Clustering and Principal Component Methods

1. Clustering Methods
2. Principal Components Methods as a Preprocessing Step
3. Graphical Complementarity

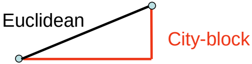
Unsupervised classification

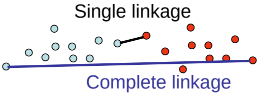
- Data set: table individuals $\times$ variables (or a distance matrix)
- Objective: to produce homogeneous groups of individuals (or groups of variables)
- Two kinds of clustering to define two structures on individuals: hierarchy or partition

Hierarchical Clustering

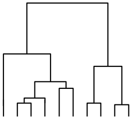
Principle: sequentially agglomerate (clusters of) individuals using

- a distance between individuals: City block, Euclidean
- an agglomerative criterion: single linkage, complete linkage, average linkage, Ward's criterion





Representation with a dendrogram



⇒ Eulidean distance is used in principal component methods

⇒ Ward's criterion is based on multidimensional variance (inertia) which is the core of principal component methods

Ascending Hierarchical Clustering

AHC algorithm:

- Compute the Euclidean distance matrix ($I \times I$)
- Consider each individual as a cluster
- Merge the two clusters A and B which are the closest with respect to the Ward's criterion:

$$\Delta_{ward}(A, B) = \frac{I_A I_B}{I_A + I_B} d^2(\mu_A, \mu_B)$$

  with d the Euclidean distance, μ_A the barycentre and I_A the cardinality of the set A
- Repeat until the number of clusters is equal to one

Ward's criterion

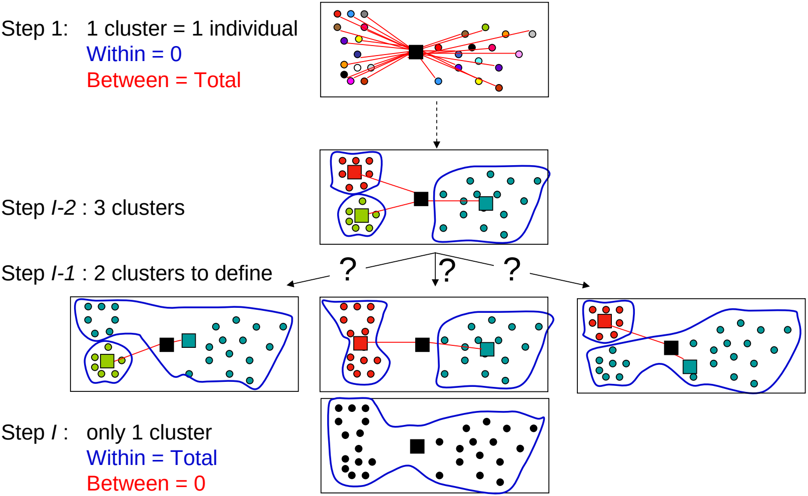
- Individuals can be represented by a cloud of points in $\mathbb{R}^K$
- Total inertia = multidimensional variance

With Q groups of individuals, inertia can be decomposed as:

$$\sum_{k=1}^{K}\sum_{q=1}^{Q}\sum_{i=1}^{I_q}(x_{iqk}-\bar{x}_k)^2 = \sum_{k=1}^{K}\sum_{q=1}^{Q} I_q(\bar{x}_{qk}-\bar{x}_k)^2+\sum_{k=1}^{K}\sum_{q=1}^{Q}\sum_{i=1}^{I_q}(x_{iqk}-\bar{x}_{qk})^2$$

Total inertia = Between inertia + Within inertia

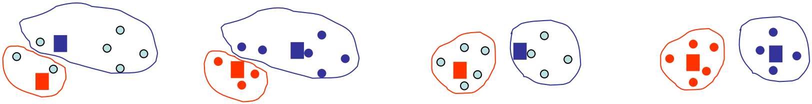
Ward's criterion



⇒ Ward minimizes the increasing of within inertia

K-means algorithm

1. Choose Q points at random (the barycentre)
2. Affect the points to the closest barycentre
3. Compute the new barycentre
4. Iterate 2 and 3 until convergence



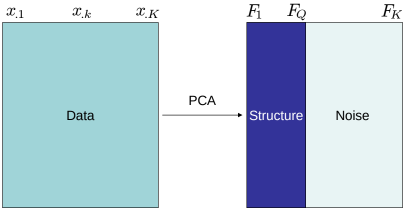
PCA as a preprocessing

With continuous variables:

⇒ AHC and k-means onto the raw data

⇒ AHC or k-means onto principal components

PCA transforms the raw variables into orthogonal principal components $F_{.1}, ..., F_{.K}$ with decreasing variance $\lambda_1 \geq \lambda_2 \geq ...\lambda_K$



⇒ Keeping the first components makes the clustering more robust

⇒ But, how many components do you keep to denoise?

MCA as a preprocessing

Clustering on categorical variables: which distance to use?

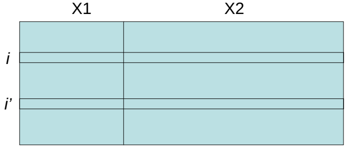
- with two categories: Jaccard index, Dice's coefficient, simple match, etc. Indices well-fitted for presence/absence data
- with more than 2 categories: use for example the χ^2-distance

Using the χ^2-distance $\Leftrightarrow$ computing distances from all the principal components obtained from MCA

In practice, MCA is used as a preprocessing in order to

- transform categorical variables in continuous ones
- delete the last dimensions to make the clustering more robust

MFA as a preprocessing



MFA balances the influence of the groups when computing distances between individuals

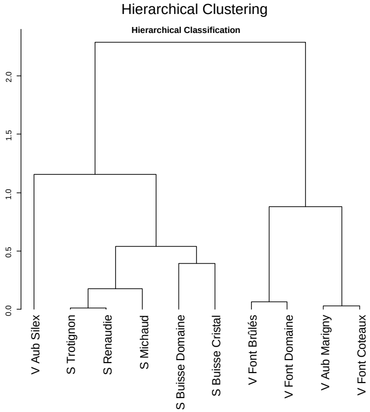
$$d^2(i,i') = \sum_{j=1}^{J} \frac{1}{\sqrt{\lambda_j}} \sum_{k=1}^{K_j} (x_{ik} - x_{i'k})^2$$

AHC or k-means onto the first principal components $(F_{.1}, ..., F_{.Q})$ obtained from MFA allows to

- take into account the groups structure in the clustering
- make the clustering more robust by deleting the last dimensions

Back to the wine data!

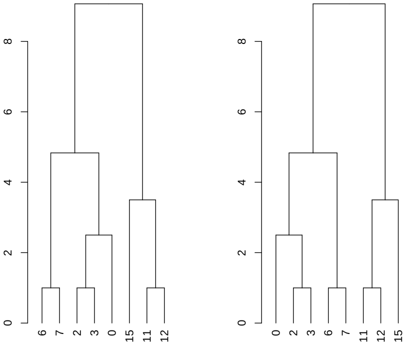
AHC onto the first 5 principal components from MFA



Individuals are sorted according to their coordinate $F_{.1}$

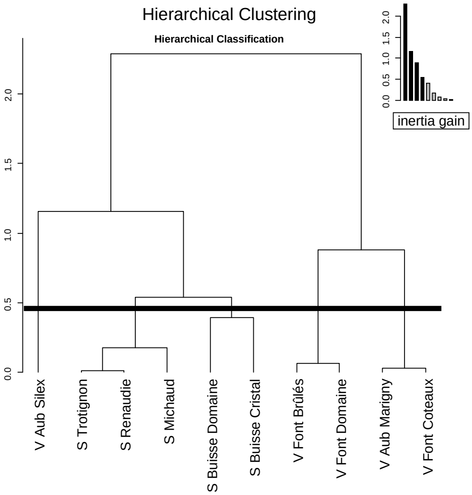
Why sorting the tree?

```
X <- c(6,7,2,0,3,15,11,12)
names(X) <- X
library(cluster)
par(mfrow=c(1,2))
plot(as.dendrogram(agnes(X)))
plot(as.dendrogram(agnes(sort(X))))
```

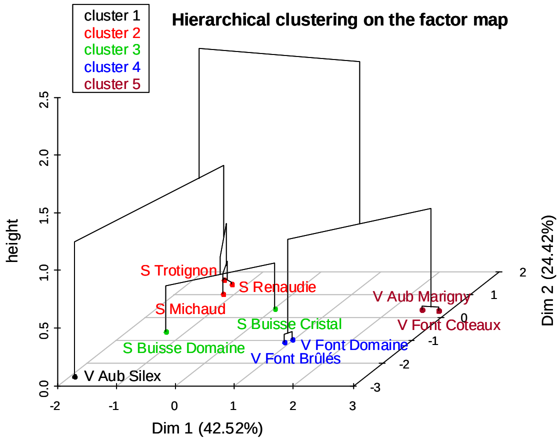


Partition from the tree

An empirical number of clusters is suggested ($\min_q \frac{W_q - W_{q+1}}{W_{q-1} - W_q}$)

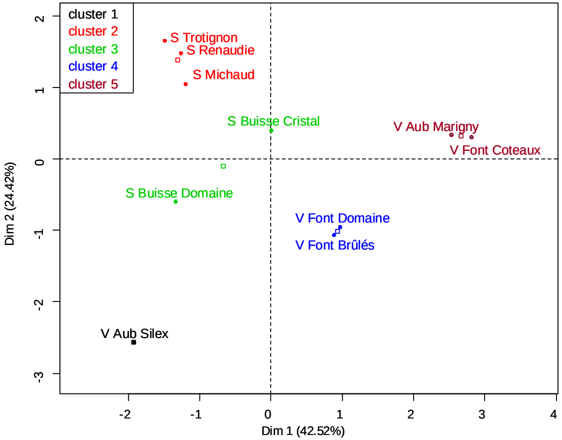


Hierarchical tree on the principal component map



Hierarchical tree gives an idea of the other dimensions

Partition on the principal component map



Continuous view (principal components) and discontinuous (clusters)

Cluster description by variables

$$\text{v.test} = \frac{\bar{x}_q - \bar{x}}{\sqrt{\frac{s^2}{I_q}\left(\frac{I-I_q}{I-1}\right)}} \sim \mathcal{N}(0,1) \qquad H_0 : \bar{x}_q = \bar{x}$$

with $\bar{x}_q$ the mean of variable x in cluster q, $\bar{x}$ (s) the mean (standard deviation) of the variable x in the data set, I_q the cardinal of cluster q

```
$desc.var$quanti$‘2‘
                v.test  Mean in Overall    sd in   Overall p.value
                        category    mean category       sd
O.passion_C       2.58      6.17    4.61     0.79     1.18    0.01
O.citrus          2.50      5.40    3.66     0.22     1.37    0.01
O.passion_S       2.45      5.69    4.18     0.54     1.20    0.01
....
Typicity         -2.42      1.36    3.91     0.72     2.07    0.02
O.candied.fruit  -2.44      0.78    2.58     0.16     1.45    0.01
O.alcohol_S      -2.48      3.98    4.33     0.13     0.28    0.01
Surface.feeling  -2.52      2.63    3.62     0.12     0.77    0.01
```

Cluster description

- by the principal components (individuals coordinates) : same description than for continuous variables

```
$desc.axes$quanti$‘2‘
      v.test  Mean in  Overall     sd in  Overall  p.value
             category     mean  category       sd
Dim.2   2.20     1.39 7.77e-17     0.253     1.24   0.0276
```

- by categorical variables : chi-square and hypergeometric test

⇒ Active and supplementary elements are used
⇒ Only significant results are presented

Cluster description by individuals

- parangon: the closest individuals to the barycentre of the cluster

$$\min_{i \in q} d(x_{i.}, \mu_q) \text{ with } \mu_q \text{ the barycentre of cluster } q$$

- specific individuals: the furthest individuals to the barycentres of the other clusters (the individuals sorted according to their distance from the highest to the smallest to the closest barycentre)

$$\max_{i \in q} \min_{q' \neq q} d(x_{i.}, \mu_{q'})$$

```
desc.ind$para
cluster: 2
 S Renaudie    S Trotignon      S Michaud
    0.1002890      0.3101154      0.3640145
-------------------------------------------
desc.ind$dist
cluster: 2
S Trotignon    S Renaudie      S Michaud
     1.934103      1.687849      1.265386
-------------------------------------------
```

Complementarity between hierarchical clustering and partitioning

- Partitioning after AHC: the k-means algorithm is initialized from the barycentres of the partition obtained from the tree
  - consolidate the partition
  - loss of the hierarchy

- AHC with many individuals: time-consuming
  $\Rightarrow$ partitioning before AHC
  - compute k-means with approximately 100 clusters
  - AHC on the weighted barycentres obtained from the k-means
    $\Rightarrow$ top of the tree is approximately the same

Practice with R

```
res.hcpc <- HCPC(res.mfa)
```

```
##### Example of clustering on categorical data
data(tea)
res.mca <- MCA(tea,quanti.sup=19,quali.sup=20:36)
plot(res.mca,invisible=c("var","quali.sup","quanti.sup"),cex=0.7)
plot(res.mca,invisible=c("ind","quali.sup","quanti.sup"),cex=0.8)
plot(res.mca,invisible=c("quali.sup","quanti.sup"),cex=0.8)
dimdesc(res.mca)

res.mca <- MCA(tea,quanti.sup=19,quali.sup=20:36, ncp=10)
res.hcpc <- HCPC(res.mca)
```

CARME conference

International conference on Correspondence Analysis and Related MEthods
Agrocampus Rennes (France), February 8-11, 2011

R tutorials for corresp. ana. and related methods of visualization:

- S. Dray: multivariate analysis of ecological data with `ade4`
- O. Nenadić & M. Greenacre: correspondence analysis with `ca`
- S. Lê: from one to multiple data tables with `FactoMineR`
- J. de Leeuw & P. Mair: multidimensional scaling using majorisation with `smacof`

Invited speakers: Monica Bécue, Cajo ter Braak, Jan de Leeuw, Stéphane Dray, Michael Friendly, Patrick Groenen, Pieter Kroonenberg

Bibliography

- Escofier B. & Pagès J. (1994). Multiple factor analysis (AFMULT package). *Computational Statistics and Data Analysis*, 121-140.
- Greenacre M. & Blasius J. (2006). *Multiple Correspondence Analysis and related methods*. Chapman & Hall/CRC.
- Husson F., Lê S. & Pagès J. (2010). *Exploratory Multivariate Analysis by Example Using R*. Chapman & Hall.
- Jolliffe I. (2002). *Principal Component Analysis*. Springer. 2nd edn.
- Lebart L., Morineau A. & Warwick K. (1984). *Multivariate descriptive statistical analysis*. Wiley, New-York.
- Le Roux B. & Rouanet H. (2004). *Geometric Data Analysis, From Correspondence Analysis to Structured Data Analysis*. Dordrecht: Kluwer.

Packages' bibliography

http://cran.r-project.org/web/views/Multivariate.html
http://cran.r-project.org/web/views/Cluster.html

- *ade4* package: data analysis functions to analyse Ecological and Environmental data in the framework of Euclidean Exploratory methods
http://pbil.univ-lyon1.fr/ADE-4
- *ca* package (Greenacre and Nenadic) deals with simple, multiple and joint correspondence analysis
- *cluster* package: basic and hierarchical clustering
- *dynGraph* package: visualization software to explore interactively graphical outputs provided by multidimensional methods
http://dyngraph.free.fr
- *FactoMineR* package
http://factominer.free.fr
- *hopach* package: builds hierarchical tree of clusters
- *missMDA* package: imputes missing values with multivariate data analysis methods

FactoMineR

A website with documentation, examples, data sets:
`http://factominer.free.fr`

How to install the Rcmdr menu:
copy and paste the following line of code in a R session

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

A book:
Husson F., Lê S. & Pagès J. (2010). *Exploratory Multivariate Analysis by Example Using R*. Chapman & Hall.