# Statistics and learning

## Multivariate statistics 2 and clustering

Emmanuel Rachelson and Matthieu Vignes

ISAE SupAero

Wednesday 2nd October 2013

# Link to the previous session

Goal of multivariate (exploratory) statistics: understanding high-dimensional data sets, reducing their 'useful' dimensions, representing them, seeking hidden or latent factors ...
Today we will:

- review PCA needed ?

# Link to the previous session

Goal of multivariate (exploratory) statistics: understanding high-dimensional data sets, reducing their 'useful' dimensions, representing them, seeking hidden or latent factors ...
Today we will:

- review PCA needed ?
- introduce Multidimensional scaling (MDS) as a factor analysis of a distance matrix

# Link to the previous session

Goal of multivariate (exploratory) statistics: understanding high-dimensional data sets, reducing their 'useful' dimensions, representing them, seeking hidden or latent factors ...
Today we will:

- review PCA needed ?
- introduce Multidimensional scaling (MDS) as a factor analysis of a distance matrix
- introduce Canonical correlation analysis (CCA): for $p$ quantitative variables and $q$ quantitative variables)

# Link to the previous session

Goal of multivariate (exploratory) statistics: understanding high-dimensional data sets, reducing their 'useful' dimensions, representing them, seeking hidden or latent factors ...
Today we will:

- review PCA needed ?
- introduce Multidimensional scaling (MDS) as a factor analysis of a distance matrix
- introduce Canonical correlation analysis (CCA): for $p$ quantitative variables and $q$ quantitative variables)
- introduce Correspondence analysis (CA): for $2$ qualitative variables with several (many) levels.

# Link to the previous session

Goal of multivariate (exploratory) statistics: understanding high-dimensional data sets, reducing their 'useful' dimensions, representing them, seeking hidden or latent factors ...
Today we will:

- review PCA needed ?
- introduce Multidimensional scaling (MDS) as a factor analysis of a distance matrix
- introduce Canonical correlation analysis (CCA): for $p$ quantitative variables and $q$ quantitative variables)
- introduce Correspondence analysis (CA): for $2$ qualitative variables with several (many) levels.
- introduce clustering methods like hierarchical clustering or Kmeans-like algorithms.

# Multidimensional scaling (MDS)

- now only an index between individual is known, variables are not observed anymore: $n \times n$ matrix (think of distances).

# Multidimensional scaling (MDS)

- now only an index between individual is known, variables are not observed anymore: $n \times n$ matrix (think of distances).
- **Goal**: represent the cloud of points in a low-dimensional subspace.

# Multidimensional scaling (MDS)

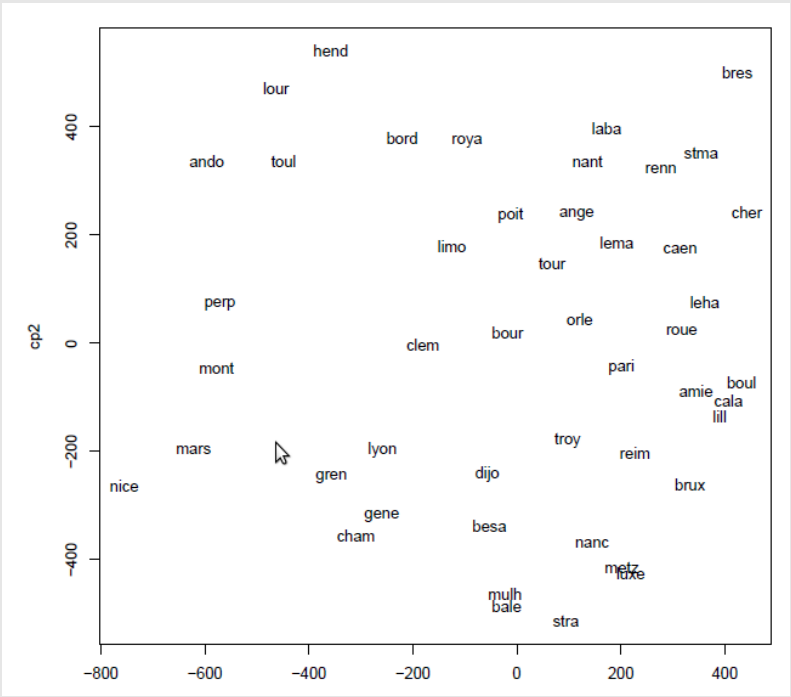
- now only an index between individual is known, variables are not observed anymore: $n \times n$ matrix (think of distances).
- **Goal**: represent the cloud of points in a low-dimensional subspace.
- MDS $=$ PCA on distance matrix !

# Multidimensional scaling (MDS)

- now only an index between individual is known, variables are not observed anymore: $n \times n$ matrix (think of distances).
- **Goal**: represent the cloud of points in a low-dimensional subspace.
- MDS $=$ PCA on distance matrix !

## Easy example

Road distances between $47$ French cities. Is it Euclidian ?



# Canonical correlation analysis (CCA)

- Uses techniques close to PCA to achieve a kind of multiple output multivariate regression

# Canonical correlation analysis (CCA)

- Uses techniques close to PCA to achieve a kind of multiple output multivariate regression
- **Goal**: Linking 2 groups of variables ($X$ and $Y$) measured on the same individuals

# Canonical correlation analysis (CCA)

- Uses techniques close to PCA to achieve a kind of multiple output multivariate regression
- **Goal**: Linking 2 groups of variables ($X$ and $Y$) measured on the same individuals
- Example from yesterday on the study of fatty acids and gene levels on mice: are some acids more present when some genes are over-expressed ? Or conversely ? $\rightarrow$ Practical session !

# Canonical correlation analysis (CCA)

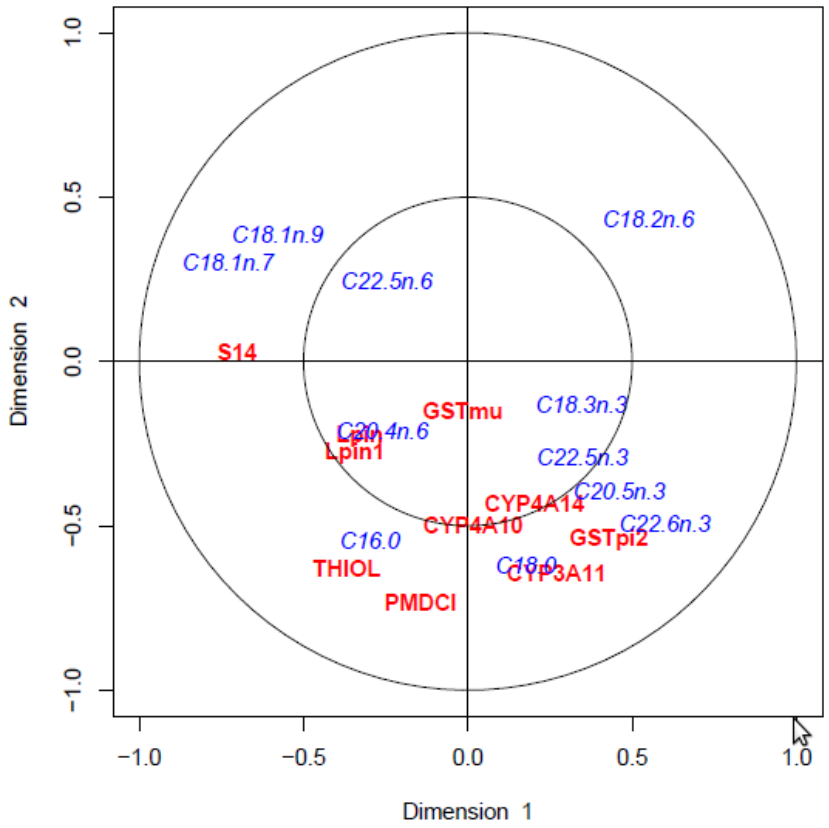
- Uses techniques close to PCA to achieve a kind of multiple output multivariate regression
- **Goal**: Linking 2 groups of variables ($X$ and $Y$) measured on the same individuals
- Example from yesterday on the study of fatty acids and gene levels on mice: are some acids more present when some genes are over-expressed ? Or conversely ? $\rightarrow$ Practical session !
- Consists in looking for a couple of vectors, one related to $X$ (gene expressions) and one to $Y$ (metabolite levels) which are maximally conected. And iteratively (without correlation between iterations).

# Canonical correlation analysis (CCA)

- Uses techniques close to PCA to achieve a kind of multiple output multivariate regression
- **Goal**: Linking 2 groups of variables ($X$ and $Y$) measured on the same individuals
- Example from yesterday on the study of fatty acids and gene levels on mice: are some acids more present when some genes are over-expressed ? Or conversely ? $\rightarrow$ Practical session !
- Consists in looking for a couple of vectors, one related to $X$ (gene expressions) and one to $Y$ (metabolite levels) which are maximally conected. And iteratively (without correlation between iterations).
- Variables can be represented in either basis, it does not change the interpretation.

# CCA (cont'd)

Need to have $p, q \leq n$. We kept $10$ genes and $11$ fatty acids.




More interpretation ? $\rightarrow$ Practical session

# Correspondence analysis (CA)

- Becomes AFC in French

# Correspondence analysis (CA)

- Becomes AFC in French
- similar concept to PCA: represent the distribution of the 2 variables and plots the individuals. but applies to qualitative rather than quantitative data $\rightarrow$ contingency table $(n_{i,j})$

# Correspondence analysis (CA)

- Becomes AFC in French
- similar concept to PCA: represent the distribution of the 2 variables and plots the individuals. but applies to qualitative rather than quantitative data $\rightarrow$ contingency table $(n_{i,j})$
- This is double PCA (line and column profiles) on $(X_{ij}) = (\frac{f_{i,j}}{f_{i,.} f_{.j}} - 1)$, with $f_{i,j} = n_{i,j}/n$.

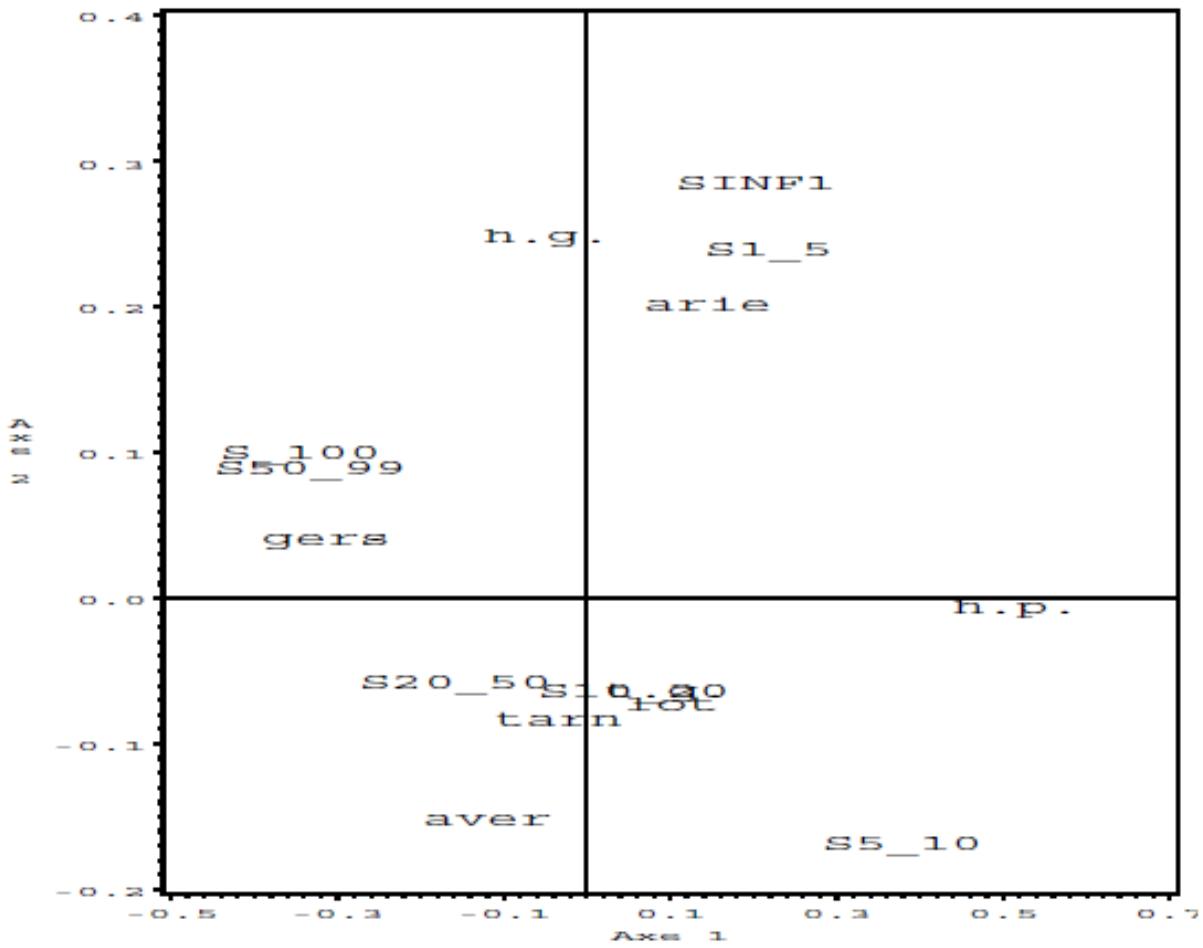
# Correspondence analysis (CA)

- Becomes AFC in French
- similar concept to PCA: represent the distribution of the 2 variables and plots the individuals. but applies to qualitative rather than quantitative data $\rightarrow$ contingency table $(n_{i,j})$
- This is double PCA (line and column profiles) on $(X_{ij}) = (\frac{f_{i,j}}{f_{i,.} f_{.j}} - 1)$, with $f_{i,j} = n_{i,j}/n$.
- Note that $\chi^2$ writes $n \sum_i \sum_j \tilde{f_{i,j}} x_{i,j}^2$

# CA: an example

Cultivated area in the Midi-Pyrénées region

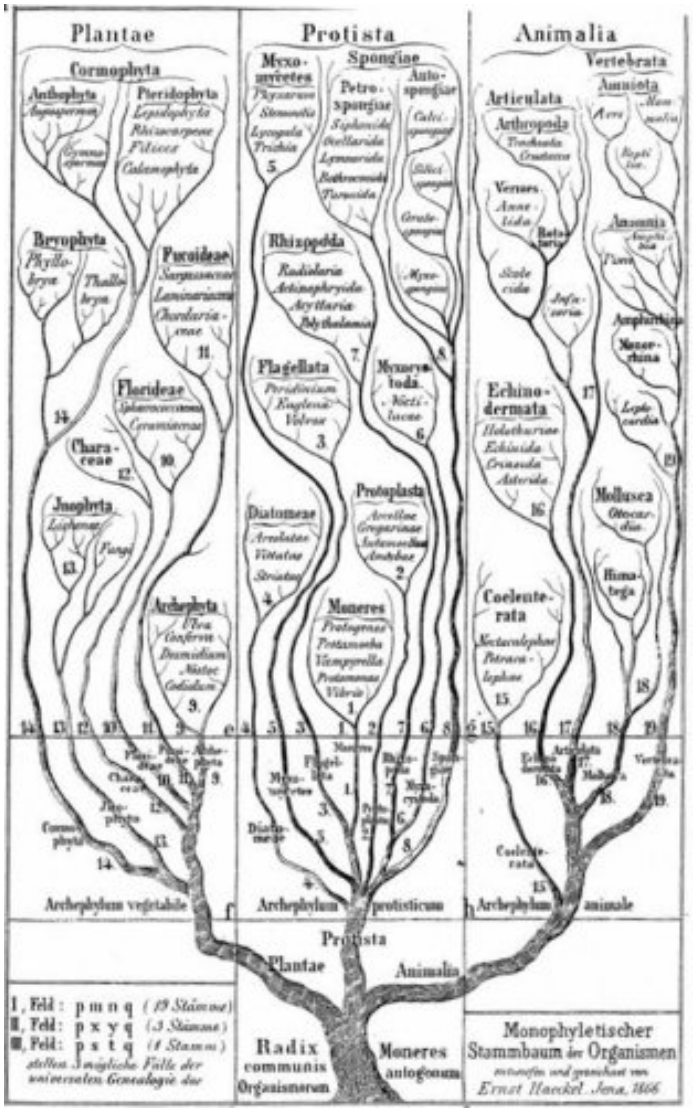
Simultaneous representation of *département* and farm size (in 6 bins).



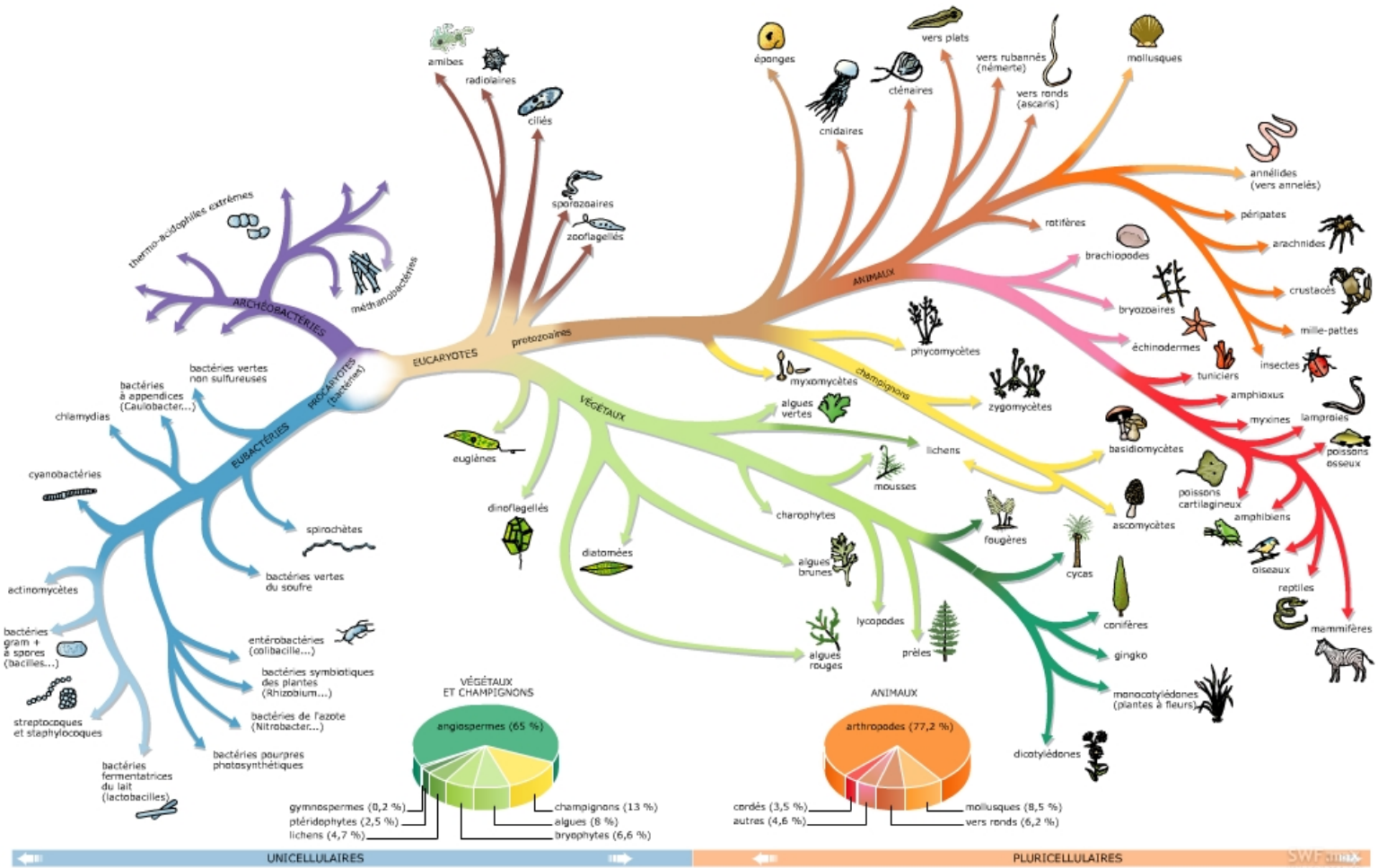
# Clustering: grouping into classes

Ever heard of that in your background ??

# Clustering: grouping into classes



# Clustering: grouping into classes



# Cluster analysis or clustering

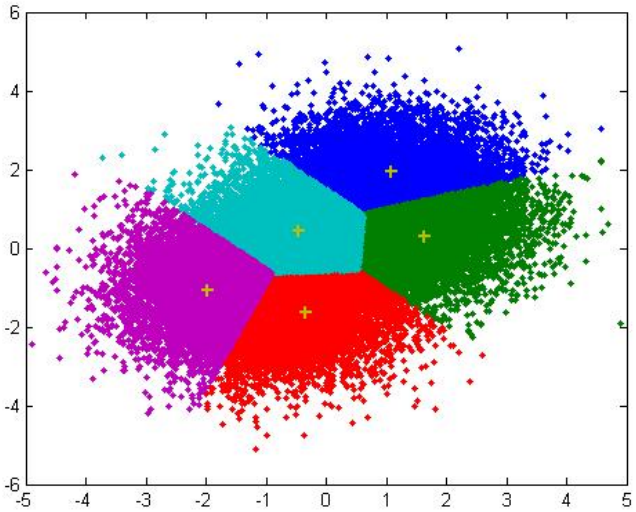
- Task of **grouping** objects so that objects belonging to the same group are 'more similar' to each other than to those in any other group $\rightarrow$ multiobjective optimisation task.

# Cluster analysis or clustering

- Task of **grouping** objects so that objects belonging to the same group are 'more similar' to each other than to those in any other group $\rightarrow$ multiobjective optimisation task.

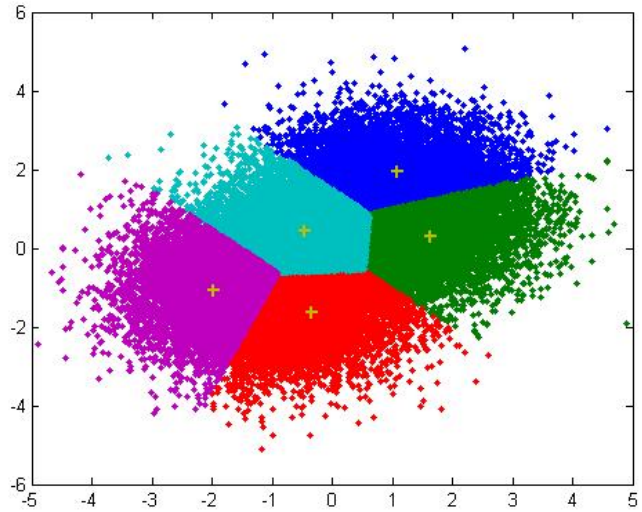
# Cluster analysis or clustering

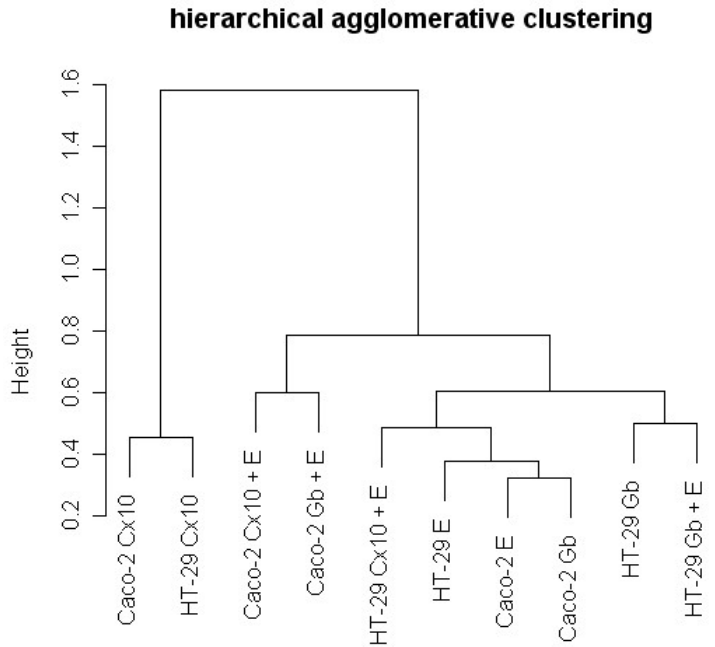
- Task of **grouping** objects so that objects belonging to the same group are 'more similar' to each other than to those in any other group $\rightarrow$ multiobjective optimisation task.



# Cluster analysis or clustering

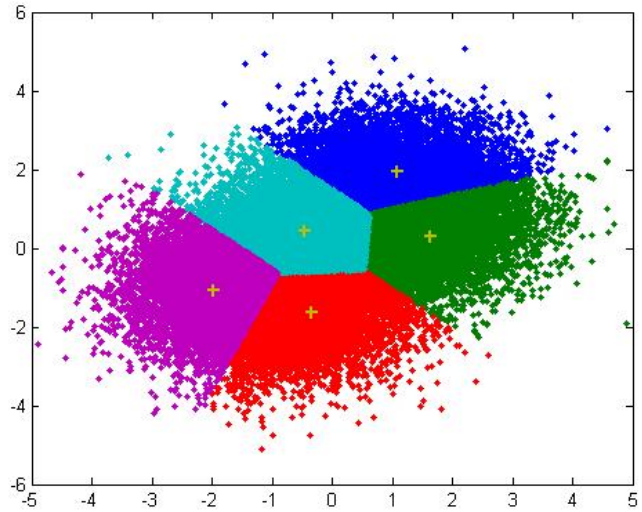
- Task of **grouping** objects so that objects belonging to the same group are 'more similar' to each other than to those in any other group $\rightarrow$ multiobjective optimisation task.

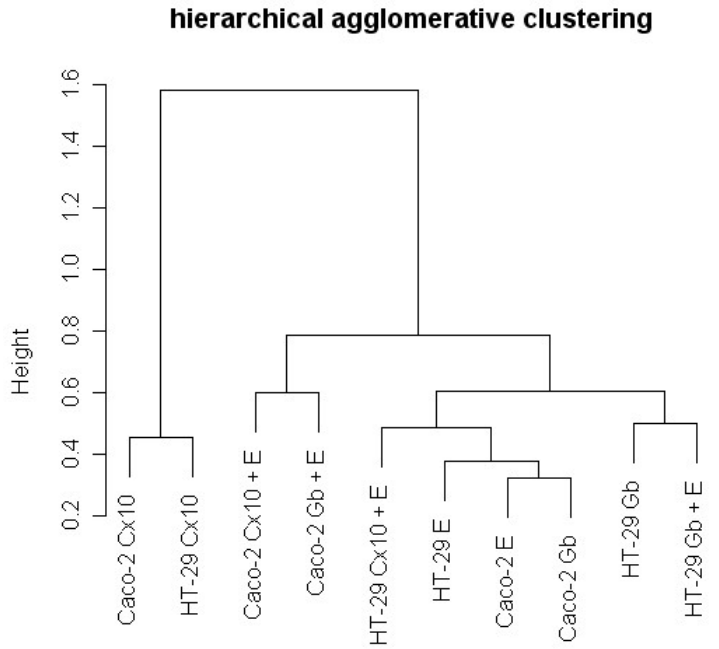


# Cluster analysis or clustering

- Task of **grouping** objects so that objects belonging to the same group are 'more similar' to each other than to those in any other group $\rightarrow$ multiobjective optimisation task.

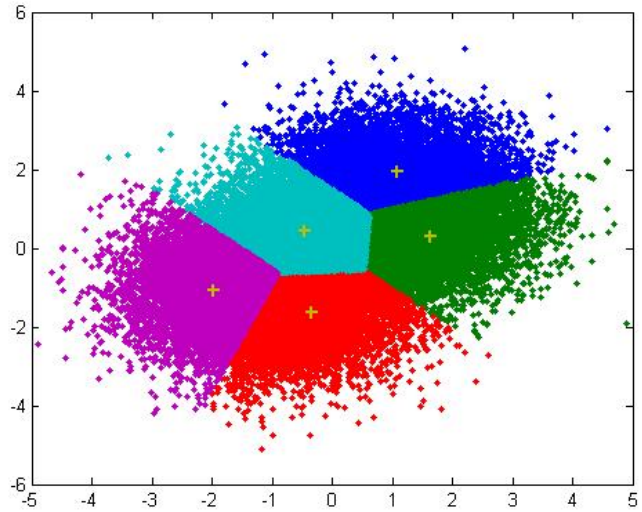


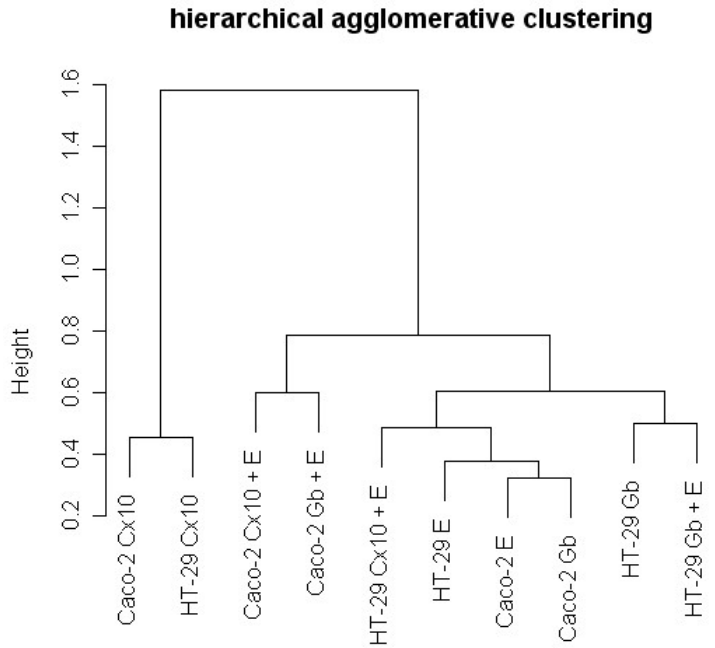




- Several algorithms can do the job, their differences mainly being about used distance.

# Cluster analysis or clustering

- Task of **grouping** objects so that objects belonging to the same group are 'more similar' to each other than to those in any other group $\rightarrow$ multiobjective optimisation task.




- Several algorithms can do the job, their differences mainly being about used distance.
- Possibly, different parameters (initialisation, distance used, ending criterion ...) lead to different representations.

# Clustering algorithms

Challenge: build your own clustering algorithm ?!

# Clustering algorithms

Challenge: build your own clustering algorithm ?!

Let's quote only few of widespread clustering algorithms:

- hierarchical clustering (single, complete, average linkages)
- centroid models (*e.g.* Kmeans clustering)
- distribution models (statistical definition *e.g.* multivariate Gaussian distribution)
- graph or density models (*e.g.* clique)
- ...

# Clustering: some formalism

- Define a similarity (symetry, self-similarity, bound) $\rightarrow$ dissimilarity
- Distance need additional property: $d(i,j) = 0 \Rightarrow i = j$ (*Euclidian* dist. from scalar product)

# Clustering: some formalism

- Define a similarity (symetry, self-similarity, bound) $\rightarrow$ dissimilarity
- Distance need additional property: $d(i,j) = 0 \Rightarrow i = j$ (*Euclidian* dist. from scalar product)

A goodness-of-fit of partitions can be defined: (i) external: TP, FP ... $\rightarrow$ precision, sensitivity or Rand/Jaccard index or (ii) internal: Dunn index $D = \min_i \min_{j \neq i} \frac{d(i,j)}{\max_k d'(k)}$.

# Homework

What do students choose after French baccalauréat ?

First describe and then represent this (simple) data set in some informative way.

Hint: CA...

| origin | counselling | | | |
|---|---|---|---|---|
| | université | prep. clas. | other | Total |
| bac lit. | 13 | 2 | 5 | 20 |
| bac éco. | 20 | 2 | 8 | 30 |
| bac scient. | 10 | 5 | 5 | 20 |
| bac tech. | 7 | 1 | 22 | 30 |
| Total | 50 | 10 | 40 | 100 |

# Finished

Next time: tests