Tips (or Palisades)
3. \texttt{FACTOMINE}\texttt{R}
4. Supplementary packages
5. Dissemination of information
6. Teaching
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 \(\Rightarrow\) GitHub or web page
  - for a large audience \(\Rightarrow\) 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

1. Introduction
2. Some Basic Tips (or Palisades)
3. \textit{factoMiner}\textregistered
4. Supplementary packages
5. Dissemination of information
6. Teaching
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)
Different methods for different data formats:

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

Clustering methods and complementary tool methods:

<table>
<thead>
<tr>
<th>Methods</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hierarchical Ascendant Clustering</td>
<td>HCPC</td>
</tr>
<tr>
<td>Description of a qualitative variable (e.g. cluster var.)</td>
<td>catdes</td>
</tr>
<tr>
<td>Description of a quantitative variable (e.g. a dimension)</td>
<td>condes, dimdesc</td>
</tr>
</tbody>
</table>
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)

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

- **PCA performed with supplementary information**
  ```r
  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**
  ```r
  plot(res.pca, habillage = 1)
  ```
Comparison of sensory panels with MFA

<table>
<thead>
<tr>
<th></th>
<th>Expert (27)</th>
<th>Consumer (15)</th>
<th>Student (15)</th>
<th>Preference (60)</th>
<th>Label (1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>wine 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>wine 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>wine 10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

Dim 1 (42.52 %)  Dim 2 (24.42 %)

-4 -2 0 2 4
-3 -2 -1 0 1 2

S Michaud  
S Renaudie  
S Trotignon  
S Buisse Domaine  
S Buisse Cristal  
V Aub Silex  
V Aub Marigny  
V Font Domaine  
V Font Brûlés  
V Font Coteaux  
Sauvignon  
Vouvray  
Expert  
Consumer  
Student
Representation of supplementary continuous variables

The favourite wine is *Vouvray Aubussière Silex*

Preferences are linked to sensory description
Plan

1. Introduction

2. Some Basic Tips (or Palisades)

3. Supplementary packages

4. Dissemination of information

5. Teaching
Supplementary packages

- **RcmdrPlugin.FactoMineR**: drop-down menu
- **Factoshiny**: GUI and interactive graphs
  \[\Rightarrow\] make it easier for users to use the packages

- **FactoInvestigate**: automatic reporting
  \[\Rightarrow\] propose an interpretation of the results

- **missMDA**: handling missing values
  \[\Rightarrow\] 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

```r
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

```r
> 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 FactoInvestigate package

Propose an interpretation of the results based on the result object

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

http://factominer.free.fr/reporting
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** $\ell$:
   (a) PCA on the completed data set $\rightarrow (F^\ell, U^\ell)$; $S$ dimensions are kept
   (b) missing values imputed with $F^\ell U^{\ell\prime}$
      $\Rightarrow X^\ell = W \ast X + (1 - W) \ast F^\ell U^{\ell\prime}$
3. *Steps of estimation and imputation are repeated*

$\Rightarrow$ gives the scores and loadings (better than Nipals)
$\Rightarrow$ gives an imputed data set
Handling missing values: PCA example

```r
> 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

⇒ Use the completed data set as an input for PCA, clustering, ...
```
Multiple imputation in PCA

⇒ iterative PCA: a single imputation method

$(\hat{F}\hat{U}')[ik]$

⇒ a single value can’t reflect the uncertainty of prediction

⇒ Multiple imputation: generate several plausible values for each missing values
Visualization of the uncertainty due to missing values

```r
> 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

1. Introduction
2. Some Basic Tips (or Palisades)
3. Supp. packages
4. Dissemination of information
5. Teaching
User aids: website

- [http://factominer.free.fr](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/](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

1. Introduction
2. Some Basic Tips (or Palisades)
3. Supplementary packages
4. Dissemination of information
5. Teaching
User aids: some videos

- available from my teaching webpage
- available from the helps of the functions of FactoMiner
User aids: Youtube channel

- [https://www.youtube.com/HussonFrancois](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
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

ADVICES

Once the video is running, you can choose the definition (1080p for HD, 720p for standard definition, and 480p for lower quality). If you have a fast enough connection, we recommend watching in HD.

It is possible to show subtitles by clicking on the speech bubble to the bottom-right of a video, and then on "English".

VIDEO: DATA - PRACTICALITIES

Two point clouds

Correspondence Analysis
Multiple Correspondence Analysis
Hierarchical Clustering
Multiple Factor Analysis
To conclude
## MOOC Exploratory Multivariate Data Analysis

### Lessons

- Principal Component Analysis
  1. Data - Practicalities
  2. Studying individuals and variables
  3. Aids for interpretation
- Correspondence Analysis
- Multiple Correspondence Analysis
- Hierarchical Clustering
- Multiple Factor Analysis
- To conclude

### Software

- FactoMineR
- Exercises
  - Exercise Echange le mai 04, 2018 at 23:30 UTC
- To go further: handling missing values

### Data

- Wines data set (data used in the course)
  - The data set
  - The script file (lines of code R)
  - The script with the outputs (pdf)
- Decathlon data set (data used in the tutorial of FactoMineR)
  - The data set
  - The script file (lines of code R)
  - The script with the outputs (pdf)
MOOC Exploratory Multivariate Data Analysis

EXERCISE 1. INTERPRETING PCA PLOTS (6 points possible)

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 applesaucs 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

**Exercise 1. Interpreting PCA plots** (6 points possible)

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:
≈ 5 200 FR students
≈ 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!

⇒ 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 group
- **Nov 2012**: Book *Análisis de datos con R*
- **Mar 2013**: Youtube channel (tutorial videos)
- **Jun 2013**: paper on MF A CT (R Journal)
- **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

Website

Google group

Videos on Youtube channel