VIETNAMESE SPRING SCHOOL
in
STATISTICS AND MACHINE LEARNING

CHAPTER III. Unsupervised classification

Agnès LAGNOUX

lagnoux@univ-tlse2.fr

LECTURE SUPPORTS

All the files of this lecture are available on my webpage :

`https://perso.math.univ-toulouse.fr/lagnoux/enseignements/`

You will find them at the bottom of the page.

LECTURE OUTLINE

1. Introduction
2. Supervised classification
   - Linear regression
   - k-nearest neighbors
   - Discriminant factor analysis
   - Naive Bayesian
   - Logistic regression
3. Unsupervised classification
   - Hierarchical clustering analysis
   - k-means

Introduction to Unsupervised Classification

- Data and objectives
- Partition or hierarchy ?
- How to measure the distance between individuals ?
- How to measure the distance between classes ?

How to evaluate the quality of a partition ?

How to compare two partitions ?

The k-means

- Choice of hyper-parameters

Hierarchical clustering analysis

- Where to cut the dendogram ?

Methodological Supplements and Class Interpretation

Plan

Introduction to Unsupervised Classification

WIKIPEDIA
Data clustering is a method in data analysis.

It aims to divide a set of data into different "packets" each subset sharing common characteristics, which most often correspond to proximity criteria (computer similarity) that are defined by introducing measures and classes of distance between objects.

Introduction to Unsupervised Classification

- Classification = partitioning of a collection of heterogeneous individuals into a set of homogeneous classes.
- unsupervised = no a priori partition of the n individuals ; number of classes K unknown.

$\Rightarrow$ The objective is to determine the K classes $\mathscr{P}_K = \{C_1, \cdots, C_K\}$ of the n individuals of X such that a class is a grouping of individuals :

- similar to each other (homogeneity in the class) ;
- different from individuals in other classes (well-separated classes).

How to automatically define groups of individuals or variables that are similar ?

Example : quantitative data describing 8 mineral French waters out of 13 variables (only 6 shown in the table).

	saveur.amère	saveur.sucrée	saveur.acide	saveur.salée	saveur.alcaline	appréciation.globale
St Yorre	3.4	3.1	2.9	6.4	4.8	2.9
Badoit	3.8	2.6	2.7	4.7	4.5	2.9
Vichy	2.9	2.9	2.1	6.0	5.0	2.8
Quézac	3.9	2.6	3.8	4.7	4.3	3.5
Arvie	3.1	3.2	3.0	5.2	5.0	2.9
Chateauneuf	3.7	2.8	3.0	5.2	4.6	3.3
Salvetat	4.0	2.8	3.0	4.1	4.5	3.4
Perrier	4.4	2.2	4.0	4.9	3.9	2.8

- From the distances between individuals : what measure of distance ?
- From the links between variables : what measure of link ?

Depends on the nature of the data : quantitative, categorial, or mixed.

Introduction to unsupervised classification

From the Euclidean distances between the individuals ?

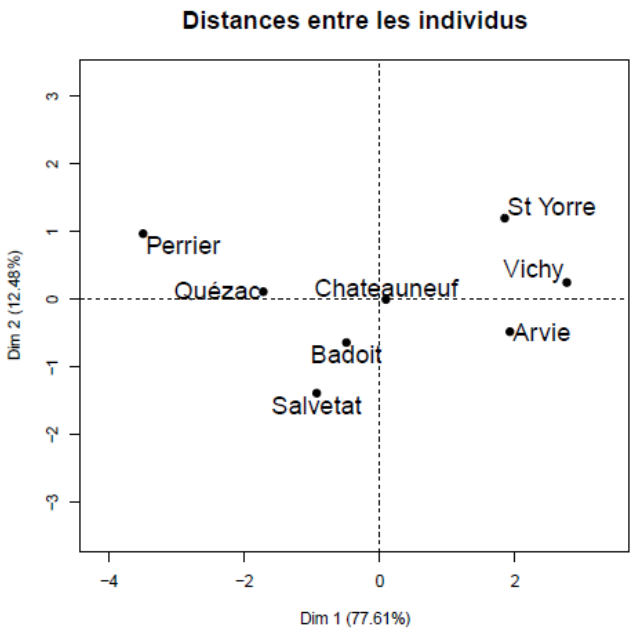
	St Yorre	Badoit	Vichy	Quézac	Arvie	Chateauneuf	Salvetat	Perrier
St Yorre	0.0	4.1	7.9	2.9	3.0	2.9	4.0	8.2
Badoit	4.1	0.0	4.8	5.3	1.8	1.8	1.2	10.6
Vichy	7.9	4.8	0.0	9.7	5.5	5.7	5.4	14.7
Quézac	2.9	5.3	9.7	0.0	4.7	4.3	4.9	6.2
Arvie	3.0	1.8	5.5	4.7	0.0	1.3	1.8	10.1
Chateauneuf	2.9	1.8	5.7	4.3	1.3	0.0	1.6	9.9
Salvetat	4.0	1.2	5.4	4.9	1.8	1.6	0.0	10.3
Perrier	8.2	10.6	14.7	6.2	10.1	9.9	10.3	0.0

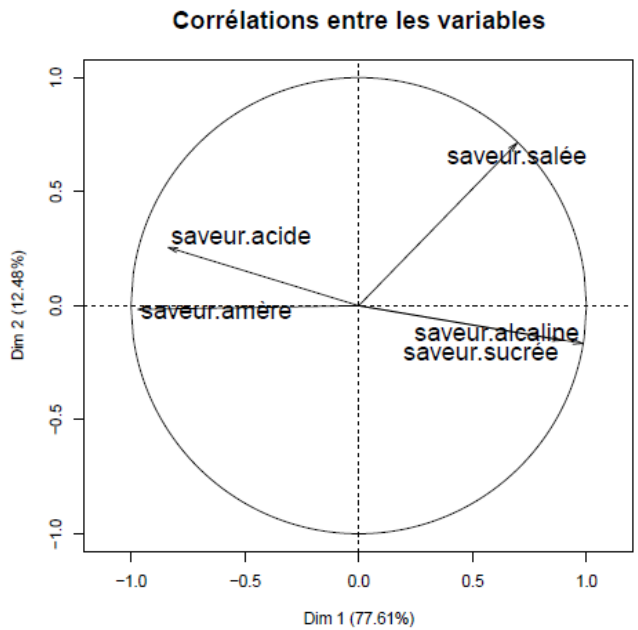
From the correlations between the variables ?

	saveur.amère	saveur.sucrée	saveur.acide	saveur.salée	saveur.alcaline
saveur.amère	1.00	-0.83	0.78	-0.67	-0.96
saveur.sucrée	-0.83	1.00	-0.61	0.49	0.93
saveur.acide	0.78	-0.61	1.00	-0.44	-0.82
saveur.salée	-0.67	0.49	-0.44	1.00	0.56
saveur.alcaline	-0.96	0.93	-0.82	0.56	1.00

Introduction to unsupervised classification

From a principal component analysis (if the data are quantitative) ?



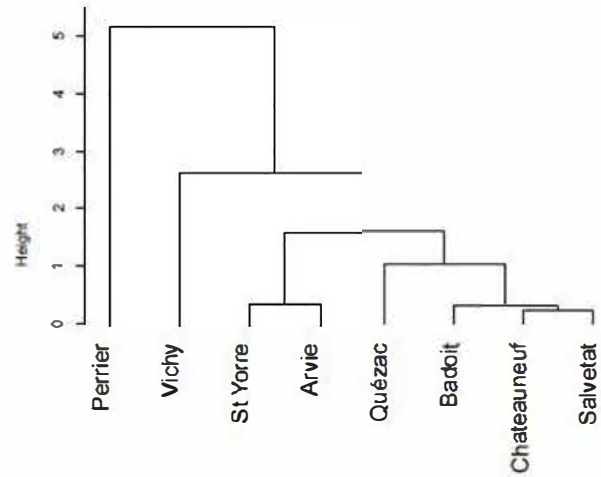


⇒ from a automatic classification (clustering) method.

Partition in 4 classes of individuals.

	P4
St Yorre	1
Badoit	2
Vichy	3
Quézac	2
Arvie	1
Chateauneuf	2
Salvetat	2
Perrier	4

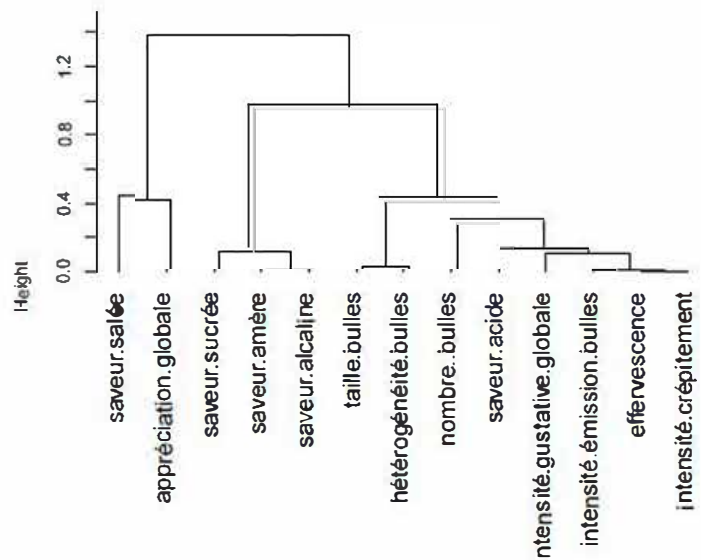
Hierarchy of individuals



Partition in 3 classes of variables.

	P3
saveur.amère	1
saveur.sucrée	1
saveur.acide	2
saveur.salée	3
saveur.alcaline	1
appréciation.globale	3
intensité.émission.bulles	2
nombre..bulles	2
taille.bulles	2
hétérogénéité.bulles	2
effervescence	2
intensité.gustative.globale	2
intensité.crépitement	2

Hierarchy of variables



Introduction to Unsupervised Clustering

There are many automatic clustering algorithms that are distinguished by :

- the nature of the objects to be clustered : individuals or variables,
- the nature of the data : quantitative, categorial, or mixed,
- the nature of the classification structure : partition or hierarchy,
- the nature of the approach used : geometric approach (distance, dissimilarity, similarity) or probabilistic approach (mixture models).

Here, we are interested in the classification of individuals described by quantitative data, using geometric approaches using distances.

Data

We consider a set $\Omega = \{1, \cdots, n\}$ of n individuals described by p quantitative variables. We therefore have the following data

$$X_n = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ & & \vdots & \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}.$$

Individuals in rows, variables in columns.

A weight w_i is associated with individual i. In general :

- $w_i = \frac{1}{n}$ for random observations,
- $w_i \neq \frac{1}{n}$ for adjusted, aggregated data...

The objectives

We therefore have a set of points of $\mathbb{R}^p$ (data) for which we do not know the labels, but that we want to group together in an "intelligent" way.

Production of a classification structure allowing to highlight :

- groups of individuals (classes - clusters) : partitioning methods
- hierarchical links between individuals : hierarchical classification methods.

Partition

A partition $\mathscr{P}$ into K classes of individuals is a set of non-empty classes, two by two disjoint and whose union is the set of individuals.

In other words,

$$\begin{aligned} &C_k \neq \emptyset, && \forall k \in \{1, \cdots, K\} \\ &C_k \cap C_{k'} = \emptyset, && \forall k, k' \in \{1, \cdots, K\} \\ &C_1 \cup \cdots \cup C_K = \Omega \end{aligned}$$

Example : if $\Omega = \{1, \cdots, 7\}$, $\mathscr{P}_3 = (C1, C_2, C_3)$ with

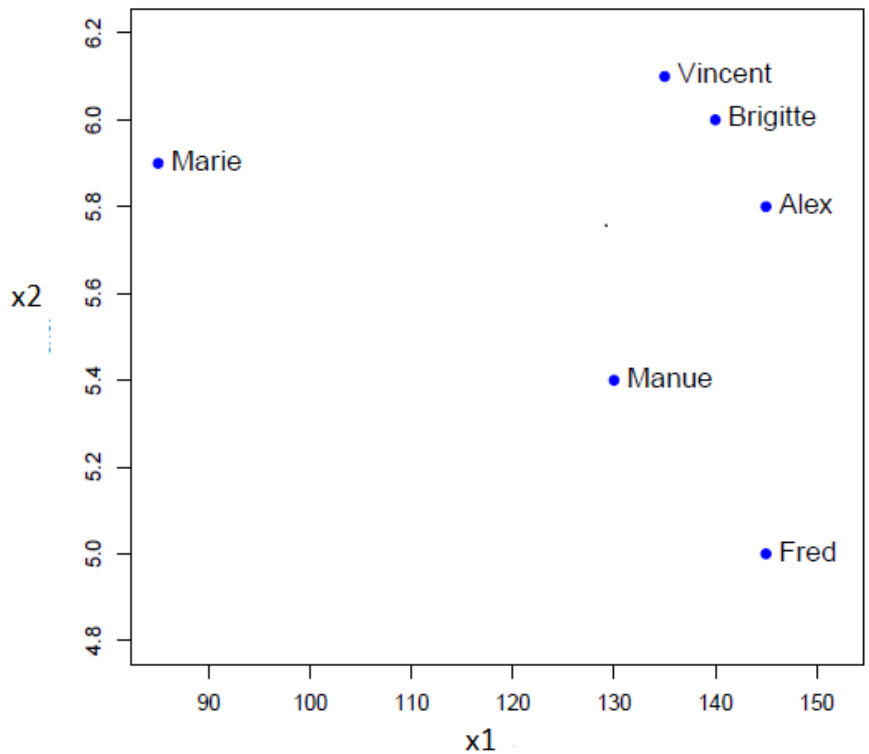
$$C_1 = \{7\},\ C2 = \{5, 1, 6\} \text{ and } C_3 = \{4, 2, 3\}$$

is a partition into three classes of Ω.

Partition or hierarchy ?

Propose a "good" and a "bad" partition $\mathscr{P}_3 = \{C_1, C_2, C_3\}$ into 3 classes of the 6 individuals below.

	x_1	x_2
Brigitte	140	6.0
Marie	85	5.9
Vincent	135	6.1
Alex	145	5.8
Manue	130	5.4
Fred	145	5.0



Hierarchy

A hierarchy H of individuals is a set of non-empty classes that satisfies :

- $\Omega \in H$ *i.e. H contains the class of all individuals,*
- $\forall i \Omega$, $\{i\} \in H$ *i.e. H contains all singletons,*
- $\forall A, B \in H, A \cap B \in \{\emptyset, A, B\}$ *i.e. two classes of H are either disjoint or contained in each other.*

Example : $H = \{\{1\}, \cdots, \{7\}, \{4,5\}, \{2,3\}, \{4,5,6\}, \{1,2,3\}, \{4,5,6,7\}, \Omega\}$.

Hierarchy

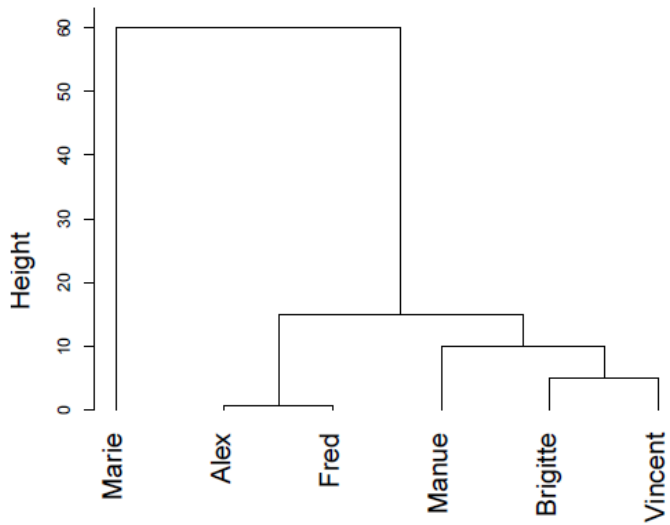
An indexed hierarchy is a pair (H, h) where H is a hierarchy and h is a function from H to $\mathbb{R}^+$ such that :

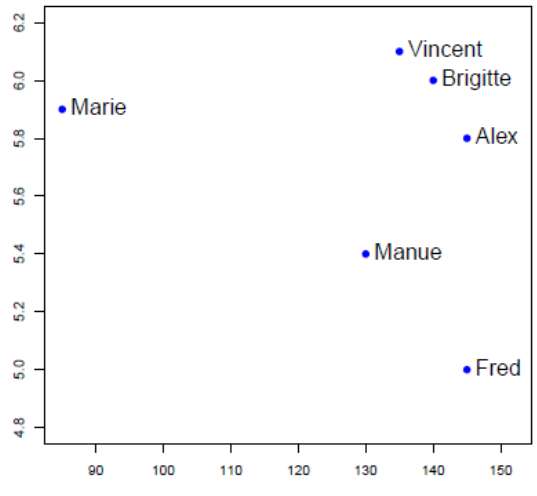
$$\forall A \in H, h(A) = 0 \Longleftrightarrow A \textit{ is a singleton}$$
$$\forall A, B \in H, A \neq B, (A \subset B) \Longrightarrow (h(A) \leqslant h(B)) \textit{ i.e. } h \textit{ increasing.}$$

A dendrogram (or hierarchical tree) is the graphical representation of an indexed hierarchy and the function h measures the height of the classes in this dendrogram.

Hierarchy

What is the hierarchy H of the dendrogram below ?



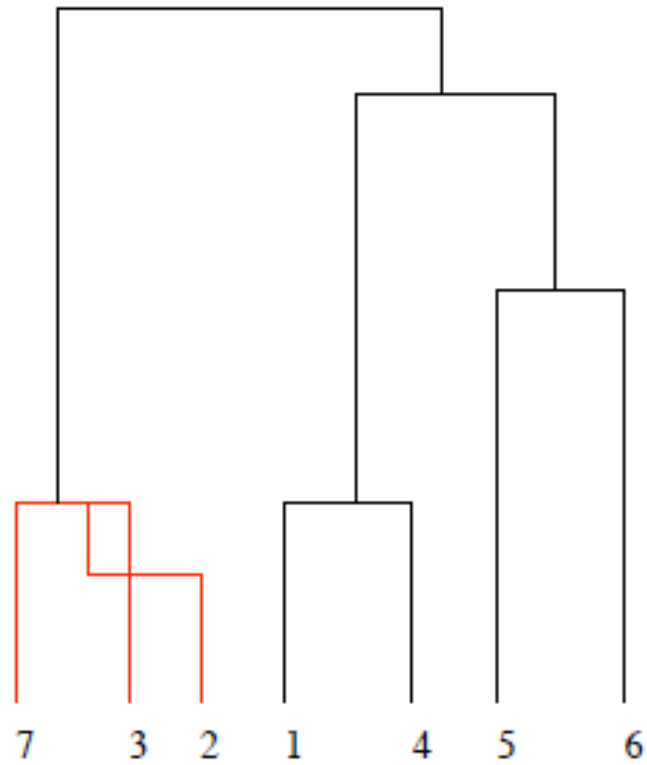


By defining a cut level, we will obtain a partition.

For example, give a hierarchy to 2 classes and another to 4 classes.

Hierarchy

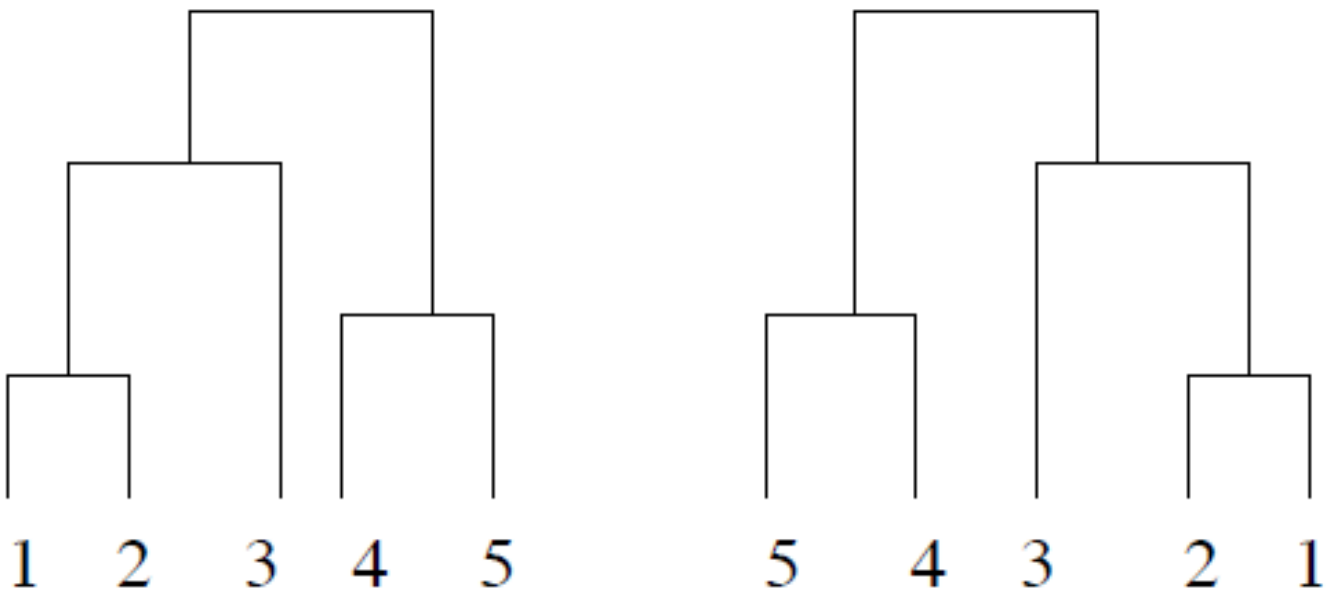
Since the function h is increasing, there is no inversion : if $C = A \cup B$, the class C is higher than classes A and B in the dendrogram.



Hierarchy

We can note that an indexed hierarchy has several equivalent representations. Indeed, the order of the representation of the n individuals at the bottom of the hierarchy can be modified and the number of possible representations is $2n-1$.

Two equivalent representations of the same indexed hierarchy :



How to measure the distance between individuals ? Similarity, dissimilarity or distance ?

Clustering methods require the ability to quantify the dissimilarity between the observations.

☞ *Dissimilarities and distances appropriate according to the type of data.*

A similarity index $s : \Omega \times \Omega \to \mathbb{R}^+$ checks $\forall i,\ i' \in \Omega$:

$$
\begin{aligned}
& s(i, i') \geqslant 0, \\
& s(i, i') = s(i', i), \\
& s(i, i) = s(i', i') = s_{\max} \geqslant s(i, i')
\end{aligned}
$$

Example : for binary data (i.e. vectors composed of 0 and 1), we construct the cross-table between two individuals i and i' :

		1	0	individual i'
individual i	1	a	b	
	0	c	d	

- a = number of attributes that are worth 1 for i and 1 for i' ;
- b = number of attributes that are worth 1 for i and 0 for i' ;
- c = number of attributes that are 0 for i and 1 for i' ;
- d = number of attributes that are 0 for i and 0 for i'.

There are then several normalized similarity indices ($s_{\max} = 1$) :

Jaccard	$\frac{a}{a+b+c}$	Russel and Rao	$\frac{a}{2a+b+c+d}$
Dice or Czekanowski	$\frac{2a}{2a+b+c}$	Ochiai	$\frac{a}{\sqrt{a+b}\sqrt{a+c}}$

How to measure the distance between individuals ? Similarity, dissimilarity or distance ?

A dissimilarity index $d : \Omega \times \Omega \to \mathbb{R}^+$ verifies

$$d(i, i') \geqslant 0, \quad d(i, i') = d(i', i), \quad d(i, i) = 0.$$

Note : it is easy to transform a similarity index s into a dissimilarity index d by setting :

$$d(i, i') = s_{\max} - s(i, i').$$

A distance is a dissimilarity that also verifies the triangle inequality :

$$d(i, j) \leqslant d(i, k) + d(k, j) \quad \forall i, j, k \in \Omega.$$

How to measure the distance between individuals ?

For quantitative data x and y of $\mathbb{R}^p$:

- simple Euclidean distance :

$$d^2(x,y) = \sum_{j=1}^{p} (x_j - y_j)^2.$$

- normalized Euclidean distance : (population $=(x_i)_{i=1,\cdots,n}$)

$$d^2(x,y) = \sum_{j=1}^{p} \frac{1}{s_j^2}(x_j - y_j)^2,$$

  where $s_j^2 = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x}^j)^2$ and $\overline{x}^j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$.
- city-block or Manhattan distance : $d(x,y) = \sum_j |x_j - y_j|$.
- Chebyshev or max distance : $d(x,y) = \max_j |x_j - y_j|$.

How to measure the distance between individuals ?

- In general, we use the simple Euclidean distance when all the variables have the same measurement scale. Indeed, if a variable has a much greater variance, the simple Euclidean distance will give much more importance to the difference between the two individuals on this variable than to the difference between the two individuals on the other variables.
- In the case of measurement scales that are too different, it is preferable to use the normalized Euclidean distance in order to give the same importance to all the variables. This is equivalent to calculating the simple Euclidean distance on the standardized data (centered-reduced).

How to measure the distance between individuals ?

For categorial data x and y with p characteristics :

- simple dissimilarity :

$$d(x,y) = \sum_{j=1}^{p} \mathbb{1}_{x_j \neq y_j},$$

- Rogers and Tanimoto dissimilarity,
- ...

For mixed data x and y : Gower metric,...

How to measure the distance between classes ?

Clustering methods require being able to quantify the dissimilarity between the classes.

Measures of the similarity D between classes :

- Minimal linkage - Single linkage : smallest distance.
- Maximal linkage - Complete linkage : largest distance.
- Average linkage : average distance.
- Ward's linkage : weighted average distance.

Minimum linkage - Simple linkage

Minimum linkage - minimum Eucidian distance between class points :

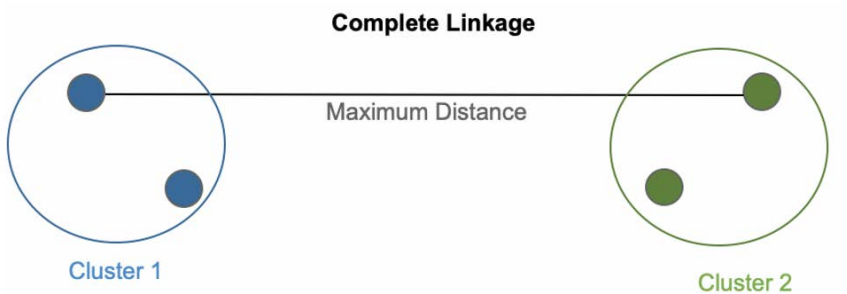
$$D(C_k, C_{k'}) = \min_{x \in C_k,\ x' \in C_{k'}} d(x, x').$$

+ Minimal spanning tree,
- Classes with very different diameters,
- Chaining effect : tendency to aggregation rather than creating new classes,
- Sensitivity to atypical individuals.

Maximal linkage - Complete linkage

Maximal linkage - maximum Eucidean distance between class points :

$$D(C_k, C_{k'}) = \max_{x \in C_k,\ x' \in C_{k'}} d(x, x').$$



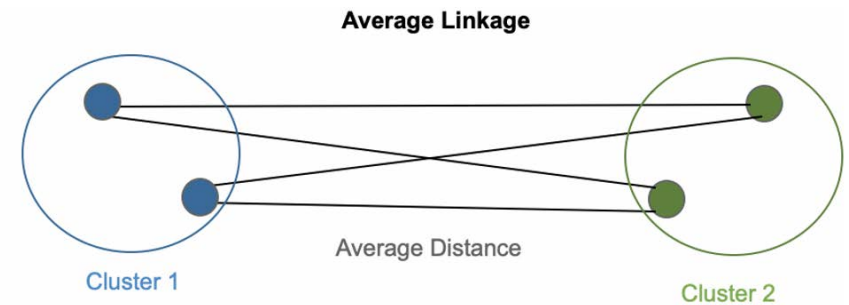
+ Creates compact classes : this merger generates the smallest increase in diameters,
- No separation control : arbitrarily close classes,
- Sensitivity to atypical individuals.

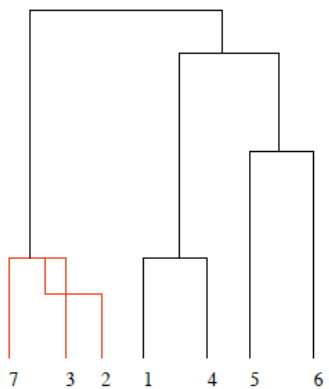
Average linkage

Average linkage - average Eucidean distance between class points :

$$D(C_k, C_{k'}) = \frac{1}{|C_k||C_{k'}|} \sum_{x \in C_k} \sum_{x' \in C_{k'}} d(x, x').$$

+ Trade-off between minimal and maximal links : good balance between class separation and class diameter,
+ Tendency to produce classes of close variance.
- inversions can be observed in the HCA tree.





Ward's linkage

Ward's linkage - weighted distance between class centers :

$$D(C_k, C_{k'}) = \frac{|C_k||C_{k'}|}{|C_k| + |C_{k'}|} d(\mu_k, \mu_{k'})^2.$$

+ Tendency to build classes of the same size for a given level of hierarchy.
+ Groups classes with close barycenters.
+ Favors spherical classes.
+ Breaks the chain effect of the minimum link.
+ No inversion in the HCA tree.
+ Favors the aggregation of low-weight classes (small numbers).

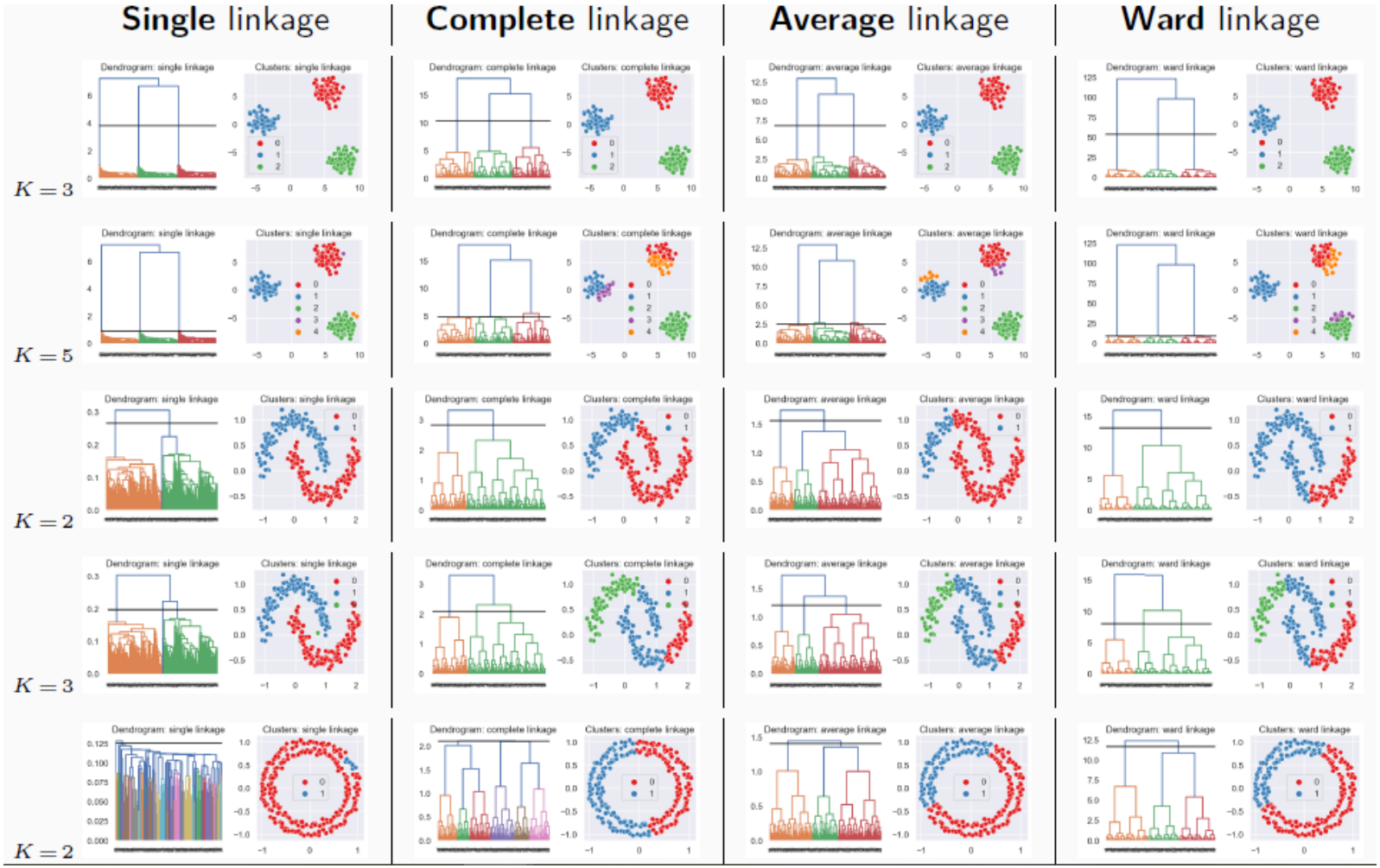


Image : Chevallier (2023).

Exhaustive search ?

Stirling number of the second kind = number of ways to partition a set of n elements into K non-empty subsets :

$$S(n,K) = \frac{1}{K!}\sum_{j=1}^{K}(-1)^{K-j}\binom{K}{j}$$

For example,

- $S(10,3) = 9330$ partitions of $n = 10$ individuals, $K = 3$ classes,
- $S(10,5) = 42525$ partitions of $n = 10$ individuals, $K = 5$ classes.
- $S(100,3) = 10^{47}$ partitions of $n = 100$ individuals, $K = 3$ classes,
- $S(100,5) = 10^{68}$ partitions of $n = 100$ individuals, $K = 5$ classes.

$\Rightarrow$ Exhaustive search impossible ! ! !

Plan

How to evaluate the quality of a partition ?

A good partition into K classes has classes

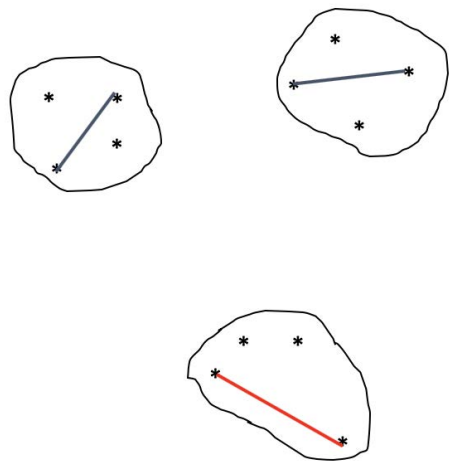
- homogeneous : individuals in the same class are similar,
- separate : individuals from two different classes are not similar.

To obtain a good partitioning, it is therefore appropriate to both :

- minimize the intra-class inertia to obtain the most homogeneous clusters possible ;
- maximize the inter-class inertia to obtain well-differentiated subsets.

How to evaluate the quality of a partition ?

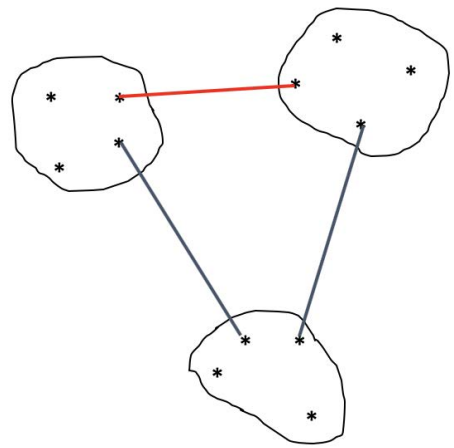
The cohesion of the classes of a partition can be measured by the largest diameter.



This criterion is minimized (approximately) by the maximum link algorithm which will build coherent but not necessarily isolated classes and which are sensitive to outliers.

How to evaluate the quality of a partition ?

The separation of the classes of a partition can be measured by the smallest minimum link.



This criterion is maximized by the minimum link algorithm which will build isolated but not necessarily coherent classes and which can be unbalanced.

How to evaluate the quality of a partition ?

We consider a partition $\mathscr{P}_K = \{C_1, \cdots, C_K\}$ in K classes.

We assume here that the data are quantitative and that the weight of the individuals is $1/n$.

We note μ the center of gravity of the point cloud

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i$$

and for each class k, μ_k the center of gravity of the class k :

$$\mu_k = \frac{1}{|C_k|} \sum_{i \in C_k} x_i \quad \text{for all } k \in K.$$

How to evaluate the quality of a partition ?

Total inertia = total variance

$$I_{Tot} = \sum_{i=1}^{n} d(\mu, x_i)^2$$

The total inertia is independent of the partition.

Inter-class inertia = variance of the class centers

$$I_{Inter} = \sum_{k=1}^{K} |C_k| d(\mu, \mu_k)^2$$

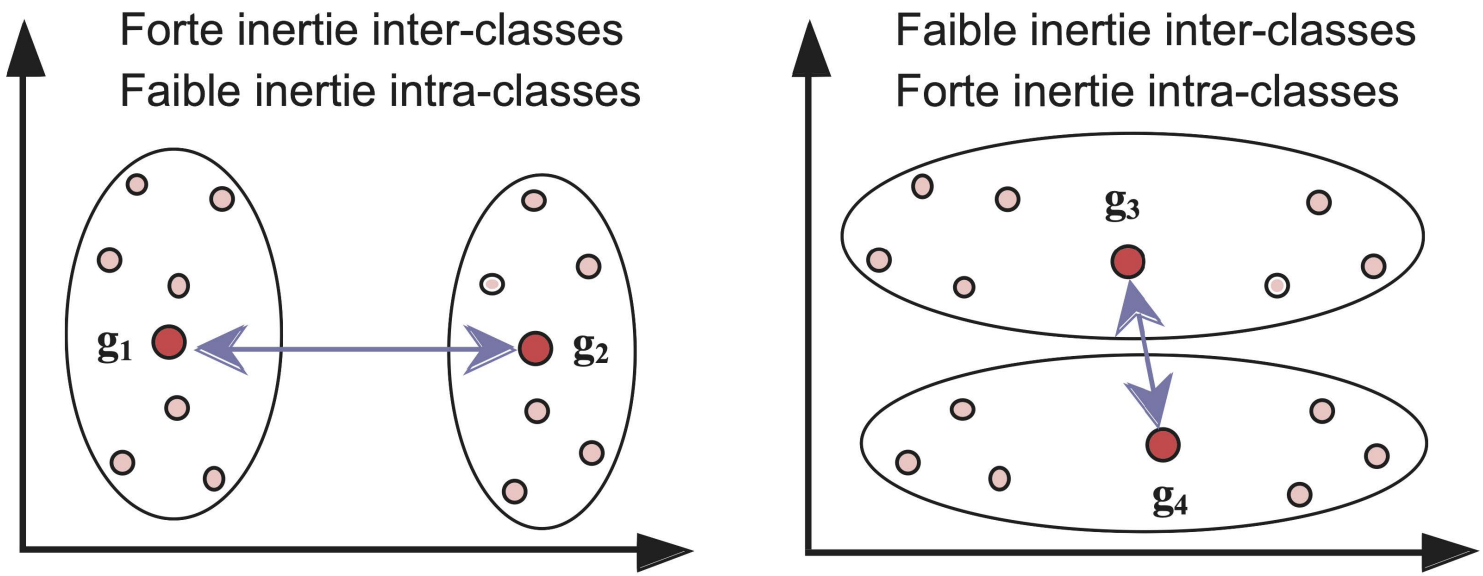
Intra-class inertia = variance of points in the same class

$$I_{Intra} = \sum_{k=1}^{K} \sum_{i \in C_k} d(\mu_k, x_i)^2$$

How to evaluate the quality of a partition ?

Clustering principle :

Minimize intra-inertia (heterogeneity of classes)
⇔ *Maximize inter-inertia (separation of classes).*

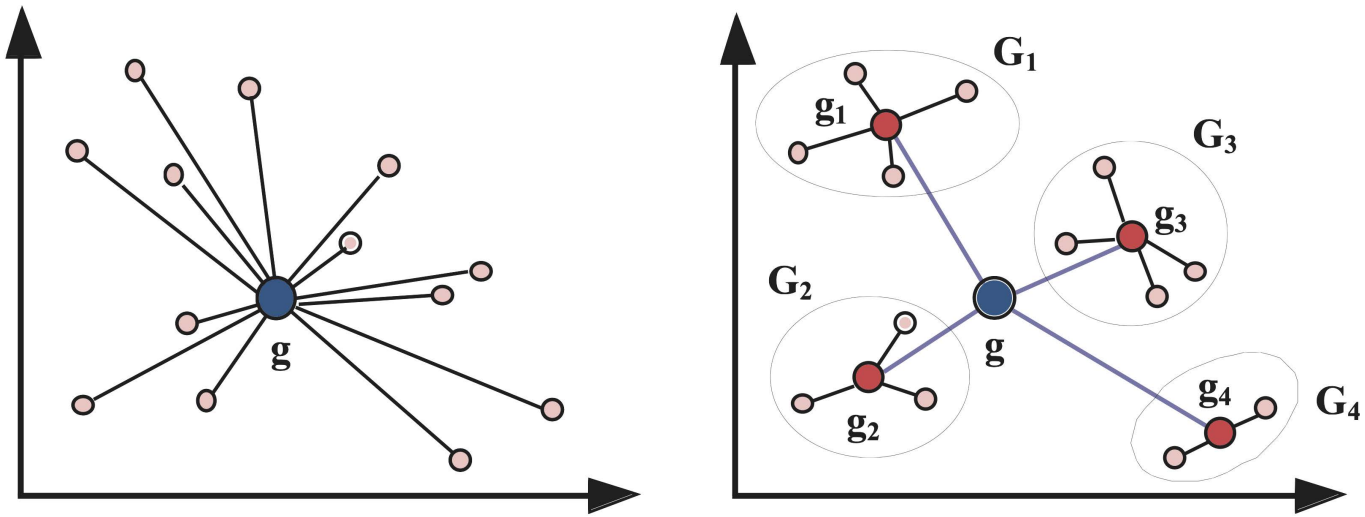


Images : Bisson 2001.

How to evaluate the quality of a score ?

By the Pythagorean theorem,

$$I_{Tot} = I_{Intra} + I_{Inter}.$$



Images : Bisson 2001.

How to evaluate the quality of a partition ?

The quality of a partition can, for example, be measured by :

$$0 \leqslant \frac{\textit{Inter inertia}}{\textit{Total inertia}} \leqslant 1$$

Interpretation of this criterion : it is the percentage explained inertia by the partition.

If $\frac{\text{Inter inertia}}{\text{Total inertia}} = 0$

- the variables have the same means in all classes (mean) ;
- the partition does not allow classification.

If $\frac{\text{Inter inertia}}{\text{Total inertia}} = 1$,

- the individuals in the same class are identical ;
- the partition is ideal for classification.

How to evaluate the quality of a partition ?

This is an external metric, as we will see later.

⚠ This criterion cannot be judged in absolute terms because it depends on the number of individuals and the number of classes.

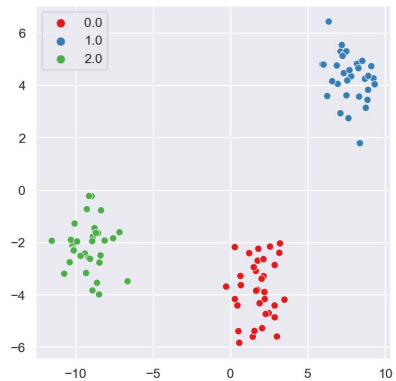
Indeed, it is equal to :

- 1 for the partition into n classes (1 individual per class),
- 0 for the partition into 1 class (containing all individuals).

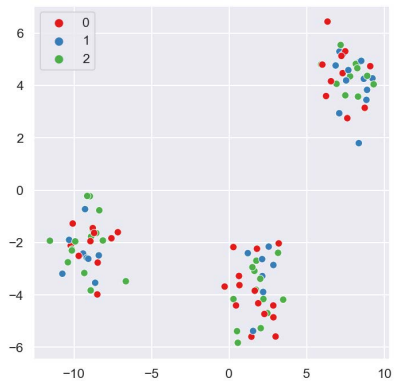
It therefore increases with the number of classes and allows for the comparison of partitions having the same number of classes.

☞ This criterion is maximized (approximately) by Ward's algorithm, which constructs isolated, coherent, and balanced classes.

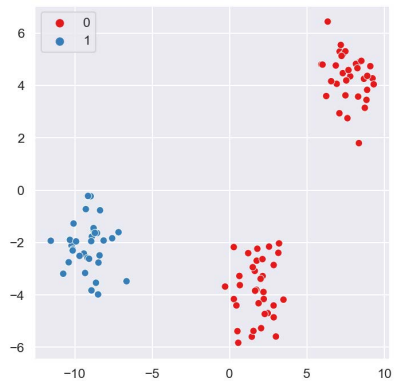
How to evaluate the quality of a score ?



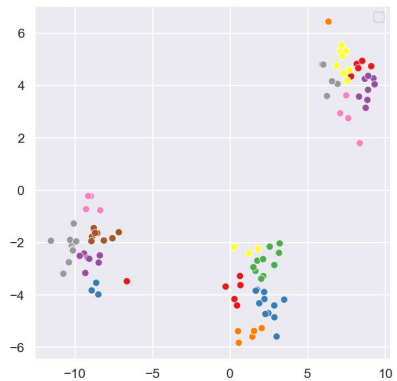
Truth



Random



C2



C20

Internal metric : practical situation - unknown truth :

- silhouette coefficient,
- R-Square (RSQ) and semi-partial R-Square (SPRSQ).

External metric : specific method if we know the truth :

- purity,
- normalized mutual information.

How to evaluate the quality of a partition ?

An example of internal metric : coefficient silhouette

We assume that we have n points and K clusters. Let x_i be a data such that $x_i \in C_k$.

Cohesion : average distance between x_i and the other points of C_k :

$$a(i) = \frac{1}{|C_k| - 1} \sum_{j \in C_k, j \neq i} d(x_i, x_j).$$

Separation : average distance between x_i and the points of the closest classes :

$$b(i) = \min_{\ell \neq k} \frac{1}{|C_\ell|} \sum_{j \in C_\ell} d(x_i, x_j).$$

How to evaluate the quality of a score ?

An example of internal metric : silhouette coefficient

Silhouette coefficient for individual i :

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \in [-1, 1].$$

Silhouette coefficient for all data :

$$S(i) = \frac{1}{n} \sum_{i=1}^{n} s(i) = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i \in C_k} s(i).$$

	T	R	C2	C20
Silhouette	0.83	-0.03	0.66	0.39

How to evaluate the quality of a partition ?

An example of internal metric : criteria based on inertia

Let $\mathscr{P}_K$ be a partition.

- R-square :

$$RSQ(\mathscr{P}_K) = \frac{I_{\text{Inter}}(\mathscr{P}_K)}{I_{\text{Tot}}} = 1 - \frac{I_{\text{Intra}}(\mathscr{P}_K)}{I_{\text{Tot}}}.$$

- Semi-partial R-square :

$$RSQ(\mathscr{P}_K) = \frac{I_{\text{Inter}}(\mathscr{P}_K) - I_{\text{Inter}}(\mathscr{P}_{K-1})}{I_{\text{Tot}}}.$$

How to evaluate the quality of a partition ?

An example of an external metric : purity

Let $\mathscr{P}_K^* = \{C_1^*, \ldots, C_{K^*}^*\}$ be the true partition of the n points.
Consider a partition $\mathscr{P}_K = \{C_1, \ldots, C_K\}$.

$$Purity\ \ (\mathscr{P}_K) = \frac{1}{n} \sum_{k=1}^{K} \max_{\ell \in \{1,\cdots,K^*\}} |C_\ell^* \cap C_k|.$$

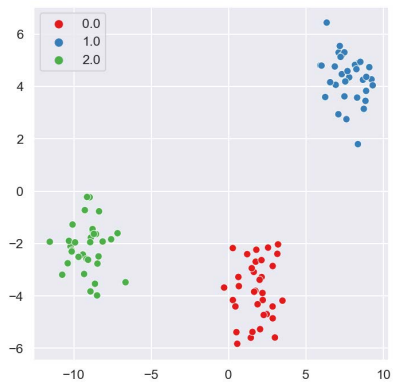
	T	R	C2	C20
Silhouette	0.83	-0.03	0.66	0.39
Purity	1	0.36	0.67	1

Plan

How to compare two partitions ?

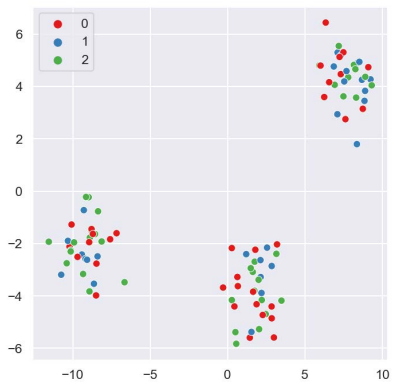
Suppose we have obtained two partitions from the same dataset :

$$\mathscr{P}_K = \{C_1, \ldots, C_K\} \quad \text{and} \quad \mathscr{Q}_L = \{D_1, \ldots, D_L\}.$$



Truth

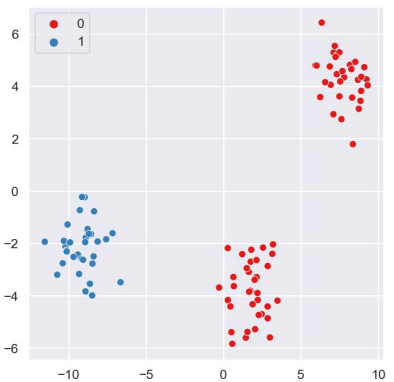
vs



Random

$K = 3$ classes

vs



C_2

$K = 2$ classes

Question : how to compare these two partitions ?

How to compare two partitions ?

Question : how to compare these two partitions ?

- contingency table,
- Rand index (RI) and adjusted Rand index (ARI),
- information variation
- ...

How to compare two partitions ? Rand Index

$\mathscr{P}_K$ vs $\mathscr{Q}_L$	Grouped in $\mathscr{Q}_L$	Separated in $\mathscr{P}_K$
Grouped in $\mathscr{P}_K$	a	b
Separated in $\mathscr{Q}_L$	c	d

$a+d$ = agreements between $\mathscr{P}_K$ and $\mathscr{Q}_L$.
$c+d$ = disagreements between $\mathscr{P}_K$ and $\mathscr{Q}_L$.

Rand index = proportion of pairs of points that are grouped the same way in both partitions :

$$RI(\mathscr{P}_K, \mathscr{Q}_L) = \frac{a+d}{a+b+c+d}.$$

How to compare two partitions ? Adjusted Rand Index

Let $n_{kl} = |C_k \cap D_l|$, $n_{k\bullet} \sum_{l=1}^{L} n_{kl}$ and $n_{\bullet l} \sum_{k=1}^{K} n_{kl}$.

Adjusted Rand index :

$$ARI(\mathscr{P}_K, \mathscr{Q}_L) = \frac{RI - \mathbb{E}[RI]}{\max(RI) - \mathbb{E}[RI]},$$

Or

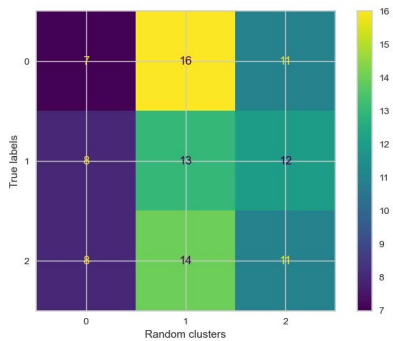
- $RI = \sum_{k,l} \binom{2}{n_{kl}}$,
- $\mathbb{E}[RI] = \frac{\sum_k \binom{2}{n_{k\bullet}} \times \sum_l \binom{2}{n_{\bullet l}}}{\binom{2}{n}}$,
- $\max(RI) = \frac{1}{2}\left(\sum_k \binom{2}{n_{k\bullet}} + \sum_l \binom{2}{n_{\bullet l}}\right)$.

The closer ARI is to 1, the more similar the partitions are.

How to compare two partitions ? Contingency table

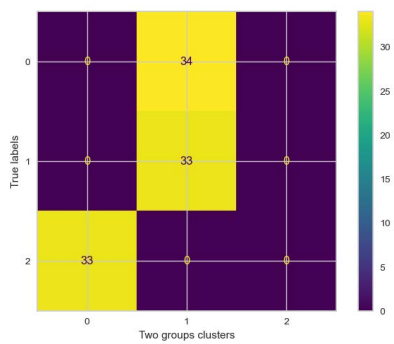
The contingency table allows to observe if classes are divided, grouped...

$\mathscr{P}_K$ \ $\mathscr{Q}_L$	D_1	D_2	$\cdots$	D_L	Total
C_1	n_{11}	n_{12}	$\cdots$	n_{1L}	$n_{1\bullet}$
C_2	n_{21}	n_{22}	$\cdots$	n_{2L}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
C_K	n_{K1}	n_{K2}	$\cdots$	n_{KL}	$n_{K\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet L}$	n



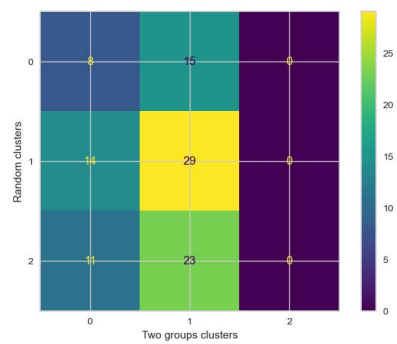
T vs R

vs



T vs C2

vs



R vs C2

Plan

k-means : the algorithm

Optimal partition

- Choose from all the partitions into K classes the one with the greatest inter inertia.
- Problem : number of partitions into K classes of the n individuals $\sim \frac{K^n}{K!}$. K ! .

$\Rightarrow$ Complete enumeration impossible.

Locally optimal partition - Heuristic of the type :

- We start from a feasible solution, i.e. a partition $\mathscr{P}_K^0$.
- At step $t+1$, we look for a partition $\mathscr{P}_K^{t+1} = g(\mathscr{P}_K^t)$ such that inertia_inter$(\mathscr{P}_K^{t+1}) \geqslant$ inertia_inter$(\mathscr{P}_K^t)$.
- Stop when no individual changes class between two iterations.

$\Rightarrow$ Method for partitioning K-means.

k-means : the algorithm

Initialization - Choosing the number of classes K + random drawing of K class centers (K individuals among n).

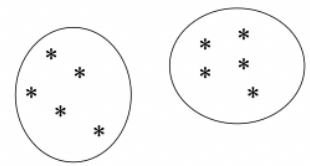
Iteration - Repeat until convergence :

- assignment step : each individual is assigned to the class whose center of gravity is the closest.
- representation step : the centers of gravity of the classes are calculated.

Tirage au hasard des centres

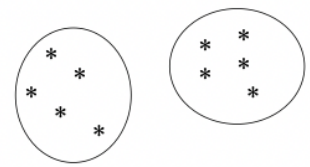
Affectation aux centres les plus proches

Calcul des centres de gravité

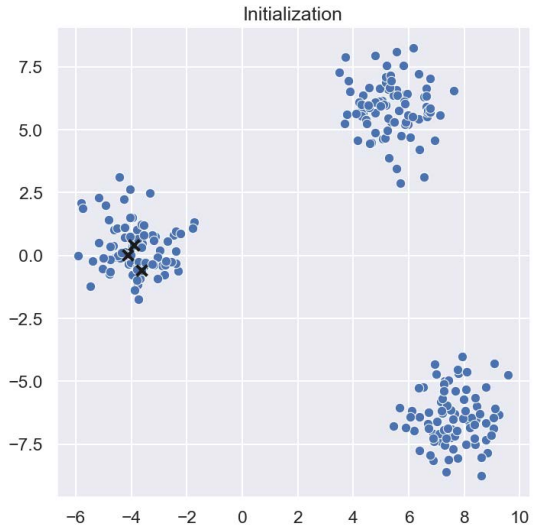


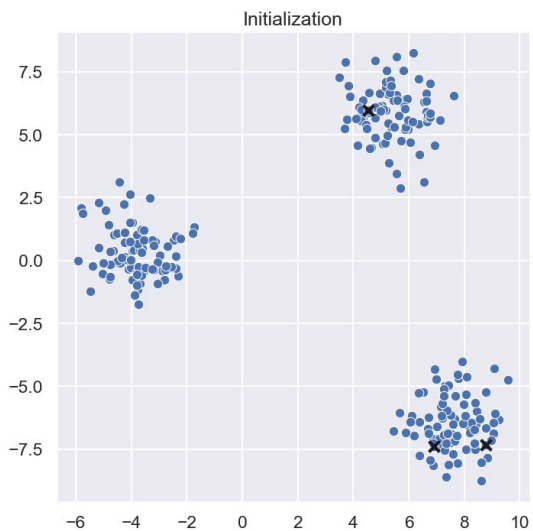
Affectation nouvelle partition

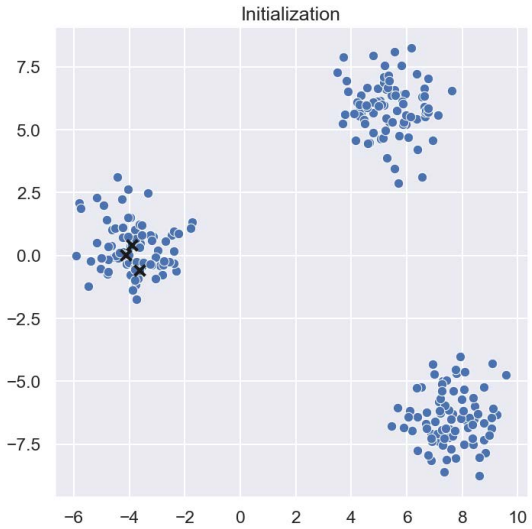
Calcul des centres de gravité

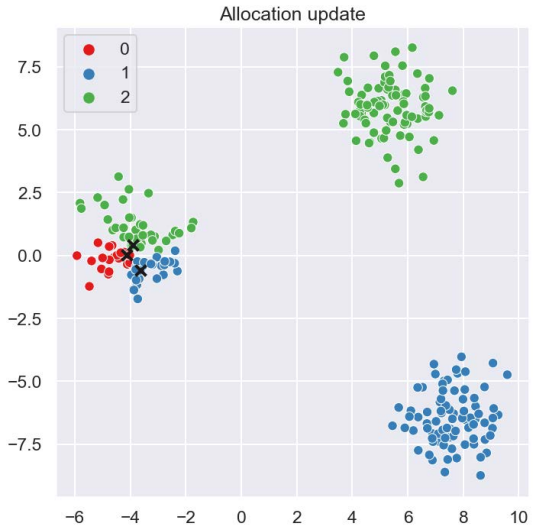


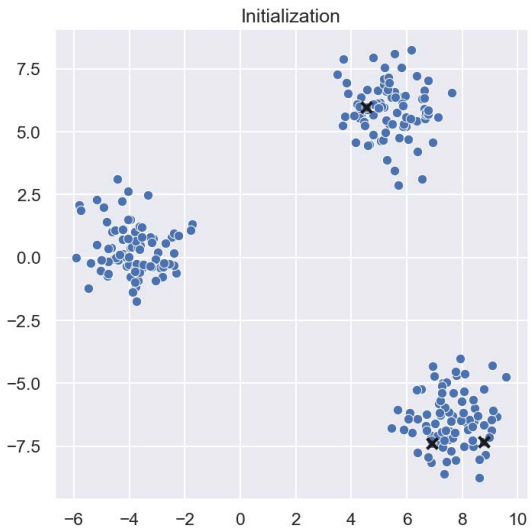
Affectation et arrêt de l'algorithme

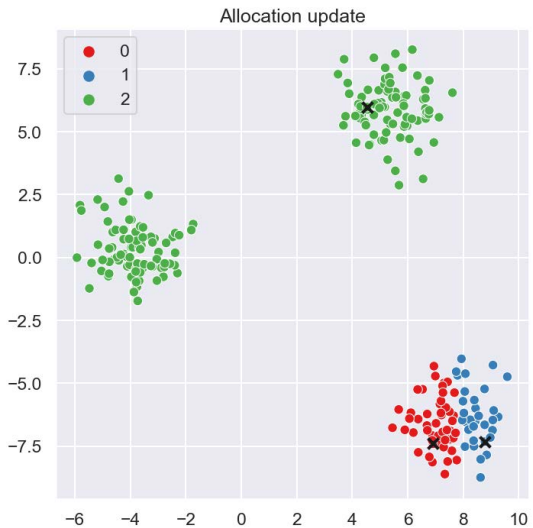

Initialization
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−6
−4
−2
0
2
4
6
8
10

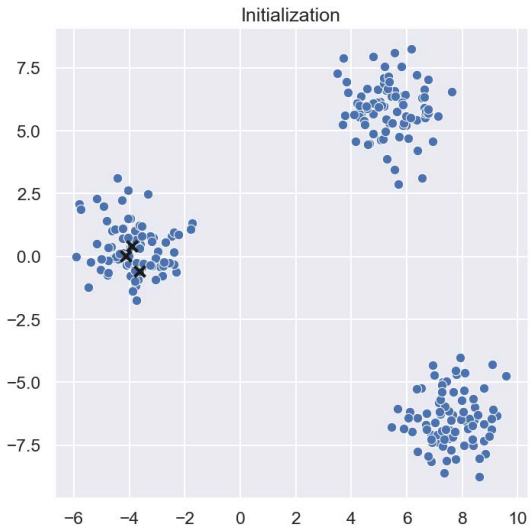

Initialization
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−6
−4
−2
0
2
4
6
8
10


Initialization
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−6
−4
−2
0
2
4
6
8
10

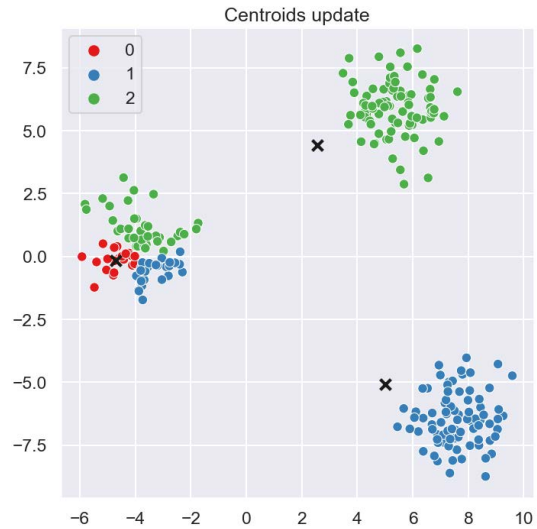

Allocation update
0
1
2
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−6
−4
−2
0
2
4
6
8
10

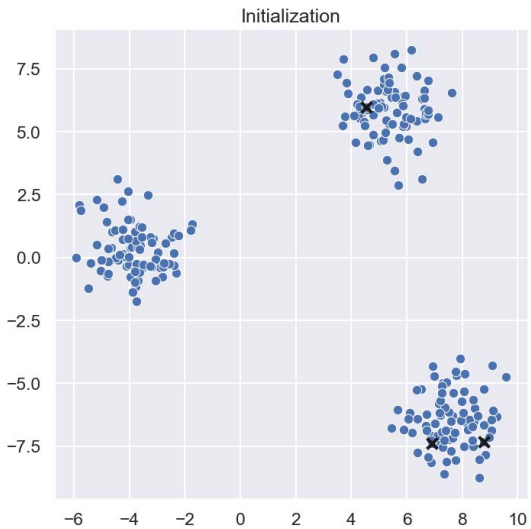

Initialization
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−6
−4
−2
0
2
4
6
8
10

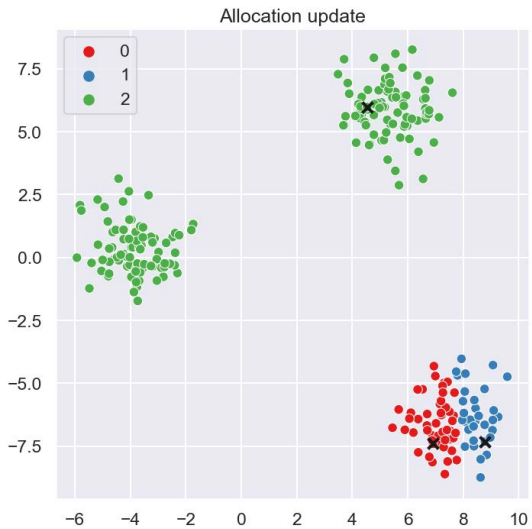

Allocation update
0
1
2
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−6
−4
−2
0
2
4
6
8
10

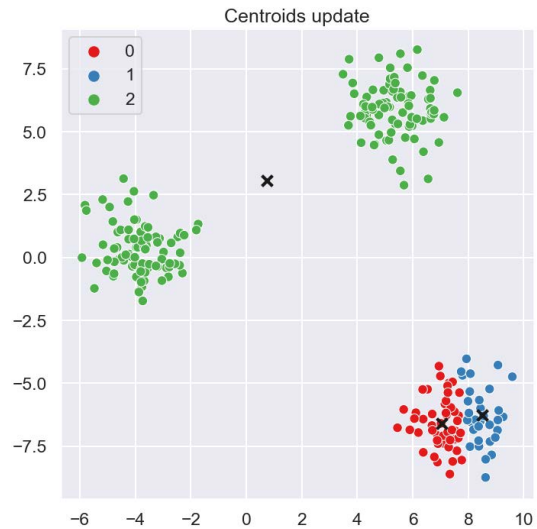

Initialization

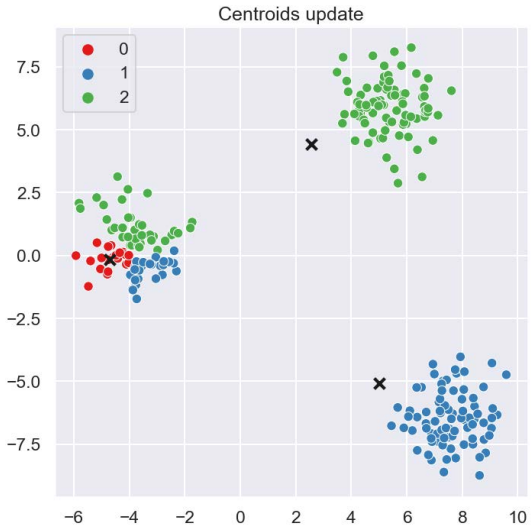

Allocation update
0
1
2

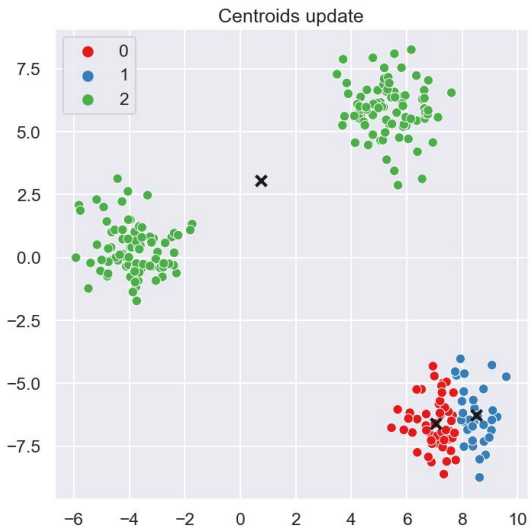

Centroids update
0
1
2

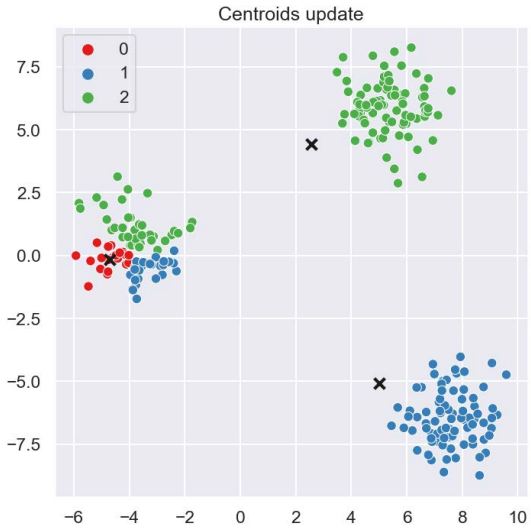

Initialization

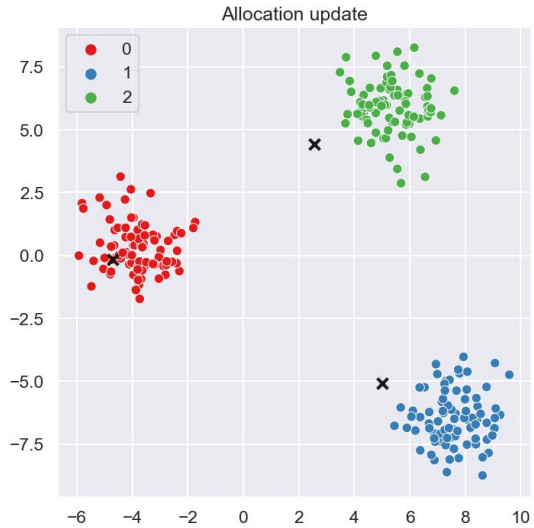

Allocation update
0
1
2

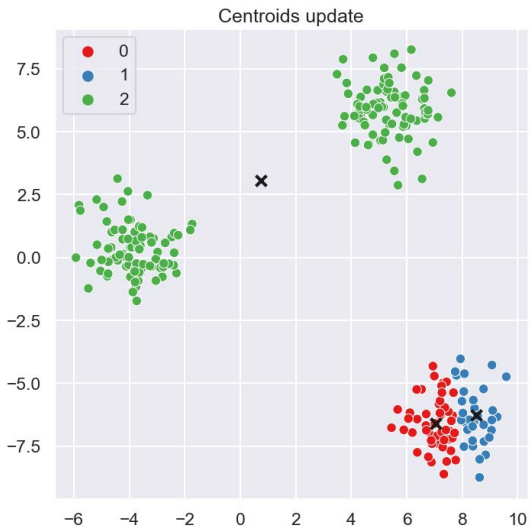

Centroids update
0
1
2

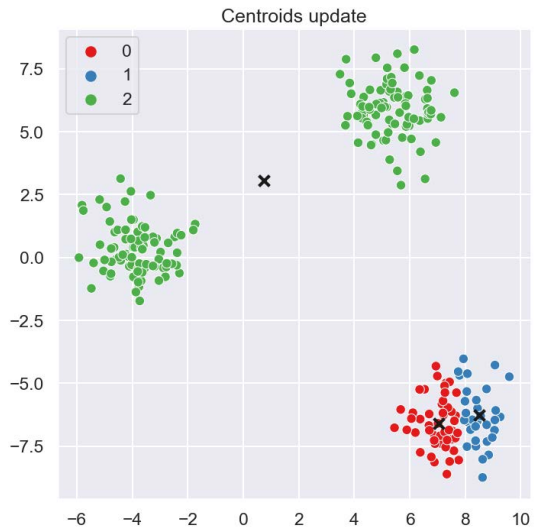

Centroids update
0
1
2
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−6
−4
−2
0
2
4
6
8
10

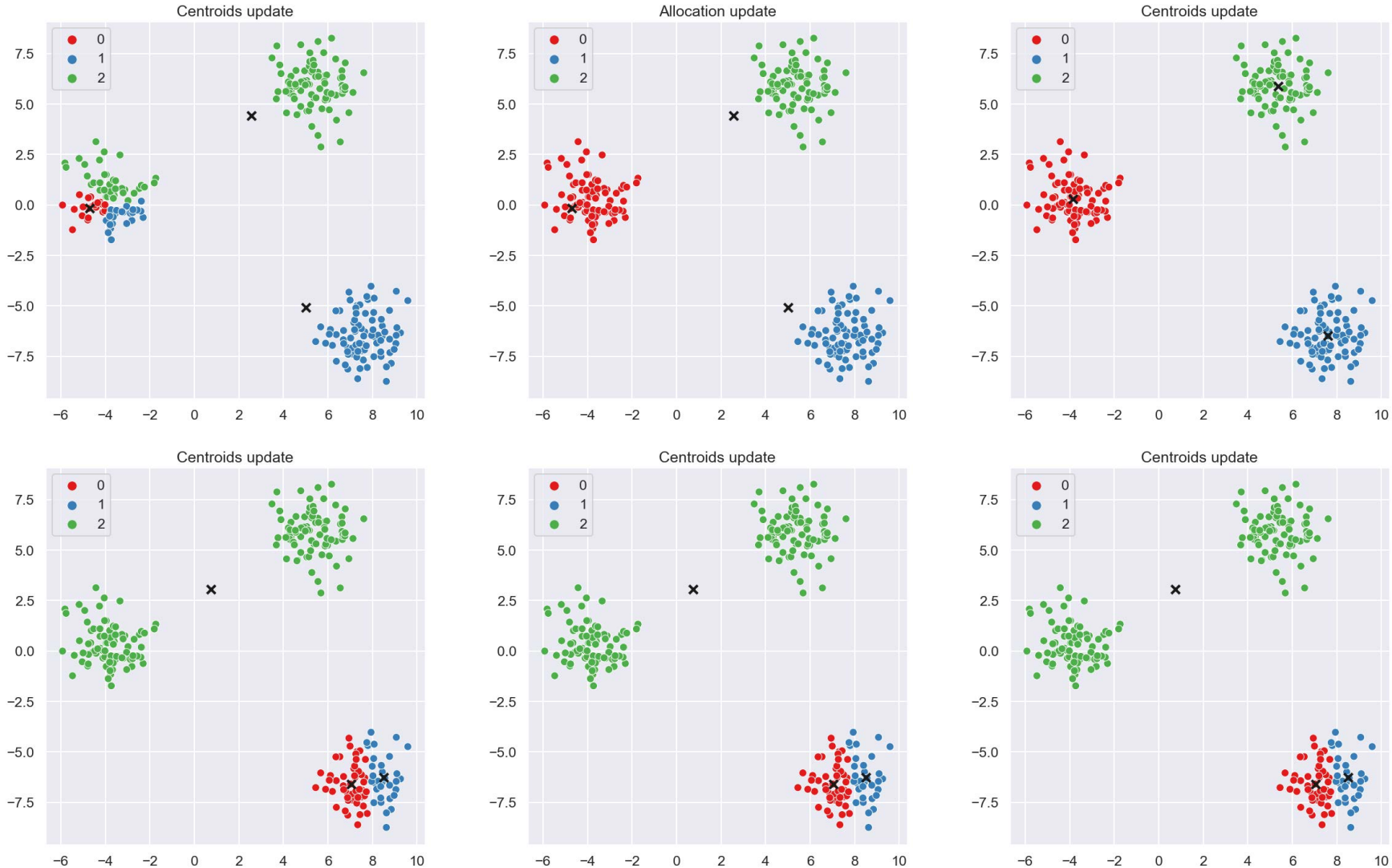

Centroids update
0
1
2
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−6
−4
−2
0
2
4
6
8
10


Centroids update
0
1
2
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−6
−4
−2
0
2
4
6
8
10


Allocation update
0
1
2
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−6
−4
−2
0
2
4
6
8
10


Centroids update
0
1
2
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−6
−4
−2
0
2
4
6
8
10


Centroids update
0
1
2
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−6
−4
−2
0
2
4
6
8
10



Images : Chevallier (2023).

HCA : another example

Example of temperature data

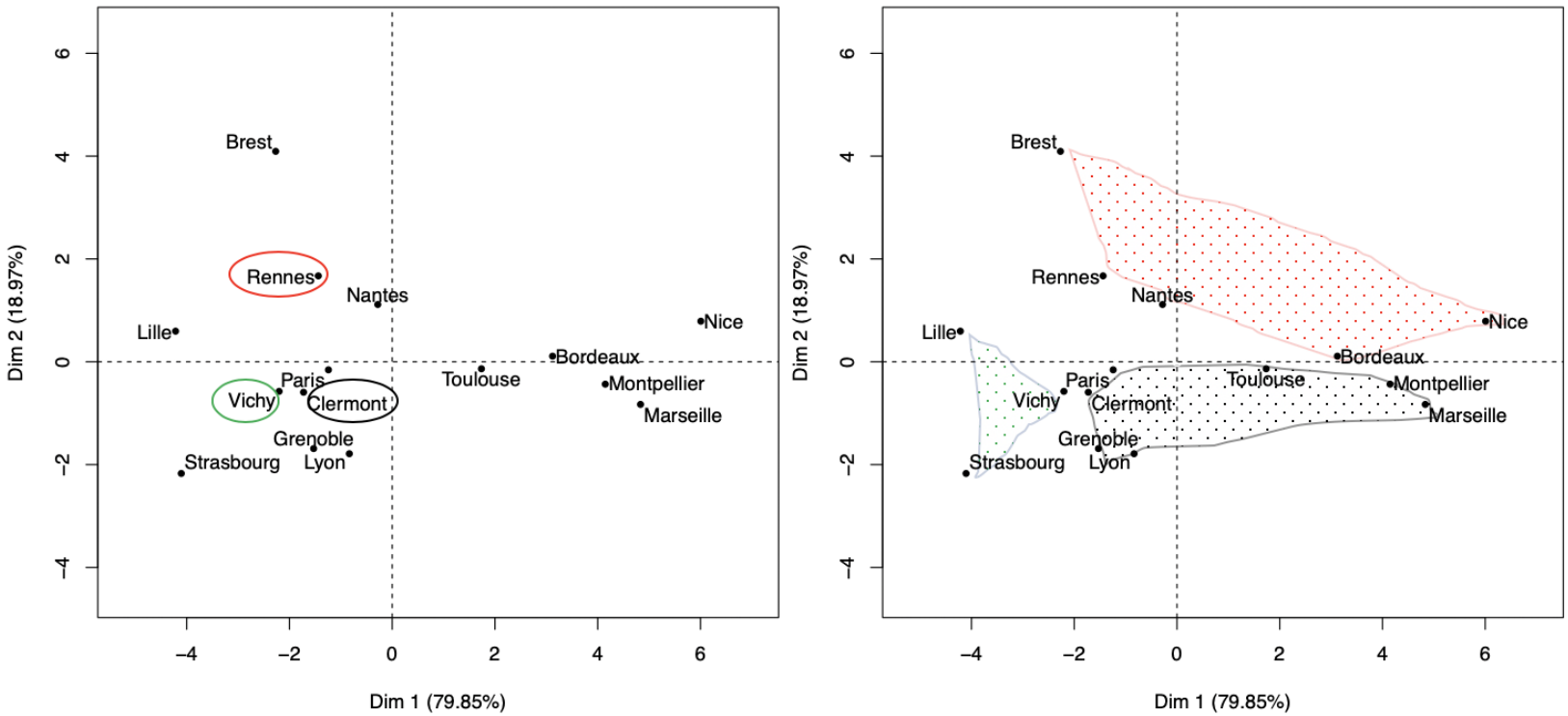
- 15 individuals : French cities,
- 12 variables : average monthly temperatures (over 30 years).

	Janv	Févr	Mars	Avri	Mai	Juin	juil	Août	Sept	Octo	Nove	Déce
Bordeaux	5.6	6.6	10.3	12.8	16	19	21	21	19	13.8	9.1	6.2
Brest	6.1	5.8	7.8	9.2	12	14	16	16	15	12.0	9.0	7.0
Clermont	2.6	3.7	7.5	10.3	14	17	19	19	16	11.2	6.6	3.6
Grenoble	1.5	3.2	7.7	10.6	14	18	20	20	17	11.4	6.5	2.3
Lille	2.4	2.9	6.0	8.9	12	15	17	17	15	10.4	6.1	3.5
Lyon	2.1	3.3	7.7	10.9	15	18	21	20	17	11.4	6.7	3.1
Marseille	5.5	6.6	10.0	13.0	17	21	23	23	20	15.0	10.2	6.9
Montpellier	5.6	6.7	9.9	12.8	16	20	23	22	19	14.6	10.0	6.5
Nantes	5.0	5.3	8.4	10.8	14	17	19	19	16	12.2	8.2	5.5
Nice	7.5	8.5	10.8	13.3	17	20	23	22	20	16.0	11.5	8.2
Paris	3.4	4.1	7.6	10.7	14	18	19	19	16	11.4	7.1	4.3
Rennes	4.8	5.3	7.9	10.1	13	16	18	18	16	11.6	7.8	5.4
Strasbourg	0.4	1.5	5.6	9.8	14	17	19	18	15	9.5	4.9	1.3
Toulouse	4.7	5.6	9.2	11.6	15	19	21	21	18	13.3	8.6	5.5
Vichy	2.4	3.4	7.1	9.9	14	17	19	19	16	11.0	6.6	3.4

Which cities have similar meteorological profiles ?

k-means : another example

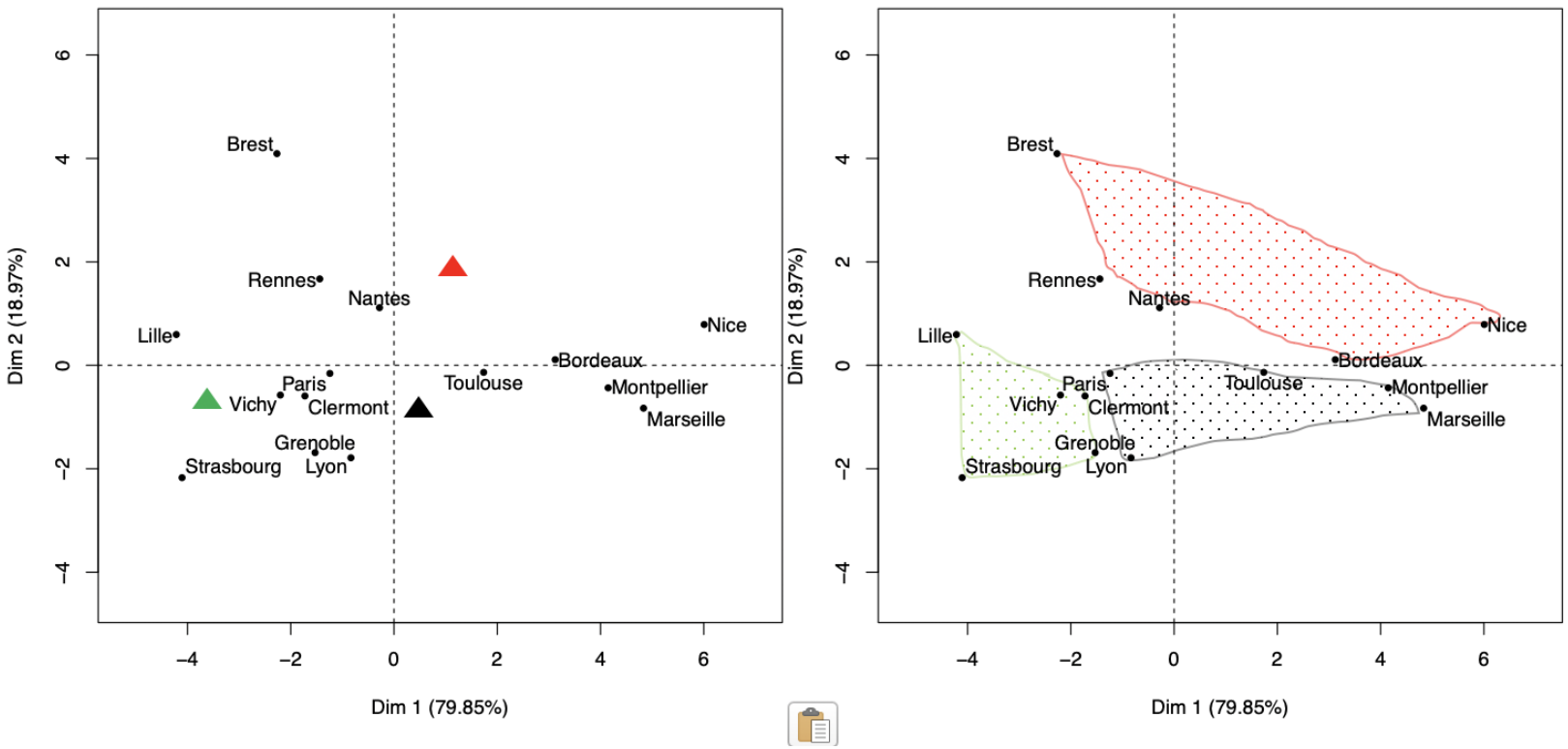
Example of temperature data



Images : Chavent (2020).

k-means : another example

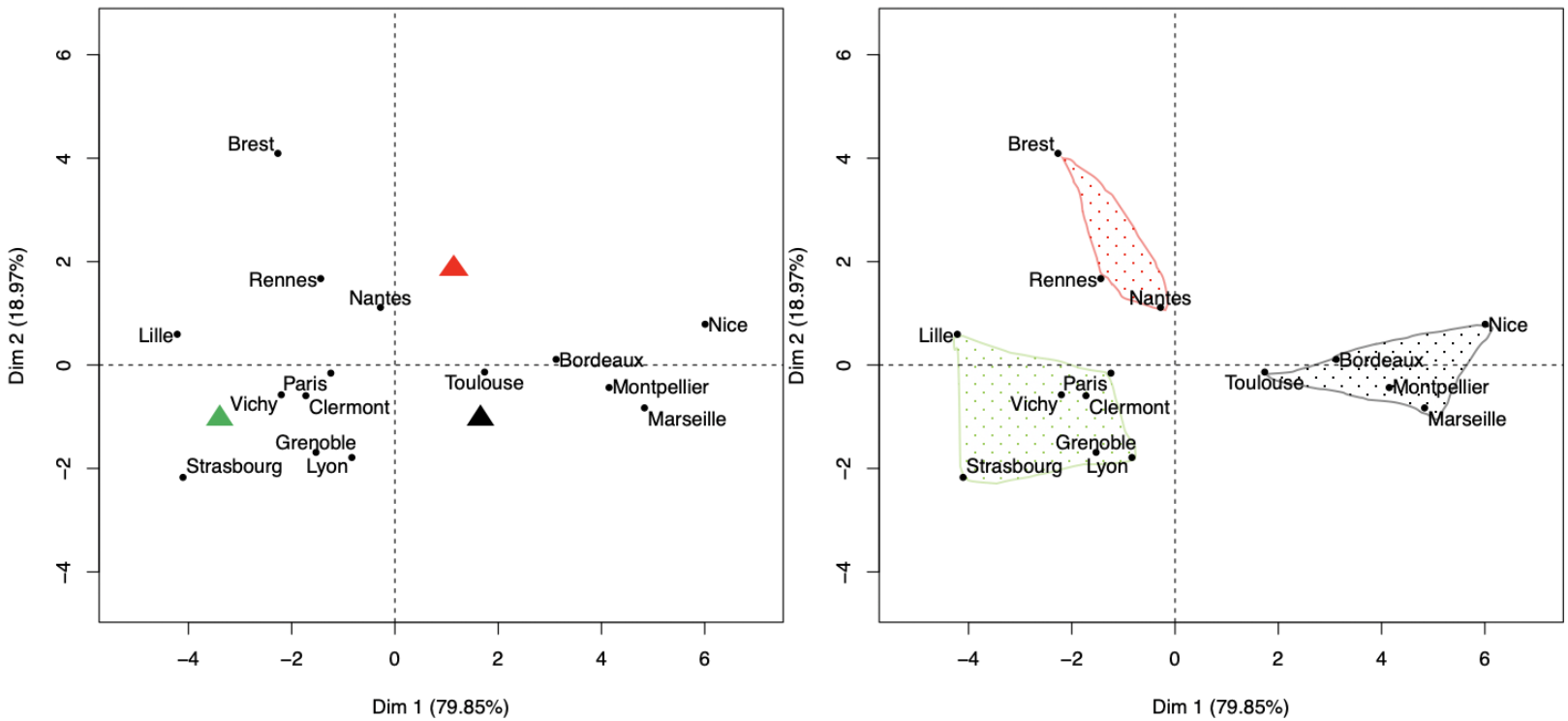
Example of temperature data



Images : Chavent (2020).

k-means : another example

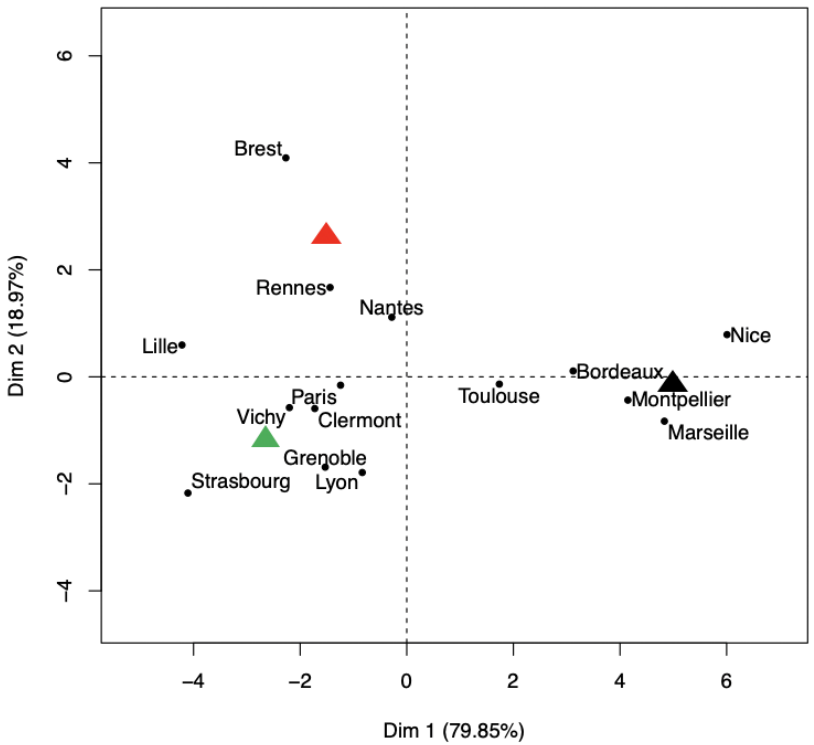
Example of temperature data

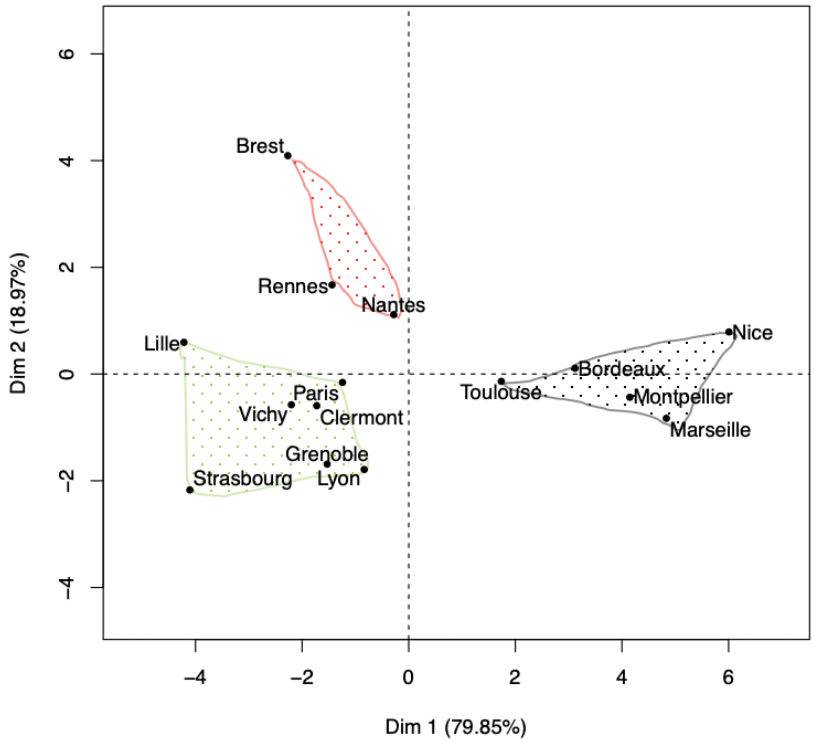


Images : Chavent (2020).

k-means : another example

Example of temperature data

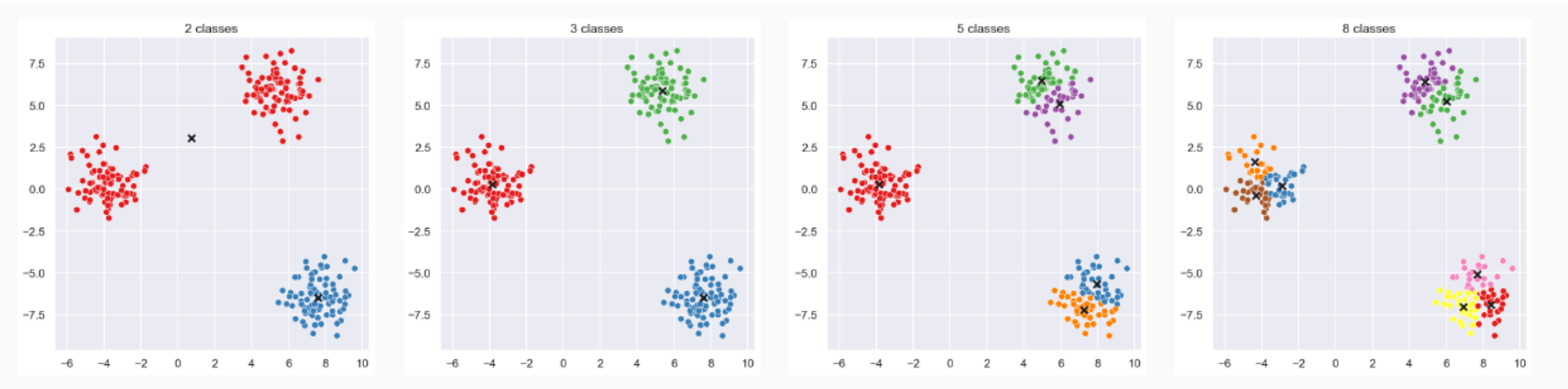




Images : Chavent (2020).

Is this partition into 3 classes different from that resulting from Ward's HCA ?

k-means : choice of the number of classes



Elbow method for intra-class inertia I_{intra} :

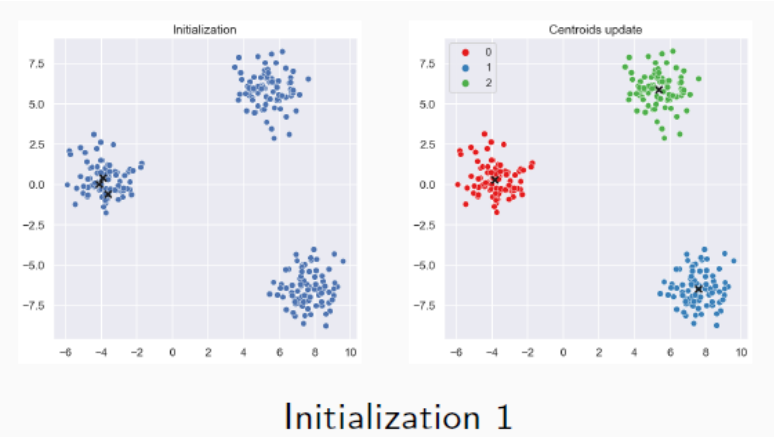
- For each value of $K \in \{2, \cdots, Kmax\}$, we obtain a classification $\mathscr{P}_K$.
- We choose the one where we observe a significant jump in intra-class inertia.

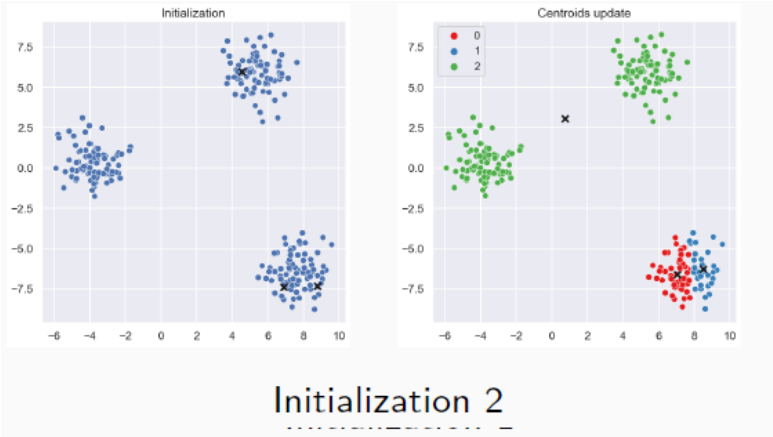
R-square : elbow on the curve $K \mapsto RSQ(K)$.

Semi-partial R-square : greater reduction of SPRSQ.

Silhouette criterion : closer to 1.

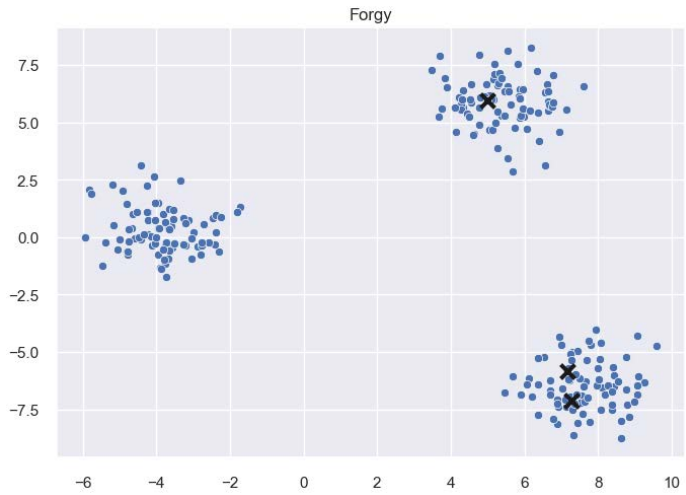
k-means : choice of initial centroids

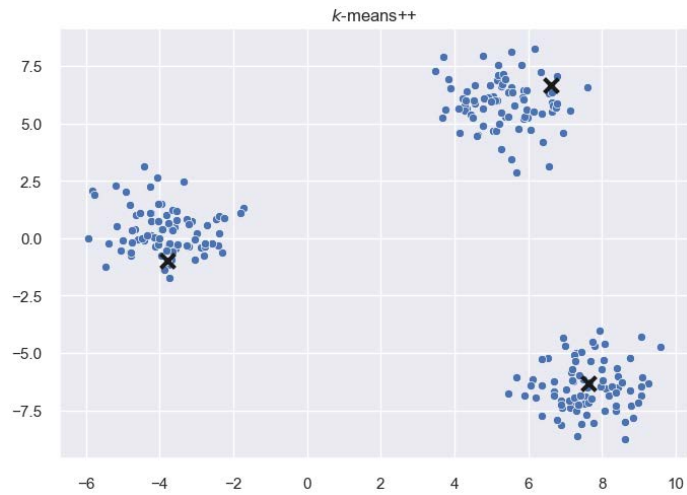




- A judicious choice can favor convergence towards a global minimum !
- Selection based on additional knowledge, or on a preliminary study of the data : histograms, etc.
- Repeat the method N times, then select the partition $\mathscr{P}_K$ with the lowest intra-class inertia.

k-means : choosing initial centroids





Random assignement - K random points in the scatterplot.

k-means++ assignement

Choose 1 centroid randomly.

2nd centroid at a large distance from the 1st with large prob. : sample a point according to 1 prob. law propor. to the distance to the 1st centroid...

k-means : strengths and weaknesses

The final partition depends on the initial partition : if we restart the algorithm with other initial centers, the final partition can be different.

In practice,

- we run the algorithm N times with different random initializations.
- we retain among the N final partitions, the one with the largest percentage of explained inertia.

k-means : strengths and weaknesses

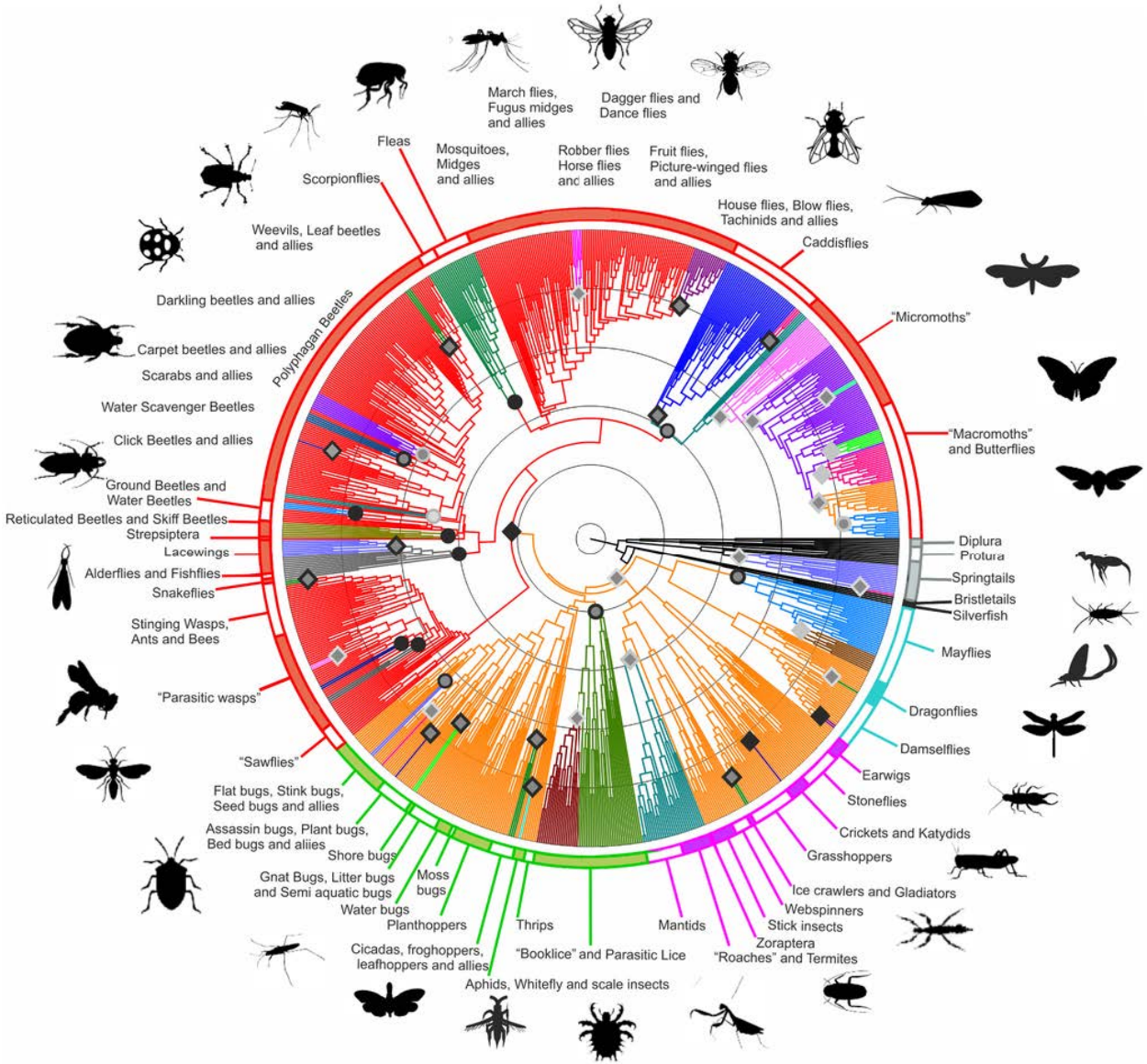
Advantages

- Relatively efficient (fast).
- Linear complexity in the number of individuals.
- Tends to reduce intra-class inertia at each iteration.
- Tends to increase inter-class inertia at each iteration.
- Forms compact and well-separated classes.

Disadvantages

- Choice of the number of classes.
- Influence of the choice of initial centroids.
- Convergence towards a local minimum of intra-class inertia.
- Convergence towards a local maximum of the inter-class inertia.
- Requires the notion of center of gravity.
- Influence of outliers (due to the mean).

Plan



Phylogenetic tree (image : Wikipedia).

Hierarchical clustering analysis(HCA)

Objective : Build a hierarchy.

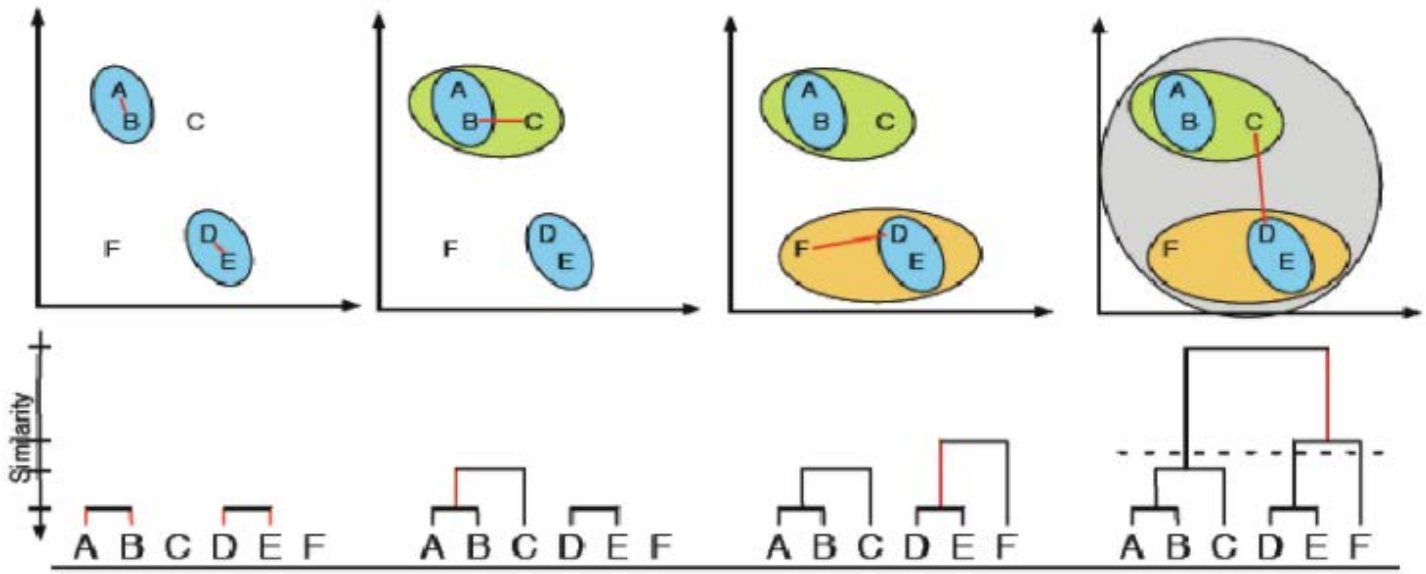
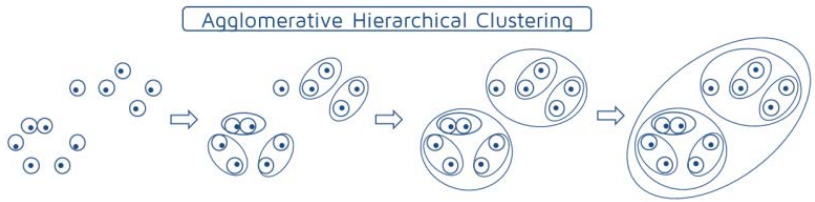


Image : Janssen (2012).

Hierarchical clustering analysis (HCA)

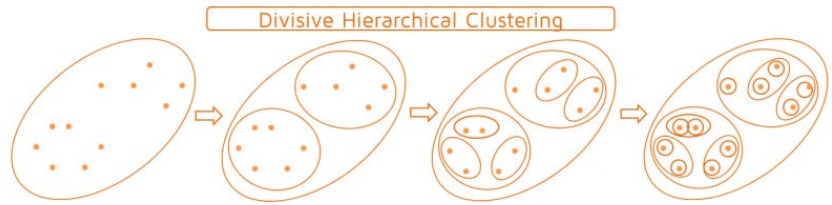
First strategy Agglomerative hierarchical clustering
- Start from the bottom of the dendrogram (singletons),
- Add the closest parts two by two until you get a single class.

⚠ How to choose the classes to aggregate ?



Second strategy Divide the hierarchical clustering
- Start from the top of the dendrogram (one unique class),
- Successive divisions until you get classes reduced to singlets.

⚠ How to choose the classes to divide ?



Images : Chevallier (2023).

HCA : the algorithm

Initialization - We consider an aggregation measure D and we start from the initial partition of singletons $\mathscr{P}_n^{(0)} = \{\{x_1\}, \ldots, \{x_n\}\}$.

Iteration t - From the partition $\mathscr{P}_K^{(t)} = \{C_1, \ldots, C_K\}$ into K classes,

- aggregate the two classes C_k and $C_{k'}$ that minimize D : $C_{k \cup k'} = C_k \cup C_{k'}$
- form the partition with $K-1$ classes : $\mathscr{P}_{K-1}^{(t+1)} = \{C_1, \ldots, C_{k \cup k'}, \ldots, C_K\}$.

End - Repeat until you get the single-class partition.

HCA : the algorithm

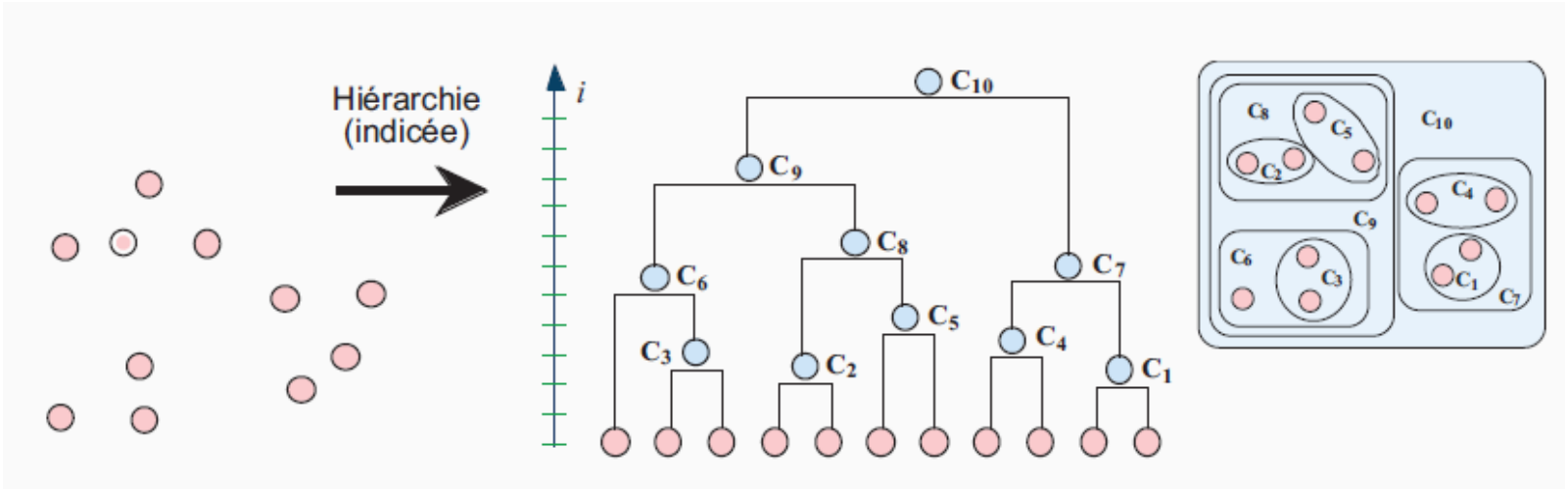


Image : Bisson (2001).

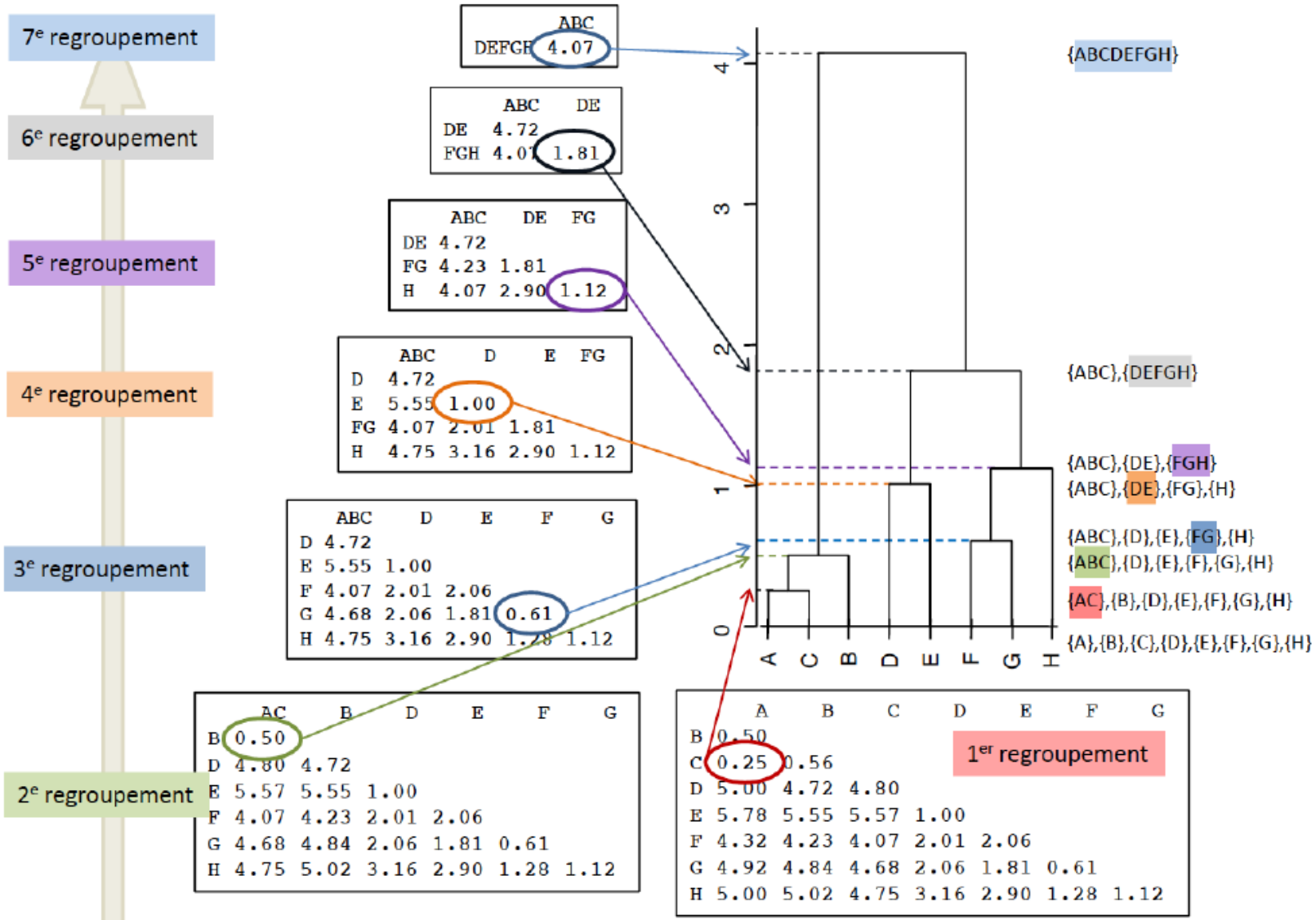
Missing building blocks for implementation

1. Choosing a dissimilarity/distance d between points.
   To be done according to the type of data : categorial, quantitative, etc.
2. Choosing an aggregation measure D between classes.
3. Construction of a dendrogram (not unique !).
4. Criterion to cut the dendrogram to deduce a classification of the data.

HCA : an example with the minimum link

Example : 8 points of $\mathbb{R}^2$ - calculation of Euclidean distances

	A	B	C	D	E	F	G	H
A	0							
B	0.50	0						
C	0.25	0.56	0					
D	5.00	4.72	4.80	0				
E	5.78	5.55	5.57	1.00	0			
F	4.32	4.23	4.07	2.00	2.10	0		
G	4.92	4.84	4.68	2.10	1.80	0.61	0	
H	5.00	5.02	4.75	3.20	2.90	1.28	1.12	0



Data analysis MOOC of François Husson (in French).

HCA : an example with the minimum link

Some details After the first step :

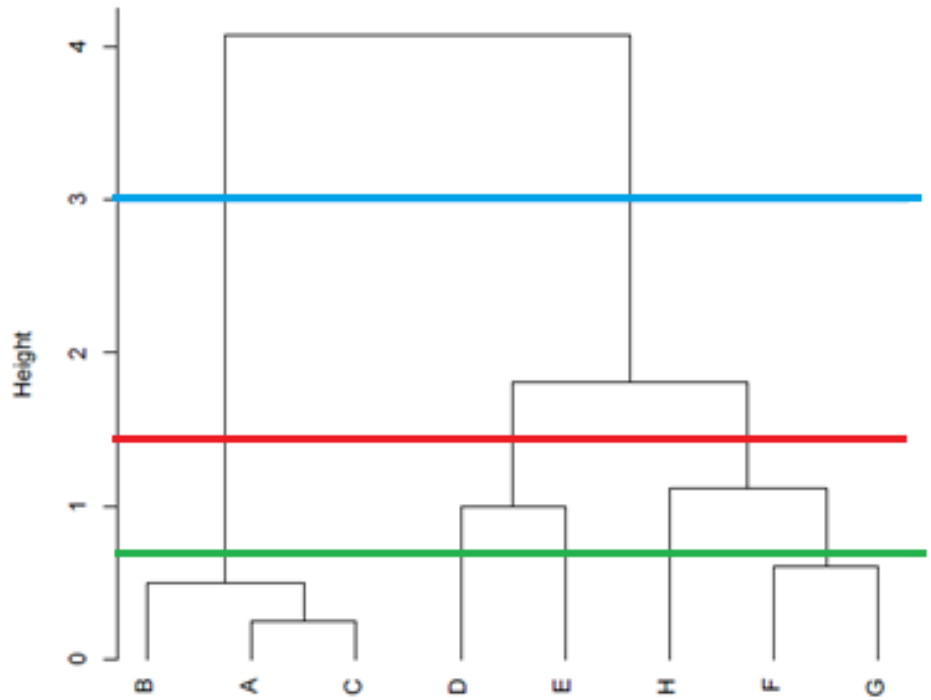
- $D(AC, B) = \min(d(A, B), d(C, B)) = \min(0.5, 0.56) = 0.5$;
- $D(AC, D) = \min(d(A, D), d(C, D)) = \min(5, 4.8) = 4.8$;
- ...

After the second step :

- $D(ABC, D) = \min(D(AC, D), d(B, D)) = \min(4.8, 4.72) = 4.8$;
- $D(ABC, E) = \min(D(AC, E), d(B, E)) = \min(5.57, 5.55) = 5.57$;
- ...

Here, index h (height of a class in the dendrogram) = minimum link between the two subclasses.

HCA : where to cut the dendogram ?



Here $\mathscr{P} = \{\{A, B, C\}, \{D, E, F, G, H\}\}$,
$\mathscr{P} = \{\{A, B, C\}, \{D, E\}, \{F, G, H\}\}$, and
$\mathscr{P} = \{\{A, B, C\}, \{D\}, \{E\}, \{F, G\}, \{H\}\}$.

HCA : another example

Example of temperature data

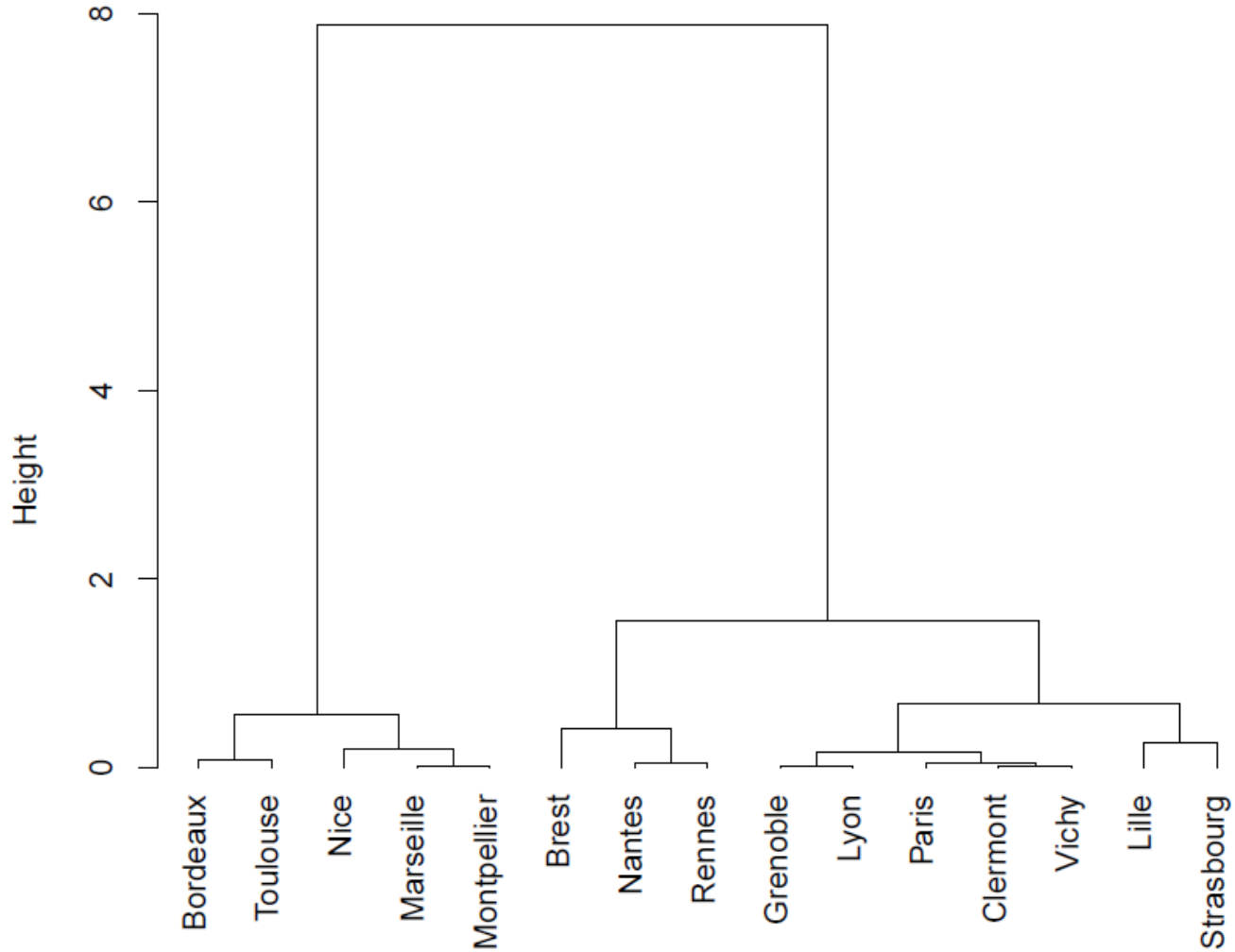
- 15 individuals : French cities,
- 12 variables : average monthly temperatures (over 30 years).

	Janv	Févr	Mars	Avri	Mai	Juin	juil	Août	Sept	Octo	Nove	Déce
Bordeaux	5.6	6.6	10.3	12.8	16	19	21	21	19	13.8	9.1	6.2
Brest	6.1	5.8	7.8	9.2	12	14	16	16	15	12.0	9.0	7.0
Clermont	2.6	3.7	7.5	10.3	14	17	19	19	16	11.2	6.6	3.6
Grenoble	1.5	3.2	7.7	10.6	14	18	20	20	17	11.4	6.5	2.3
Lille	2.4	2.9	6.0	8.9	12	15	17	17	15	10.4	6.1	3.5
Lyon	2.1	3.3	7.7	10.9	15	18	21	20	17	11.4	6.7	3.1
Marseille	5.5	6.6	10.0	13.0	17	21	23	23	20	15.0	10.2	6.9
Montpellier	5.6	6.7	9.9	12.8	16	20	23	22	19	14.6	10.0	6.5
Nantes	5.0	5.3	8.4	10.8	14	17	19	19	16	12.2	8.2	5.5
Nice	7.5	8.5	10.8	13.3	17	20	23	22	20	16.0	11.5	8.2
Paris	3.4	4.1	7.6	10.7	14	18	19	19	16	11.4	7.1	4.3
Rennes	4.8	5.3	7.9	10.1	13	16	18	18	16	11.6	7.8	5.4
Strasbourg	0.4	1.5	5.6	9.8	14	17	19	18	15	9.5	4.9	1.3
Toulouse	4.7	5.6	9.2	11.6	15	19	21	21	18	13.3	8.6	5.5
Vichy	2.4	3.4	7.1	9.9	14	17	19	19	16	11.0	6.6	3.4

Which cities have similar meteorological profiles ?

HCA : another example

Algorithme de Ward appliqué aux températures standardisées



HCA : another example

Height of a class in Ward's dendogram

The aggregation height of two classes C_k and $C_{k'}$ is :

$$\underbrace{\frac{|C_k||C_{k'}|}{|C_k|+|C_{k'}|}d(\mu_k,\mu_{k'})^2}_{\textit{Ward's measure}} = \underbrace{I(C_k \cup C_{k'}) - I(C_k) - I(C_{k'})}_{\textit{Loss of explained inertia}}$$

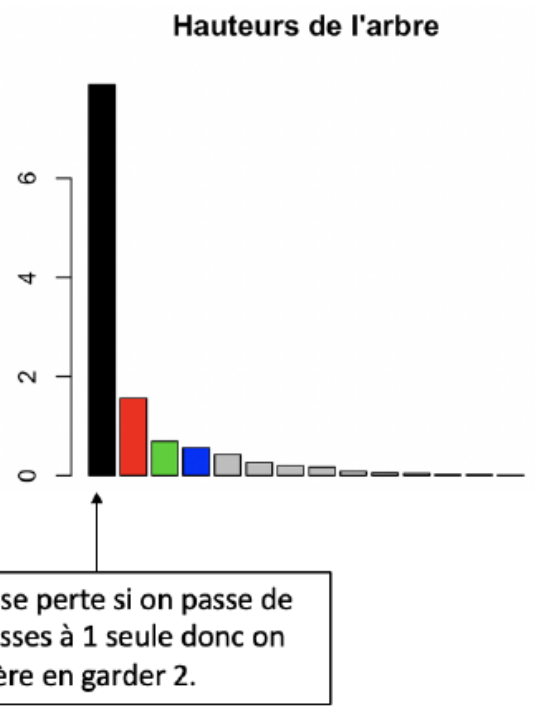
$$= \underbrace{|C_k|d(\mu_k,\mu)^2 + |C_{k'}|d(\mu_{k'},\mu)^2}_{\textit{Inter inertia}}.$$

HCA : another example

For the temperature data

Pertes d'inertie lors du passage de

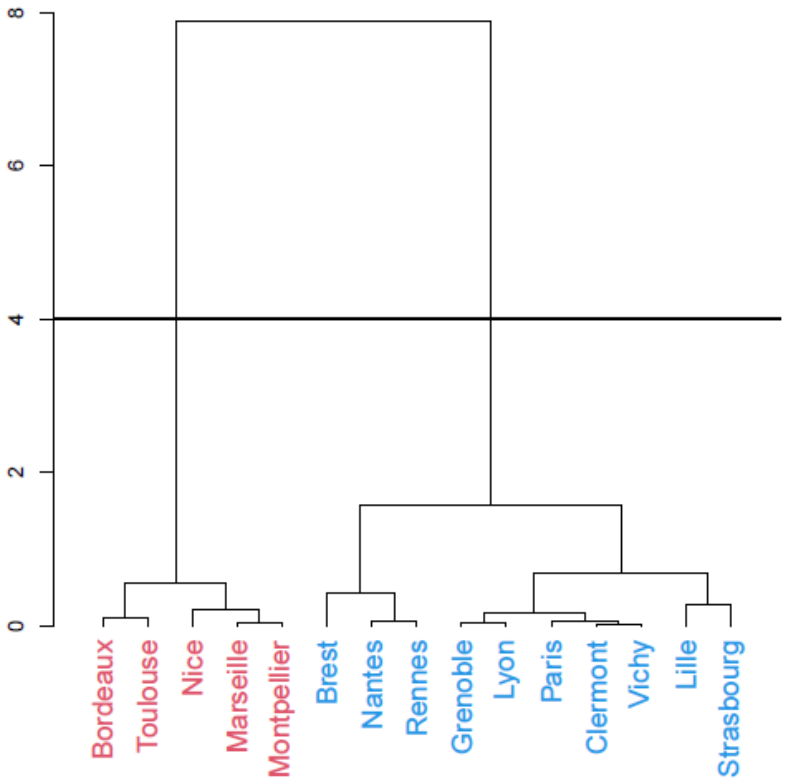
15 classes à 14 classes : 0.01
14 classes à 13 classes : 0.02
13 classes à 12 classes : 0.03
12 classes à 11 classes : 0.05
11 classes à 10 classes : 0.06
10 classes à 9 classes : 0.09
9 classes à 8 classes : 0.17
8 classes à 7 classes : 0.19
7 classes à 6 classes : 0.26
6 classes à 5 classes : 0.42
5 classes à 4 classes : 0.56
4 classes à 3 classes : 0.69
3 classes à 2 classes : 1.56
2 classes à 1 classes : 7.88



Sum of the intertia losses $= 12$ (total inertia).

HCA : another example

Cut the dendogram to get a partition



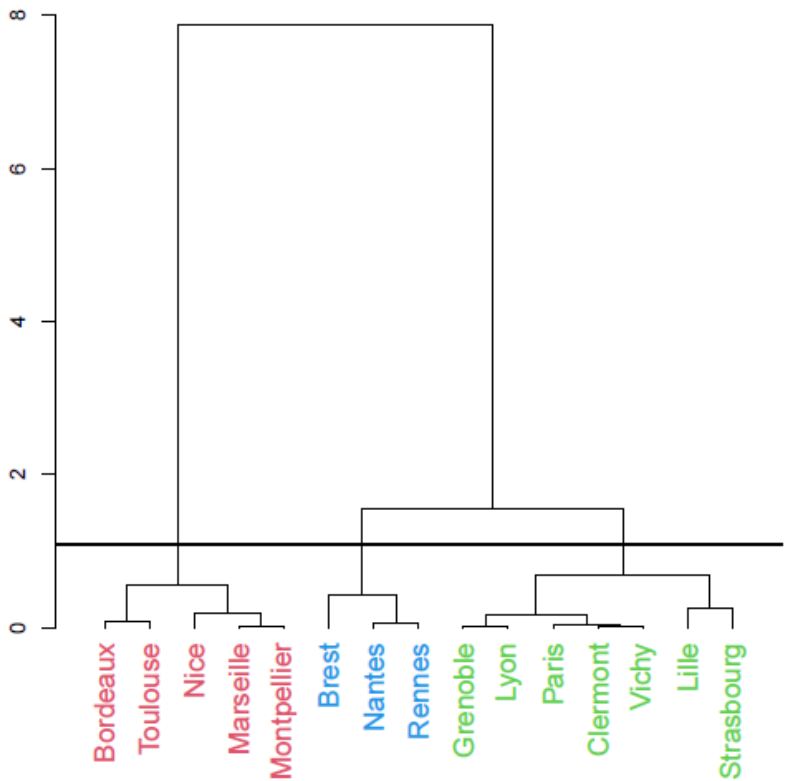
Clustering in 2 classes

$$\frac{\textit{Inter inertia}}{\textit{Total inertia}} = \frac{7.88}{12} = 66\%$$

☞ 66% of explained inertia with the partition in 2 classes.

HCA : another example

Separate the cold cities into two groups.



Clustering in 3 classes

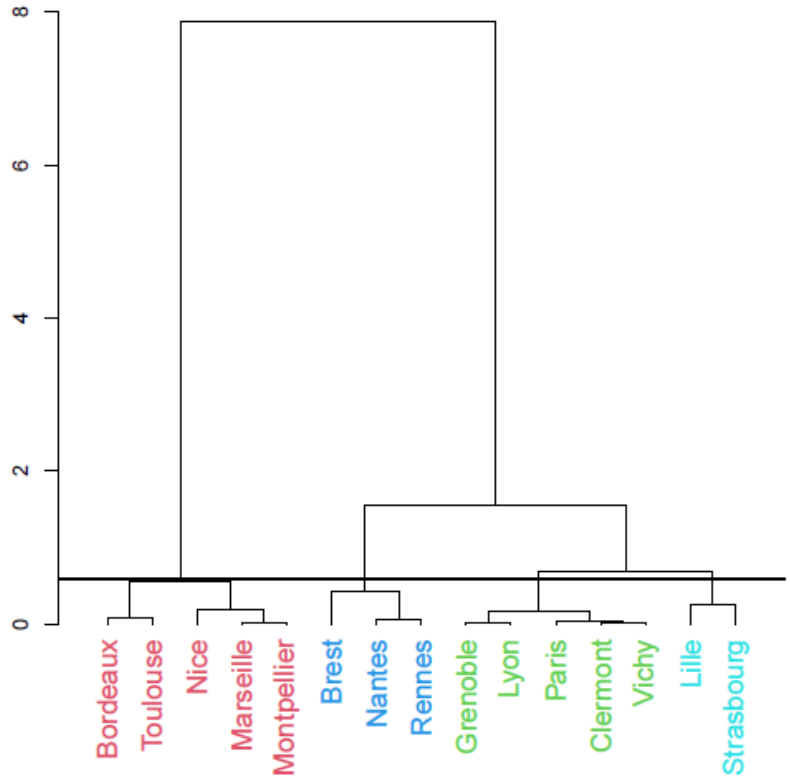
$$\frac{\textit{Inertia loss}}{\textit{Total inertia}} = \frac{1.56}{12} = 13\%$$

Gain of 13% of inertia considering 3 classes instead of 2 (going from 2 to 3).

☞ $66\% + 13\% = 79\ \%$ of explained inertia with the partition in 3 classes.

HCA : another example

Separate the eastern cold cities into two groups.



Clustering in 4 classes

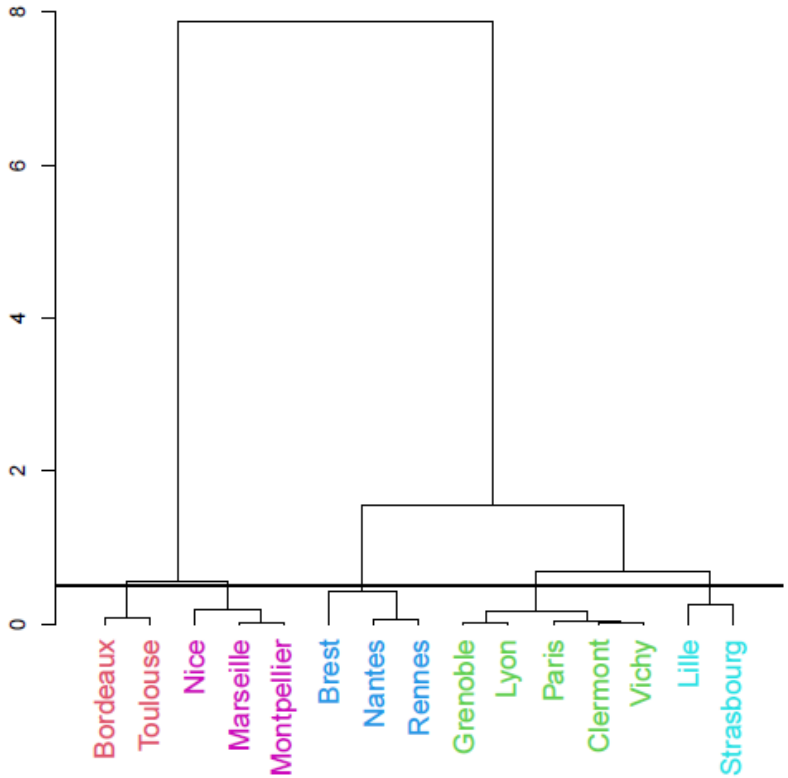
$$\frac{\textit{Inertia loss}}{\textit{Total inertia}} = \frac{0.69}{12} = 5.75\%$$

Gain of 5.75% of inertia considering 4 classes instead of 3 (going from 3 to 4).

☞ $79\% + 5.7\% = 84.7\ \%$ of explained inertia with the partition in 4 classes.

HCA : another example

Separate the hot cities into two groups.



Clustering in 5 classes

$$\frac{\textit{Inertia loss}}{\textit{Total inertia}} = \frac{0.56}{12} = 4.7\%$$

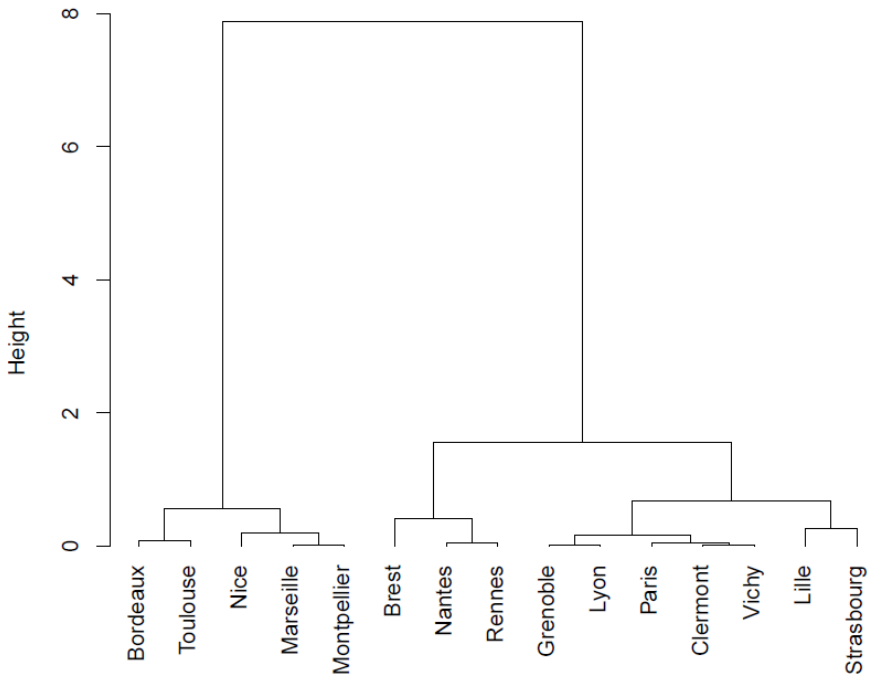
Gain of 4.7% of inertia considering 5 instead of 4 classes (going from 4 to 5).

☞ Gain close to the gain of the passage from 3 to 4 classes.

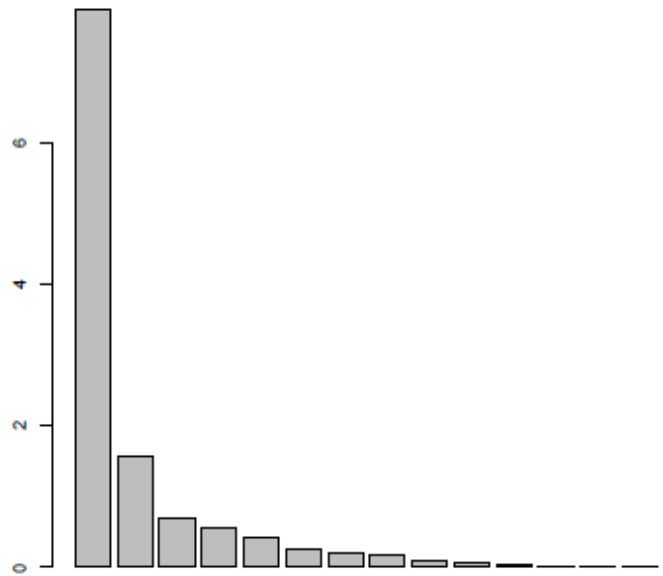
HCA : another example

Determination of the number of clases

From the dendogram



From the heights



From the interpretability of classes.

HCA : another example

Proprerties of Ward's algorithm

- The partition constructed at each step maximizes the inter-inertia among the partitions resulting from the aggregation of two classes from the previous partition.
- The sum of the heights of the Ward dendrogram is the total inertia.
- The sum of the $K-1$ largest heights is the inter-inertia of the partition into K classes of the tree.
- The complexity of the algorithm : quadratic with respect to the number of individuals.

$\Rightarrow$ Problem for datasets with a very large number of individuals.

HCA : where to cut the dendogram ?

By defining a cut level, we build a partition. The cut level determines the number of classes and the classes are then unique.

The cut must be made

- after the aggregations corresponding to low values of the index ;
- before the aggregations corresponding to high values of the index.

In most cases, several thresholds and therefore several possible choices of partitions.

HCA : where to cut the dendogram ?

Rule of thumb - Selection of a cut when there is a significant jump in the index by visual inspection of the tree. This jump reflects the sudden passage from a certain homogeneity of classes to much less homogeneous classes.

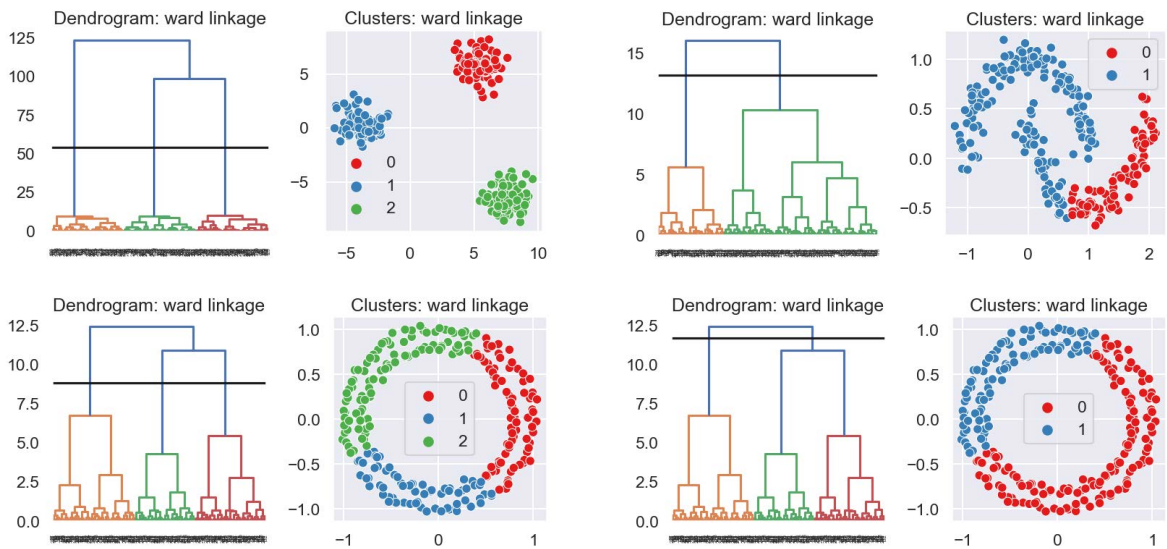


Image : Chevallier (2023).

HCA : where to cut the dendrogram ?

The cut of the dendrogram can be defined by determining a priori the number of classes into which we want to divide the data set.

It is also possible to use the criteria seen previously :

- R-square : elbow on the curve $K \mapsto RSQ(K)$,
- Semi-partial R-square : greater reduction of the SPRSQ,
- Silhouette criterion,
- ...

HCA : strengths and weaknesses

Advantages

- Simple considerations of distances between individuals and dissimilarities between clusters.
- No assumption on the number of classes.
- Can correspond to significant taxonomies.

Disadvantages

- Choice of the dendogram cutoff.
- The partition obtained at a step depends on that of the previous step.
- Once a decision is made to combine classes, it cannot be undone.
- Too slow for large datasets.

Plan

Methodological Supplements and Class Interpretation

The dataset shows the amount of protein consumed in 9 food types in 25 (former) European countries : 25 individuals (the first 10 below) and 9 quantitative variables.

	Red.Meat	White.Meat	Eggs	Milk	Fish	Cereals	Starchy.Foods	Nuts	Fruite.veg.
Alban	10.1	1.4	0.5	8.9	0.2	42	0.6	5.5	1.7
Aust	8.9	14.0	4.3	19.9	2.1	28	3.6	1.3	4.3
Belg	13.5	9.3	4.1	17.5	4.5	27	5.7	2.1	4.0
Bulg	7.8	6.0	1.6	8.3	1.2	57	1.1	3.7	4.2
Czech	9.7	11.4	2.8	12.5	2.0	34	5.0	1.1	4.0
Den	10.6	10.8	3.7	25.0	9.9	22	4.8	0.7	2.4
E_Ger	8.4	11.6	3.7	11.1	5.4	25	6.5	0.8	3.6
Finl	9.5	4.9	2.7	33.7	5.8	26	5.1	1.0	1.4
Fr	18.0	9.9	3.3	19.5	5.7	28	4.8	2.4	6.5
Greece	10.2	3.0	2.8	17.6	5.9	42	2.2	7.8	6.5

We apply k-means to these data to illustrate two methodological aspects :

- Why is it sometimes necessary to standardize data ?
- How to interpret classes using PCA ?

Why is it sometimes necessary to standardize data ?

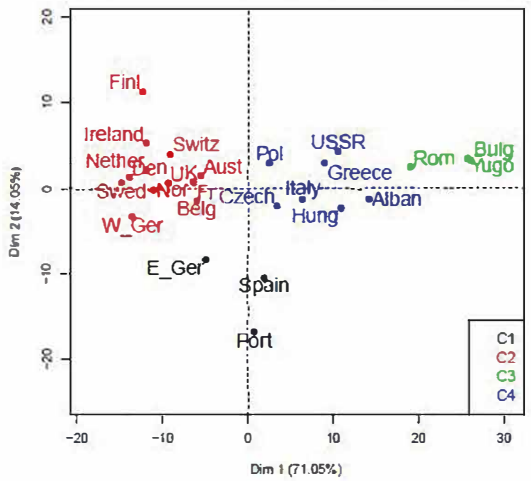
Raw data : partition of k-means into 4 classes after N=5 initializations.

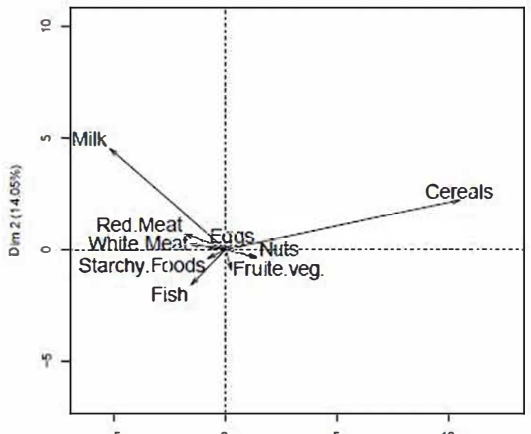
	P4
Alban	4
Aust	2
Belg	2
Bulg	3
Czech	4
Den	2
E_Ger	1
Finl	2
Fr	2
Greece	4
Hung	4
Ireland	2
Italy	4
Nether	2
Nor	2
Pol	4
Port	1
Rom	3
Spain	1
Swed	2
Switz	2
UK	2
USSR	4
W_Ger	2
Yugo	3

	écart-type
Red.Meat	3.4
White.Meat	3.7
Eggs	1.1
Milk	7.1
Fish	3.4
Cereals	11.0
Starchy.Foods	1.6
Nuts	2.0
Fruite.veg.	1.8

Interpretation via PCA

not norméd.



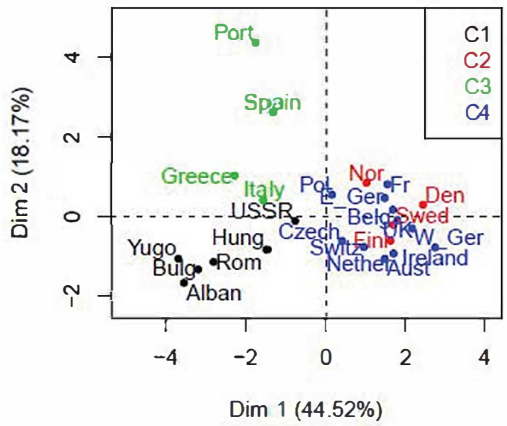


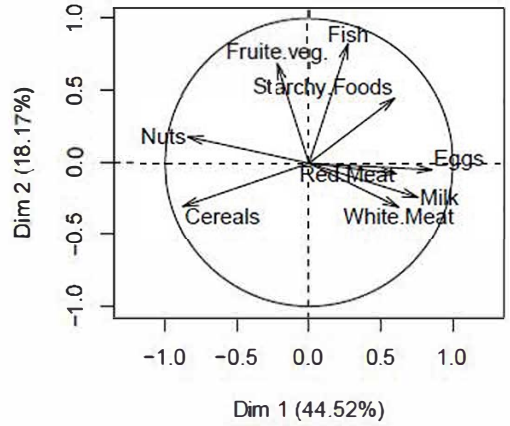
Why is it sometimes necessary to standardize data ?

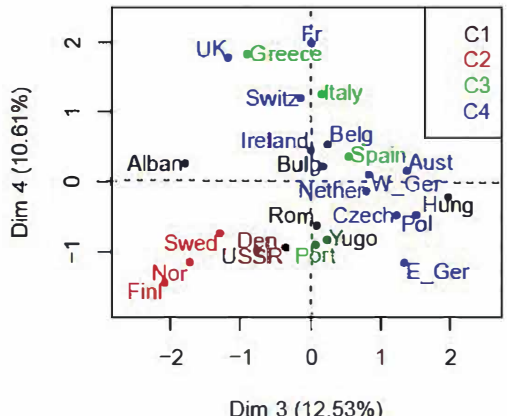
Standardized data : partitioning the K-means into 4 classes after $N = 5$ initializations.

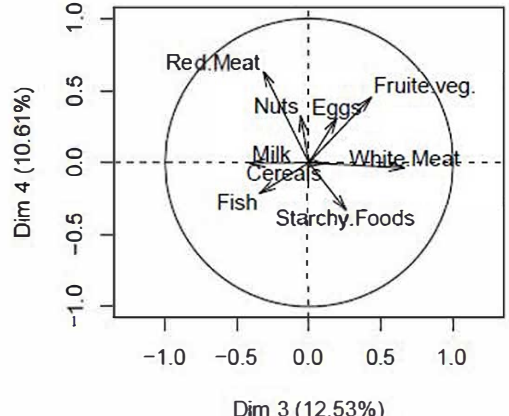
	écart-type
Red.Meat	1
White.Meat	1
Eggs	1
Milk	1
Fish	1
Cereals	1
Starchy.Foods	1
Nuts	1
Fruite.veg.	1

Interpretation via PCA normé.









Consolidation of a partition obtained by HCA

The partition obtained by HCA is not optimal and can be improved, consolidated, by k-means.

Consolidation algorithm :

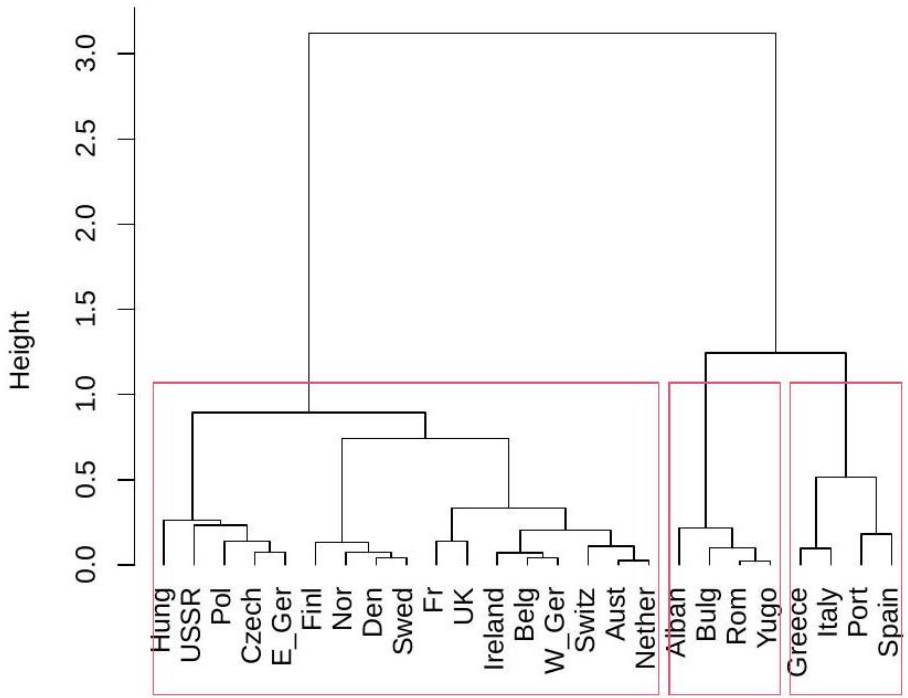
- the partition obtained by HCA is used as initialization of the partitioning algorithm,
- a few k-means steps are iterated.

$\Rightarrow$ Improvement of the partition (often not decisive).

Advantage : Consolidation of the partition.

Disadvantage : Loss of hierarchy information.

Example of standardized protein data.



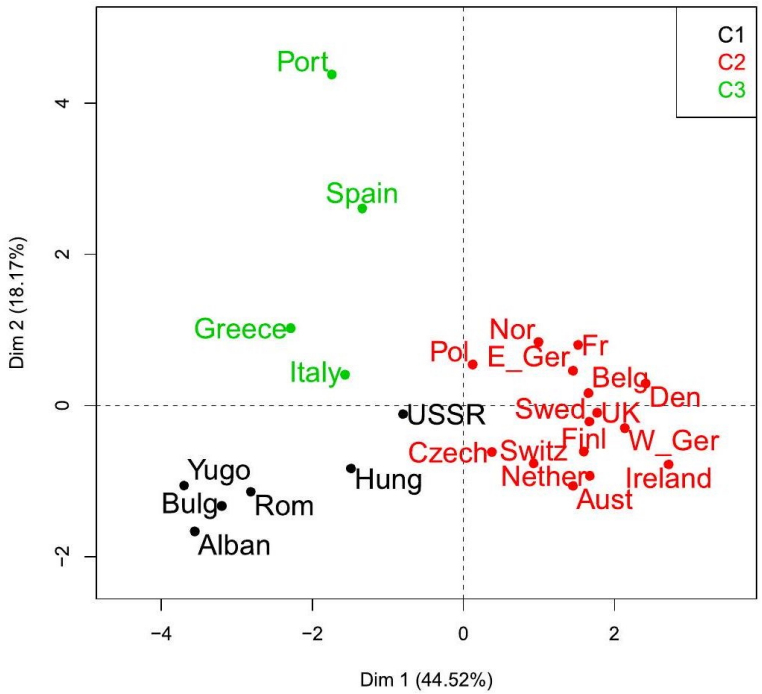
The dendrogram suggests splitting the tree into 3 or 5 classes.

Let's look at the partition into 3 classes :

$$\frac{\textit{Inter inertia}}{\textit{Total inertia}} = 0.485.$$

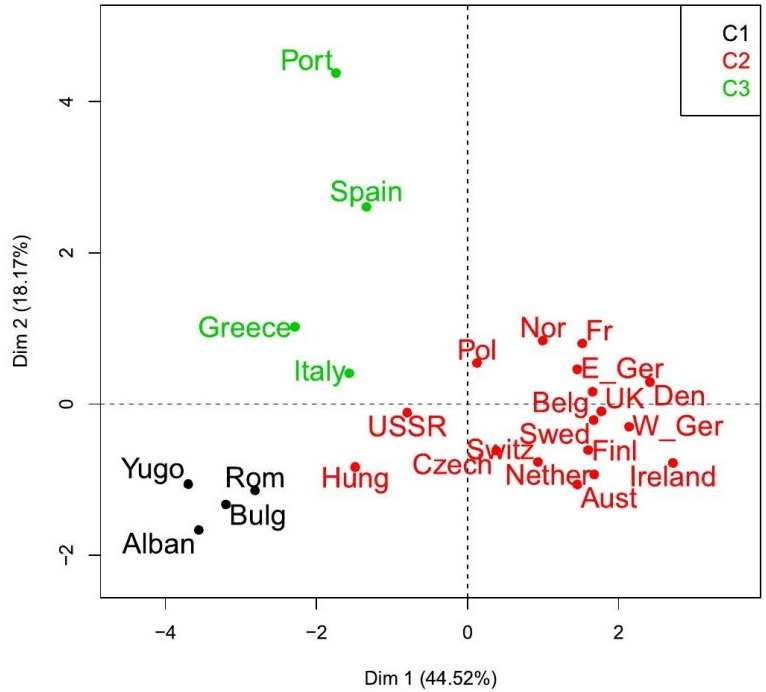
The partition into Ward's 3 classes explain 48.5% of the inertia.

Before consolidation



48.5% of the inertia explained by

After consolidation



50.9% of the inertia explained by the

the ⇒ partition slightly improved (2 individuals changed class).

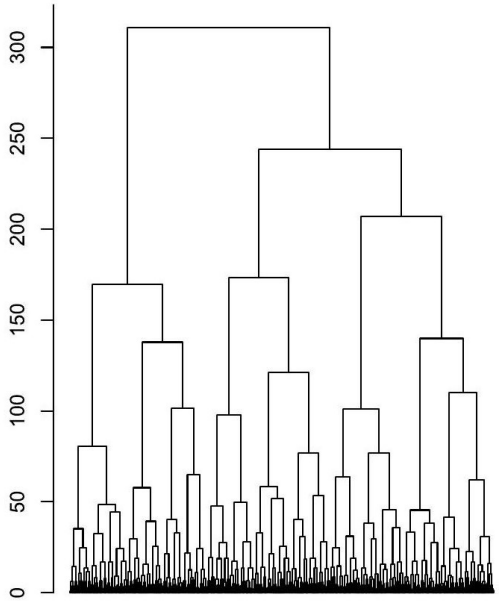
HCA with many individuals

If there are many individuals, the HCA algorithm becomes too long.

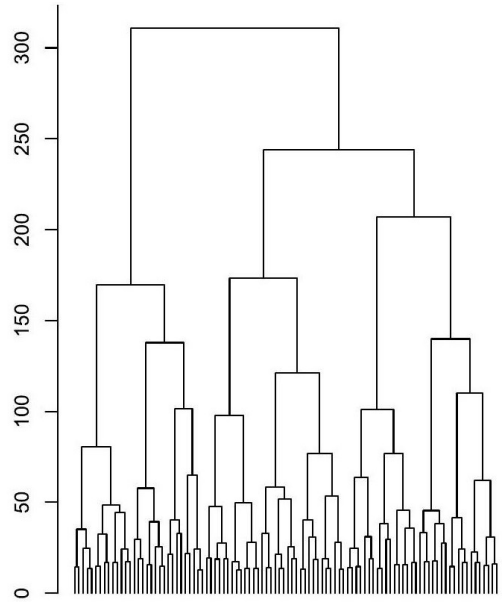
- Partition (by k-means) into about a hundred classes.
- Construct the HCA from the classes (use the class size in the calculation).

$\Rightarrow$ Obtain the "top" of the HCA tree.

Tree on 1000 raw data



Tree from 100 classes



HCA or k-means with many variables

If there are many variables, reduce the dimension.

- Perform a PCA and retain the first q principal components.
  ☞ If we retain all the principal components of the normalized (or unnormalized) PCA, we find the same classification as with the standardized (or raw) data.
- Perform a clustering of variables into q classes and retain the q synthetic variables (first principal components of the classes).

$\Rightarrow$ Difficulty in choosing q.

HCA or k-means on categorial or mixed variables variables

- Refer to quantitative variables :
  - make a MCA (or mixed PCA) and retain all major components (or the first q) ;
  - make a variable clustering (which manages categorial and mixed data) and keep the q synthetic variables of the classes.

  $\Rightarrow$ Difficulty choosing q.
- Use measures appropriate for categorial or mixed data :
  similarity index, Jaccard index... then apply a HCA algorithm to this similarity matrix (dissimilarity, distance).
  $\Rightarrow$ What does Ward mean if the distance is not Euclidean ?

Interpreting Individual Classes

We can interpret the classes of a partition based on

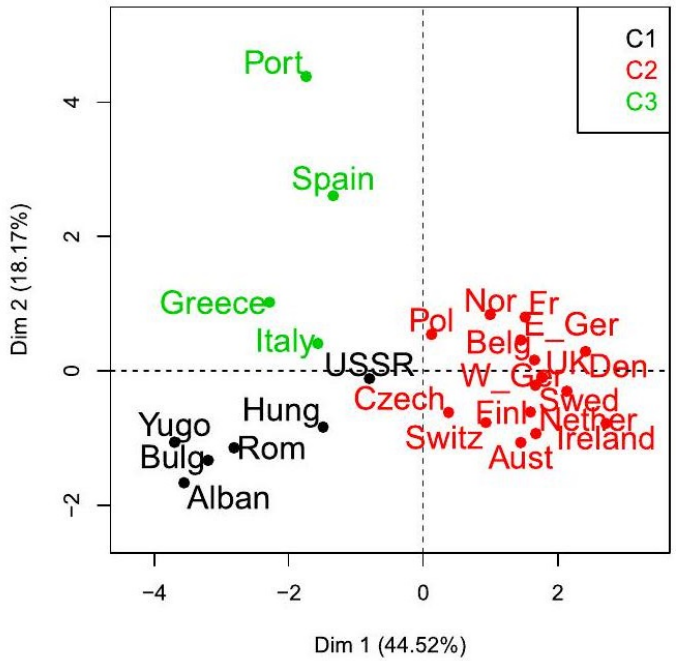
- active variables : variables used in the clustering process,
- illustrative variables : used solely to describe the classes.

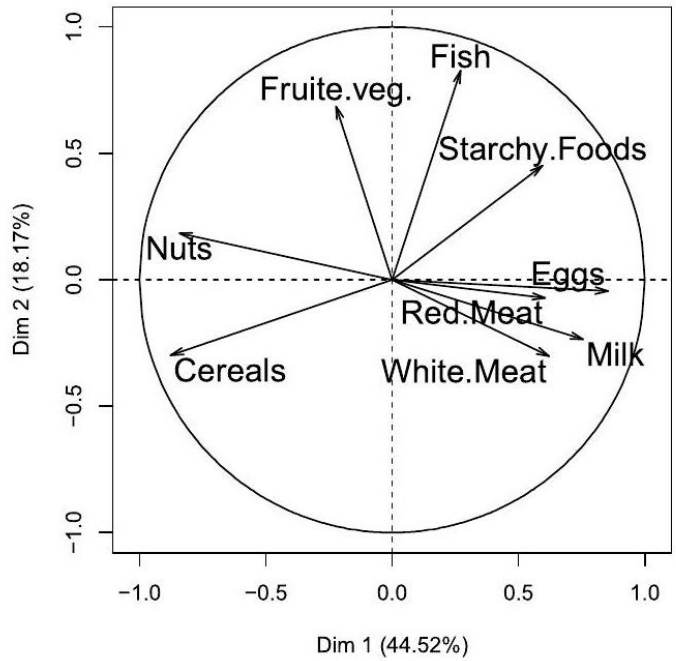
These variables can be quantitative or
pcategorial.
In practice, we will often characterize classes (or groups of individuals) by :

- the modalities of categorial variables : is one modality more frequent in the class, does the class contain all the individuals possessing this modality, ... ?
- the quantitative variables : is the average in the class different from the average across all individuals... ?

Interpreting Individual Classes

Example of the partition into 3 classes after consolidation of protein data.





Questions

- Which variables best characterize the partition ?
- How can we characterize the cities in a particular class ?

Which variables best characterize the partition ?

For each quantitative variable :

- calculate the correlation ratio η^2 between the partition (categorial variable with k modalities) and the quantitative variables.
- perform the Fisher test of the effect of the partition on the quantitative variable (analysis of variance model),
- sort the variables by increasing critical probability (p-value).

	η^2	p-value
Nuts	0.79	3.0e – 08
Cereals	0.75	2.1e – 07
Eggs	0.58	8.1e – 05
Fruite.veg.	0.54	1.7e – 04
Milk	0.53	2.3e – 04
White.Meat	0.43	1.9e – 03
Fish	0.40	4.0e – 03
Red.Meat	0.33	1.3e – 02

Which variables best characterize the partition ?

For each categorial variable : the same with a test of χ^2 of independence between the partition and the categorial variable.

	zone		zone
Alban	east	Nether	west
Aust	west	Nor	north
Belg	west	Pol	east
Bulg	east	Port	south
Czech	east	Rom	east
Den	north	Spain	south
E_Ger	east	Swed	north
Finl	north	Switz	west
Fr	west	UK	west
Greece	south	USSR	east
Hung	east	W_Ger	west
Ireland	west	Yugo	east
Italy	south		

	p-value	df
zone	1e-06	6

Which quantitative variables characterize a class ?

Output of the `catdesc` function from the R package FactoMineR for C1.

```
#Countries in C1:
pays <- rownames(protein)
pays[which(P5=="C1")]

## [1] "Alban" "Bulg"  "Rom"   "Yugo"

res <- catdes(data.frame(P5,Z),num.var=1)
res$quanti$C1

##               v.test Mean in category Overall mean sd in category Overall sd p.value
## Cereals          3.8               1.8      2.4e-16           0.54          1 0.00017
## Nuts             2.2               1.0     -1.6e-17           0.41          1 0.02972
## Fish            -2.3              -1.1      6.3e-17           0.12          1 0.02342
## Milk            -2.4              -1.1     -2.1e-16           0.15          1 0.01862
## Starchy.Foods   -3.1              -1.5      1.4e-16           0.70          1 0.00190
## Eggs            -3.4              -1.6      3.1e-17           0.39          1 0.00070
```

Which quantitative variables characterize a class ?

Which quantitative variable X best characterizes class C_k ?

The test value of a quantitative variable X in class C_k measures the difference between the mean of X in C_k and the mean of X across all data :

$$test\ value = \frac{\overline{X}_k - \overline{X}}{\sqrt{\frac{s^2}{n_k}\frac{n-n_k}{n-1}}}$$

where $s^2 = \frac{n_k}{n}\left(1 - \frac{n_k}{n}\right)$ is the variance of X in C_k.

☞ If the test value of a variable in a class is large (in absolute value), this variable characterizes the class.

Which quantitative variables characterize a class ?

Which quantitative variable X best characterizes class C_k ?

Hypothesis tested $\mathscr{H}_0$: the n_k values of X are selected randomly from n. Under $\mathscr{H}_0$, the mean in the class is the same as in the population and

$$test\text{-}value = \frac{\overline{X}_k - \overline{X}}{\sqrt{\frac{s^2}{n_k}\frac{n-n_k}{n-1}}} \sim \mathscr{N}(0,1)$$

for n tending to infinity.

☞ If the p-value of this test is small (less than 0.05, for example), this variable characterizes the class.

```
#Countries in C1:
pays <- rownames(protein)
pays[which(P5=="C1")]

## [1] "Alban" "Bulg"  "Rom"   "Yugo"

res <- catdes(data.frame(P5,Z),num.var=1)
res$quanti$C1

##                v.test Mean in category Overall mean sd in category Overall sd p.value
## Cereals           3.8               1.8      2.4e-16           0.54          1 0.00017
## Nuts              2.2               1.0     -1.6e-17           0.41          1 0.02972
## Fish             -2.3              -1.1      6.3e-17           0.12          1 0.02342
## Milk             -2.4              -1.1     -2.1e-16           0.15          1 0.01862
## Starchy.Foods    -3.1              -1.5      1.4e-16           0.70          1 0.00190
## Eggs             -3.4              -1.6      3.1e-17           0.39          1 0.00070
```

The first column gives the `v.test` (test-values) quantitative variables in the class.

The last column gives a `p.value` (*p*-value) and by default only variables with a `p.value` greater than 0.05 are displayed.

```
res <- catdes(data.frame(P5,Z),num.var=1)
res$quanti[2:5]

## $C2
##             v.test Mean in category Overall mean sd in category Overall sd p.value
## Red.Meat      3.5             1.03      1.7e-16           0.94          1 0.00052
## Eggs          3.2             0.96      3.1e-17           0.52          1 0.00125
## White.Meat    2.5             0.76      8.6e-18           0.70          1 0.01091
## Cereals      -2.4            -0.70      2.4e-16           0.28          1 0.01832
##
## $C3
##                v.test Mean in category Overall mean sd in category Overall sd p.value
## Starchy.Foods      2              0.8      1.4e-16           0.59          1   0.049
##
## $C4
##              v.test Mean in category Overall mean sd in category Overall sd p.value
## Milk            2.9             1.37     -2.1e-16           0.59          1  0.0033
## Fish            2.5             1.18      6.3e-17           0.51          1  0.0115
## Nuts           -2.1            -0.98     -1.6e-17           0.18          1  0.0371
## Fruite.veg.    -2.4            -1.14     -4.5e-17           0.28          1  0.0150
##
## $C5
##              v.test Mean in category Overall mean sd in category Overall sd p.value
## Fruite.veg.    3.6              1.7     -4.5e-17           0.31          1 0.00038
## Nuts           2.9              1.3     -1.6e-17           0.70          1 0.00423
## Fish           2.1              1.0      6.3e-17           1.20          1 0.03214
## White.Meat    -2.4             -1.1      8.6e-18           0.22          1 0.01554
```

Which modalities of categorical variables characterize a class ?

The catdesc function also describes the categorical variates.

```
zone <- c("east","west","west","east","east","north","east","north","west","south",
          "east","west","south","west","north","east","south","east","south","north",
          "west","west","east","west","east")
res <- catdes(data.frame(P5,zone),num.var=1)
res$category$C1

##           Cla/Mod Mod/Cla Global p.value v.test
## zone=east      44     100     36    0.01    2.6
```

Cla/Mod = proportion of modality m in class k = $\frac{n_{mk}}{n_m}$
Mod/Cla = proportion of class k in modality m = $\frac{n_{mk}}{n_k}$
Global = proportion of modality m in the data = $\frac{n_m}{n}$

Which modalities of categorical variables characterize a class ?

Which modality m of X best characterizes class C_k ?

	C1	C2	C3	C4	C5	Total
east	$n_{mk} = 4$	0	5	0	0	$n_m = 9$
north	0	0	0	4	0	4
south	0	0	0	0	4	4
west	0	8	0	0	0	8
Total	$n_k = 4$	8	5	4	4	$n = 25$

Thus, for class 1,

$Cla/Mod = \frac{4}{9} \approx 0.44 = 44\%$, $Mod/Cla = \frac{4}{4} = 1 = 100\%$,
$Global = \frac{9}{25} = 0.36 = 36\%$

$\Rightarrow$ the "east" modality characterizes the class (over-represented).

```
res$category

## $C1
##           Cla/Mod Mod/Cla Global p.value v.test
## zone=east      44     100     36    0.01    2.6
##
## $C2
##           Cla/Mod Mod/Cla Global p.value v.test
## zone=west     100     100     32 9.2e-07    4.9
## zone=east       0       0     36 1.2e-02   -2.5
##
## $C3
##           Cla/Mod Mod/Cla Global p.value v.test
## zone=east      56     100     36  0.0024      3
##
## $C4
##            Cla/Mod Mod/Cla Global p.value v.test
## zone=north     100     100     16 7.9e-05    3.9
##
## $C5
##            Cla/Mod Mod/Cla Global p.value v.test
## zone=south     100     100     16 7.9e-05    3.9
```

Which modalities of categorical variables characterize a class ?

Which modality m of X best characterizes the class C_k ?

Under $\mathscr{H}_0$: $\frac{n_{mk}}{n_k} = \frac{n_m}{n}$ i.e. the frequency in the class is the same as in the population

$$\text{test-value} = \frac{\frac{n_{mk}}{n_k} - \frac{n_k}{n}}{\sqrt{\frac{n-n_k}{n-1}\frac{s^2}{n_k}}} \sim \mathscr{N}(0,1)$$

with $s^2 = \frac{n_m}{n}\left(1 - \frac{n_m}{n}\right)$, for n tending towards infinity.

Cảm ơn sự chú ý của bạn !

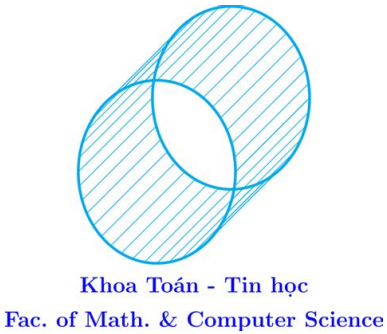
Câu hỏi ?

Major sources of the whole lecture

- Marie Chavent's lectures
  `https://marie-chavent.perso.math.cnrs.fr/`
- Juliette Chevallier's lectures
  `https://juliette-chevallier.pages.math.cnrs.fr/`


VIASM
VIETNAM INSTITUTE FOR
ADVANCED STUDY IN MATHEMATICS


TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHTN
ĐHQG-HCM
HCMUS
Viet Nam National University
Ho Chi Minh City
University of Science


Khoa Toán - Tin học
Fac. of Math. & Computer Science


IRL
FVMA
International Research Laboratory
FRANCE - VIETNAM
in MATHEMATICS & ITS APPLICATION


INSTITUT
de MATHEMATIQUES
de TOULOUSE


UNIVERSITÉ TOULOUSE
Jean Jaurès


INSA
TOULOUSE


université
PARIS-SACLAY


ensIIE


cimi
TOULOUSE


Mint
Mathematics and Interactions in Toulouse


ANITI
Université
Fédérale
Toulouse
Midi-Pyrénées