Hierarchical clustering methods

Mohamed Nadif

Université Paris Descartes, France

Outline

Plan for the week

Clustering and Visualization

- Course 1: Hierarchical clustering methods
- Course 2: Nonhierarchical clustering methods
- Course 3: Finite Mixture Models and Clustering
- Course 4: Block Clustering models and algorithms
- Course 5: Intoduction to the Visualization

Principals Points

- Advantages and disadvantages of methods
- Illustrations by examples
- Importance of Softwares
- Use of SAS, **R** and discussion about other tools

Clustering

- Aim: It seeks to obtained a reduced representation of the initial data
- Organization of data into homogeneous subsets "clusters" or "classes"
- Terminoloy can depend on the field:
  - Taxonomy science of clustering of human being
  - Nosology science of clustering of diseases in medecine
  - Unsupervised learning or unsupervised classification in pattern recognition and machine learning
- Not confuse with the classification
- History: first clustering of the animals and the vegetables by Linné (18th century)
- Naming objects is a form of clustering

Structure of clustering

- It can take different forms: partitions, sequence of encased partitions or hierarchical, overlapping clusters, clusters with high density, fuzzy clusters.
- In this chapter we focus on the hierarchical methods

Characteristics of these methods

- Simple and is a tool of data visualization but ...

Data matrix: x

- **x** represents n objects (persons, countries, genes) with p variables (weight, income, religion, sex)

$$\mathbf{x} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$

Distance or dissimilarity matrix: d

- Each value $d(i,j)$ is the measured difference or dissimilarity between objects i and j with $d(i,i) = 0$. In general $d(i,j)$ is a nonnegative number and this matrix is symmetric.

$$\mathbf{d} = \begin{pmatrix} 0 & \cdots & & \\ d(2,1) & 0 & \cdots & \\ \vdots & \ddots & \vdots & \\ d(n,1) & d(n,2) & \cdots & 0 \end{pmatrix}$$

Types of variables

- Continuous variable: weight, height etc.
  - Standardization is often necessary (centered and reduced, etc.)
  - Other transformations can be useful (log, exp, 1/x etc.)
- Binary variable
  - Nominal categorical with two categories, the binary variable is called *symmetric*
  - Can be considered as continuous variable
  - If the outcomes of a binary variable are not equally important, the variable is *asymmetric*. Importance of 1, examples include presence-absence data in ecology. Measure of dissimilarity adapted such as Jaccard'is index
- Categorical Variable
  - Nominal: generally we use the *complete disjunctive table*. Ex. let be a variable with 3 categories (modalities) 1,2 and 3. These categories are coded respectively by the binary vectors : $(1,0,0)$, $(0,1,0)$ and $(0,0,1)$
  - Ordinal: generally we use the *disjunctive additive table*. Ex. let be a variable with 3 categories 1,2 and 3. These categories are coded respectively by the binary vectors : $(1,0,0)$, $(1,1,0)$ and $(1,1,1)$. Sometimes we can use the following transformation $\frac{r_{ij}-1}{k_j-1} \in [0,1]$ where $r_{ij} = 1, \ldots k_j$ is the rank of a value of a variable j and k_j is the number of distinct values, and consider j as a continuous variable

Proximity

- The evaluation of proximity between objects depend on the nature of variables
- The evaluation between variables is *more complex*

Distance d: $A \times A \rightarrow \mathbb{R}$

- $\forall x, y \in A,\ d(x, y) = 0 \Leftrightarrow x = y$
- $\forall x, y \in A,\ d(x, y) = d(y, x)$
- $\forall x, y, z \in A, d(x, z) \leq d(x, y) + d(y, z)$

Ultrametric

- $\forall x, y \in A,\ d(x, y) = 0 \Leftrightarrow x = y$
- $\forall x, y \in A,\ d(x, y) = d(y, x)$
- $\forall x, y, z \in A,\ d(x, z) \leq max(d(x, y), d(y, z))$

Dot or Scalar product: $E \times E \rightarrow \mathbb{R}$ (E: vector space)

- $\forall x \in E, < x, x >= 0 \Rightarrow x = 0$
- $\forall x, y \in E, < x, y >=< y, x >$
- $\forall x \in E, < x, x > \geq 0$

Scalar product: Matrix representation expresssion in $\mathbb{R}^p$

- $< x, y >_M = x^T My$ where the matrix $(p \times p)$ M is
  - symmetric $M^T = M$
  - definite $\forall x \in \mathbb{R}^p,\ x^T Mx = 0 \Rightarrow x = 0$
  - positive $\forall x, y \in E,\ x^T Mx \geq 0$

Norm: E (vector space) $||.|| : E \longrightarrow \mathbb{R}^+$

- $\forall \boldsymbol{x} \in E, \lambda \in \mathbb{R}, ||\lambda x|| = |\lambda||x||$
- $\forall \boldsymbol{x} \in E, ||\boldsymbol{x}|| = 0 \Rightarrow x = 0$
- $\forall \boldsymbol{x}, \boldsymbol{y} \in E, ||\boldsymbol{x} + \boldsymbol{y}|| \leq ||\boldsymbol{x}|| + ||\boldsymbol{y}||$

Euclidean Norm and distance

- When E Euclidean space, we define the euclidean norm $||\boldsymbol{x}||_M = \sqrt{< x, x >_M}$
- We can show that $d(\mathbf{x}, \mathbf{y}) = ||\mathbf{x} - \mathbf{y}||$ is a distance in E
- $d_M(\mathbf{x}, \mathbf{y}) = ||\mathbf{x} - \mathbf{y}||_M = \sqrt{< \mathbf{x} - \mathbf{y}, \mathbf{x} - \mathbf{y} >_M} = \sqrt{(\mathbf{x} - \mathbf{y})^T M (\mathbf{x} - \mathbf{y})}$
- For example, $M = I$ $d^2_M(\mathbf{x}, \mathbf{y}) = \sum_j (x_j - y_j)^2$, $M = (1/s_j^2)$, $d^2_M(\mathbf{x}, \mathbf{y}) = \sum_j (\frac{x_j}{s_j} - \frac{y_j}{s_j})^2$

Distance on vector spaces

- Manahattan distance: $d(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{p} |x_j - y_j|$
- Minkowski distance: $d(\mathbf{x}, \mathbf{y}) = (\sum_{j=1}^{p} |x_j - y_j|^p)^{1/p}$
- Mahalanobis distance: it takes into account the correlations between the variables (Σ a variance matrix) $d^2_{\Sigma^{-1}}(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})\Sigma^{-1}(\mathbf{x} - \mathbf{y})$
- etc.

Illustration: 4 measures on 3 objets

ident	z1	z2	z3	z4
p8	22	30	19	20
p15	22	36	24	20
p22	26	34	22	21

Compute distances between objects

- $d^2(p22, p15) = 4^2 + 2^2 + 2^2 + 1 = 25$, $d^2(p22, p8) = 4^2 + 4^2 + 3^2 + 1 = 42$
- p22 is closer p15 than p8

data normalized: how ?

ident	z1	z2	z3	z4
p8	0.24176	0.32967	0.20879	0.21978
p15	0.21569	0.35294	0.23529	0.19608
p22	0.25243	0.33010	0.21359	0.20388

compute the distances

- $d^2(p22, p15) \geq d^2(p22, p8)$, p22 is closest p8 than p15
- Introduction to χ^2 distance

Dissimilarity

- $\forall \boldsymbol{x} \in \Omega,\ d(\boldsymbol{x}, \boldsymbol{x}) = 0$
- $\forall \boldsymbol{x}, \boldsymbol{y} \in \Omega,\ d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x})$

Example of Dissimilarity data matrix

	a	b	c	d	e
a	0				
b	0.2	0			
c	1	1.05	0		
d	0.7	0.75	0.3	0	
e	1	0.8	1.5	1.3	0

Similarity

- $\forall \boldsymbol{x} \in \Omega,\ s(x, x) = s_{max}$
- $\forall \boldsymbol{x} \in \Omega,\ d(x, y) = s_{max} - s(x, y)$

Example of Similarity data matrix

	X	Y	Z	T	W
X	40				
Y	20	40			
Z	15	39	40		
T	7	25	32	40	
W	10	38	30	10	40

Outline

Partition

- Given a finite set Ω,
- $\mathbf{z} = \{(z_1, z_2, ..., z_K); z_k \neq \emptyset; z_k \subset \Omega\}$ is a *partition* if
  - $\forall k \neq \ell$, $z_k \cap z_\ell = \emptyset$ and
  - $\cup_k z_k = \Omega$.
- For such a partition $\mathbf{z}$ into K subsets or clusters $z_1, \ldots, z_K$, each element of Ω belongs to only one cluster, then $\mathbf{z}$ can be represented by the binary classification matrix defined by :

$$\mathbf{z} = \begin{pmatrix} z_{11} & \cdots & z_{1K} \\ \vdots & \ddots & \vdots \\ z_{n1} & \cdots & z_{nK} \end{pmatrix}$$

  where $z_{ik} = 1$ if $i \in z_k$ and 0 otherwise.
- The sum of the ith row values is equal to 1 (each element belongs to only one cluster) and the sum of the kth column values is equal to n_k representing the cardinality of z_k. Here, we consider a hard clustering.

Fuzzy Partition

- Fuzzy sets (Zadeh, 1965) the fuzzy partition seems "natural"
- Fuzzy clustering developed in the beginning of 1970 by Ruspini generalizes the classical approach by extending the notion of membership
- Considering the membership degree coefficients $c_{ik} \in [0, 1]$

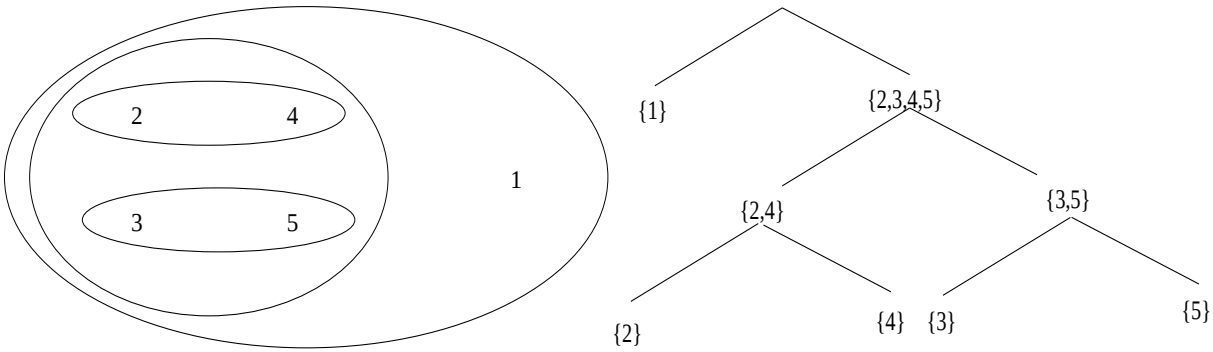
$$\mathbf{c} = \begin{pmatrix} c_{11} & \cdots & c_{1K} \\ \vdots & \ddots & \vdots \\ c_{n1} & \cdots & c_{nK} \end{pmatrix}$$

- A fuzzy partition is represented by a fuzzy classification matrix $\mathbf{c} = \{c_{ik}\}$ verifying the following conditions:
  - $\forall k, \sum_i c_{ik} > 0$
  - $\forall i, \sum_k c_{ik} = 1$
- The second condition considers no empty cluster and the third one expresses the concept of total membership

Definition

- Given Ω a finite set and H a set of non-empty subsets of Ω
- H is then a hierarchy on Ω if
  - $\Omega \in H$
  - $\forall \boldsymbol{x} \in \Omega, \{\boldsymbol{x}\} \in H$
  - $\forall h, h' \in H, h \cap h' = \emptyset$ or $h \subset h'$ or $h' \subset h$
- Example:
  - $\Omega = \{1, 2, 3, 4, 5\}$
  - $H = \{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{2, 4\}, \{3, 5\}, \{2, 3, 4, 5\}, \{1, 2, 3, 4, 5\}\}$

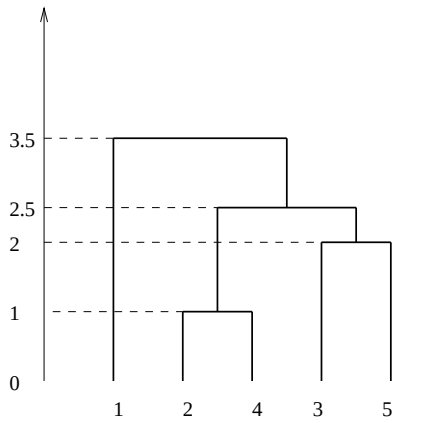
Representations of a hierarchy



These representations are seldom used. One often prefers to associate an index to the hierarchy in order to obtain a readable representation

- The *index* on a hierarchy H is a mapping noted i from H to $\mathbb{R}^+$ verifying the following proprieties :
  - $h \subset h'$ and $h \neq h' \Rightarrow i(h) < i(h')$ (i is a strictly increasing function)
  - $\forall x \in \Omega \quad i(\{x\}) = 0$.
- In the following we note (H, i) the hierarchy with the index i
- Example: By associating to clusters {1},{2},{3},{4},{5},{2,4},{3,5},{2,3,4,5}, {1,2,3,4,5} of the previous hierarchy the values 0,0,0,0,0,1,2,2.5,3.5, one obtains (H, i) which can be represented by a tree data structure often called dendrogram

Representation of (H, i) by a dendrogram



If $\mathbf{z} = (z_1, z_2, \ldots, z_K)$ a partition of Ω, H formed by the clusters z_k, singletons of Ω and Ω itself constitute a hierarchy. Let us notice that conversely, it is possible to associate at each level of (H, i) a partition. Hence, (H, i) corresponds then to a set of encased clusters.

- The number of possible hierarchies and partitions to be defined on Ω quickly becomes enormous when the cardinality of Ω increases
- For instance, the number of partitions of n objects into K clusters is giving by the following formula

$$S(n,K) = \frac{1}{K!}\sum_{k=0}^{K}(-1)^{k-1}C_k^K k^n$$

- When n and K become large, we have $S(n,K) \approx \frac{K^n}{K!}$, for example $S(100,5) \approx 10^{67}$

(n,K)	1	2	3	4	5	6	7	8
1	1							
2	1	1						
3	1	3	1					
4	1	7	6	1				
5	1	15	25	10	1			
6	1	31	90	65	15	1		
7	1	63	301	350	140	21	1	
8	1	127	966	1701	1050	266	28	1

Search for partitions associated to a dissimilarity measure

- Given a dissimilarity measure d on Ω, it is natural to associate for each real α positive or null the following neighbor relation V_α on Ω

$$xV_\alpha y \Leftrightarrow d(x,y) \leq \alpha$$

- We search for the conditions so that there exists a partition of Ω such as all objects belonging to a cluster are close and the elements belonging to distinct clusters are not close.
- For this, it is necessary and sufficient that V_α is an equivalence relation
- The clusters are then the equivalence classes of V_α. As the function d is a dissimilarity measure, the relation V_α is reflexive, symmetric. It is necessary and sufficient that the transitivity is verified,

$$xV_\alpha y \text{ and } yV_\alpha z \quad \Rightarrow \quad xV_\alpha z$$

$$d(x,y) \leq \alpha \quad \text{and} \quad d(y,z) \leq \alpha \Rightarrow d(x,z) \leq \alpha \tag{1}$$

- It is clear that the relation (1) is verified if d is an ultrametric.
- Conversely, so that V_α is an equivalence relation for whatever α, d must be an ultrametric. Indeed, for each triplet x, y et z of Ω, taking $\alpha = \max(d(x,y), d(y,z))$, we have $d(x,y) \leq \alpha$ and $d(y,z) \leq \alpha$ and therefore $d(x,z) \leq \alpha$, what induces the ultrametric inequality.

Ultrametric associated to (H, i): function φ

- Given (H, i) on Ω, we can define δ from $\Omega \times \Omega$ to $\mathbb{R}^+$ by assigning for each couple x, y the smallest index of all clusters of H including x et y
- The function i is increasing with the include relation $(h_1 \subset h_2 \Rightarrow i(h_1) \leq i(h_2))$
- $\delta(x, y)$ can be considered as the index of the smallest cluster according to H and containing x et y. Then we can show that δ is an ultrametric on Ω

(H, i) associated to an ultrametric: function ψ

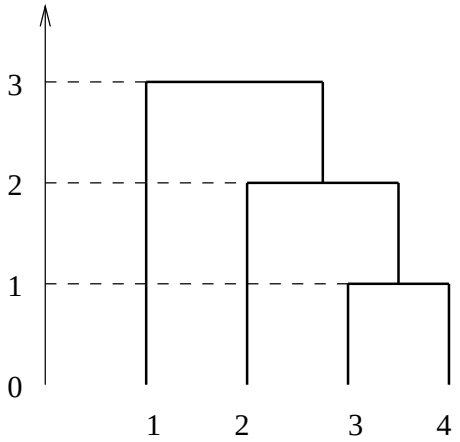
- We consider the V_α relations on Ω previously defined, but from the ultrametric δ
- We know that the V_α's are equivalence relations for each $\alpha \geq 0$
- Let be D_δ the set of values taken by δ on Ω, we define the set H as the set of all equivalence classes of V_α's when α coverts D_δ
- Taken as function i on H the function *diameter* $(i(h) = \max_{x,y \in h} \delta(x, y))$, we can show that (H, i) forms an indexed hierarchy on Ω

Hierarchy and ultrametric

- Finally, the functions φ et ψ are reciprocal and there is then an equivalence between (H, i) and an ultrametric

Function φ

- The application of the function φ on the hierarchy (H_1, i_1) (a) implies the ultrametric δ_1 of (b)



a : (H_1, i_1)

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

b : $\delta_1 = \varphi(H_1, i_1)$

Function ψ

- The application of the function ψ to the ultrametric δ_1 implies : we have $D_\delta = \{0, 1, 2, 3\}$. The equivalence classes of the 4 relations R_α are $R_0 : \{1\}, \{2\}, \{3\}, \{4\}$, $R_1 : \{1\}, \{2\}, \{3, 4\}$, $R_2 : \{1\}, \{2, 3, 4\}$ and $R_3 : \{1, 2, 3, 4\}$.
- The obtained hierarchy is then $\{\{1\}, \{2\}, \{3\}, \{4\}, \{3, 4\}, \{2, 3, 4\}, \{1, 2, 3, 4\}\}$ and the associated indexes to subsets of the hierarchy are respectively $(0, 0, 0, 0, 1, 2, 3)$. One finds then (H_1, i_1)

Outline

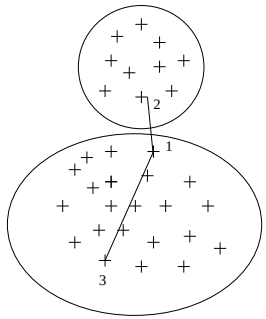
Introduction

- The aim of clustering is to organize the objects of Ω into homogeneous clusters. To define the notion of homogeneity, we often use a similarity or dissimilarity measures on Ω
- For example, if d is a dissimilarity measure, we can characterize this homogeneity by imposing to clusters of the partition to verify the following propriety

$$\forall x, y \in \text{ same cluster cluster and } \forall z, t \in \text{ distinct clusters} \Rightarrow d(x, y) < d(z, t)$$

- This property means that we aim to obtain clusters such that two objects of the same cluster are more similar than two objects not belonging to the same cluster
- In practice, this objective is not used

Illustration



Partition

- For these reasons, several approaches are then used to replace this objective difficult to reach
- We replace this too stringent condition by a numerical function which will measure the partition homogeneity quality. This function is commonly called *criterion*. The problem can then appear more simple. Indeed, for example, in the case of research for a partition, it is enough to seek among the finite set of all the partitions that which optimizes the numerical criterion
- Unfortunately, the number of partitions is very large, their enumeration is impossible in a realistic time (combinatorial problem). Generally, we use then heuristics giving not the best solution but a *good* one, close to the optimal solution
- We consider therefore an local optimization. When we have an order structure on the finite set Ω and this one must be respected by the partition, it exists a dynamic programming method by Fisher (1958) algorithm giving an optimal solution

Example of criterion

- The within-cluster variance can be used when the finite set Ω to cluster corresponds to a set of n objects described par p quantitative variables. It is then possible, as for the component principal analysis, to associate a cloud of points in $\mathbb{R}^p$ provided by a weight equal to $\frac{1}{n}$ for each element and the Euclidean metric
- Matrix of variance can then taking this form

$$S = \frac{1}{n}(\mathbf{x} - \mathbb{1}_n\overline{x})^T(\mathbf{x} - \mathbb{1}_n\overline{x}) = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{x})(\boldsymbol{x}_i - \overline{x})^T$$

  and the variance $I = \frac{1}{n}\sum_{i=1}^{n} d^2(\boldsymbol{x}_i, \overline{x})$ verifies $I = \text{trace}(S)$
- If $\mathbf{z} = (z_1, \ldots, z_K)$ is a partition of Ω into K clusters, $\mathbf{x}_k$ the matrix $\mathbf{x}$ summarized to rows corresponding to the kth cluster z_k and $\overline{x}_k$ its vector means, we can define the within variance matrix

$$W = \frac{1}{n}\sum_{k=1}^{K} n_k W_k$$

  where W_k is the variance matrix of the cluster z_k defined by

$$W_k = \frac{1}{n}(\mathbf{x}_k - \mathbb{1}_n\overline{x}_k)^T(\mathbf{x}_k - \mathbb{1}_n\overline{x}_k)$$

Within-cluster variance

- The within-cluster variance noted W is written as $W(\mathbf{z}) = \sum_{k=1}^{K} I(z_k)$ where $I(z_k) = W(z_k) = \frac{1}{n} \sum_{i \in z_k} d^2(\boldsymbol{x}_i, \overline{x})$ is the variance of the cluster z_k
- We can show the following relation $W(\mathbf{z}) = \text{trace}(W_k)$
- It is then possible to use the within-cluster variance as a clustering criterion: a partition will be as much homogeneous than within-cluster variance is close to 0; in particular, this criterion will be equal to 0 when the objects of each cluster are confused in an object.
- The problem is to define directly an algorithm giving homogeneous clusters taking into account the dissimilarity measure. It is easy to propose such algorithms but the difficulty is to prove that the obtained clusters are interesting and give answers to our aim. In fact the algorithmic and numerical approaches often converge. Several proposed algorithms without any reference to a criterion and, giving good results, optimize a numerical criterion. This is the case of the famous k-means algorithm

Hierarchy

- In the case of hierarchical clustering, we aim to obtain clusters as much more homogeneous that they are located in the bottom of the tree structure. The definition of criterion is less easy. We will see that it is possible to do it by using the concept of ultrametric (ultrametric optimal)

Outline

Objective

- The aim of hierarchical methods is to create a hierarchical decomposition (H, i) of Ω. On Ω we have a dissimilarity measure such that the closest objects are grouped in the clusters with the smallest index. It exists two principal approaches

Description

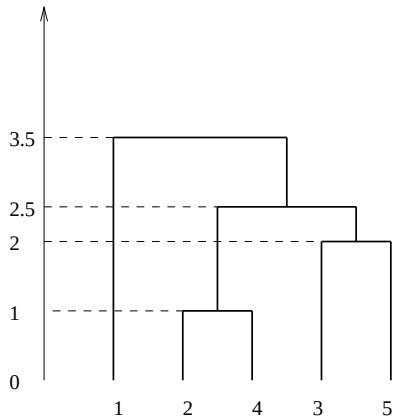
- Divisive approach: this approach is also called top-down approach. It starts with just one cluster containing all objects. In each successive iteration, we split up clusters into two or more clusters until generally each object is in one cluster. Note that other stop conditions can be used and the division into clusters are defined by the verification or not of a property. For example, in taxonomy, we split up animals into vertebrates and non vertebrates

- Agglomerative approach: opposed to the divisive approach, we start by assuming n clusters, each object forms a singleton cluster. In each successive iteration, we merge the closest clusters until obtaining one cluster which is the set Ω. In the following, we focus on this approach which is the most frequently used

Construction of the hierarchy

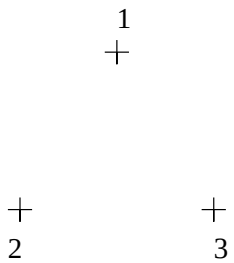
- While the process of grouping the clusters in the agglomerative approach, it is necessary to define a distance between the clusters in order to merge the closest ones. Generally, from the dissimilarity measure on Ω we define a *distance* D between the clusters. In fact, D is just a dissimilarity measure. We see later different manners to define these kinds of measures. Now, we briefly present the different steps of the algorithm :
  1. **Initialization**: Each object is a singleton partition, compute the dissimilarity measure between these objects.
  2. **Repeat**
     - merge two closest clusters according to D,
     - compute the distance between the obtained new cluster and the old clusters not merged.
  3. **Until** the number of clusters is equal to 1

It is easy to show that the set of clusters defined during the successive iterations form an hierarchy



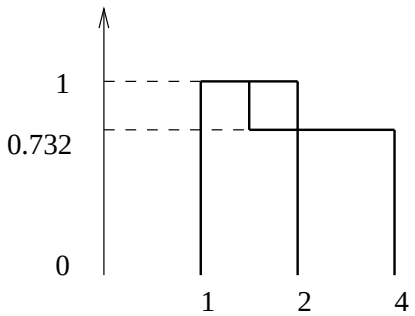
Construction of the index

- Once a hierarchy is defined, it is necessary to associate it an index. For the singleton clusters, this index is necessarily equal to 0. For the other clusters, this index is generally defined by associating to each new agglomerated cluster the dissimilarity measure D which evaluates the proximity between the merged clusters to form this new cluster. Note that, in order to have a property of index, the proposed ones have to *increase strictly* with the level of hierarchy. Then several difficulties can appear
- Inversion problem: For a certain D, the index defined is not necessarily strictly increasing function, this leads with the *inversion* problem. For example, if the data are formed by three points of the plan located at the top of an equilateral triangle with a side equal to 1 and if one takes as D the distance between the cluster means, one obtains an inversion illustrated hereafter. With the family of D studied in this course, it is possible to show that the inversion is impossible



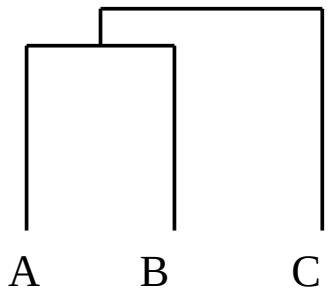
	1	2	3
1	0		
2	1	0	
3	1	1	0

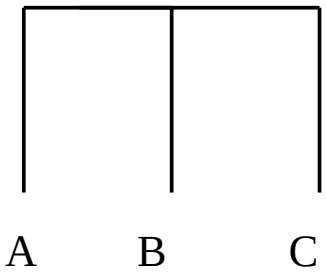
	{1,2}	3
{1,2}	0	
	0.732	0



Not strict increasing

- When there is equality of index for several encased levels, it suffices to *filter* the hierarchy, i.e. to preserve only one cluster containing all the encased classes having the same index. In the following example, the cluster $A \cup B$ having the same index as the cluster $A \cup B \cup C$ can be removed





This problem can occur with the family of D considered in the following and the associated algorithms will thus require to envisage this operation of filtering

Agglomerative criteria

- In the following, the different kinds of D are designated as agglomerative criteria or approaches. There exist several criteria but the most used are:
  - Single linkage or Nearest Neighbor approach (Sibson, 1973)
$$D(A,B) = \min\{d(i,i'), i \in A \text{ et } i' \in B\};$$
  - Complete linkage or farthest Neighbor approach (Sorenson, 1948)
$$D(A,B) = \max\{d(i,i'), i \in A \text{ et } i' \in B\};$$
  - Average linkage (Sokal and Michener, 1958)
$$D(A,B) = \frac{\sum_{i \in A} \sum_{i' \in B} d(i,i')}{n_A . n_B}$$

  where n_E represents the cardinality of the cluster E.

Recurrence formulas of Lance and Williams, 1967

- For the three agglomerative criteria, there exist calculus simplification relations of the distances between clusters necessary for the agglomerative hierarchical clustering (AHC) algorithm, without these kinds of relations it would be prohibitory in time calculation. These relations called generally recurrence formulas of Lance and Williams, are for the three agglomerative criteria single, complete and average linkage:

$$D_{\text{min}}: \quad D(A, B \cup C) = \min\{D(A, B), D(A, C)\};$$

$$D_{\text{max}}: \quad D(A, B \cup C) = \max\{D(A, B), D(A, C)\};$$

$$D_{\text{average}}: \quad D(A, B \cup C) = \frac{n_B.D(A, B) + n_C.D(A, C)}{n_B + n_C}.$$

Note that these formulas can be deducted from the general formula :

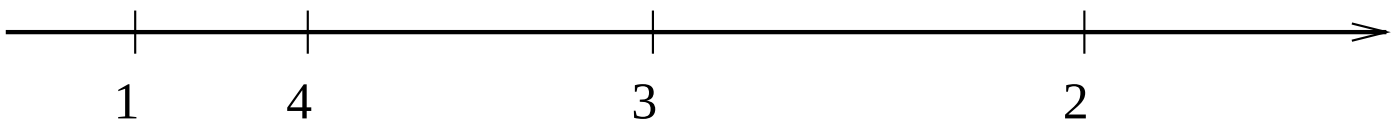
$$D(A, B \cup C) = \alpha_1 D(A, B) + \alpha_2 D(A, C) + \beta D(B, C) + \gamma |D(A, B) - D(A, C)|.$$

Then D_{min} is obtained by taking $\alpha_1 = \alpha_2 = 0.5, \beta = 0, \gamma = -0.5$, D_{max} by taking $\alpha_1 = \alpha_2 = 0.5, \beta = 0, \gamma = 0.5$ and D_{average} by taking $\alpha_1 = \frac{n_B}{n_B + n_C}$, $\alpha_2 = \frac{n_C}{n_B + n_C}$, $\beta = 0$, $\gamma = 0$

Example

- Hereafter, we consider 4 aligned points, separated successively by the distances 2, 4 and 5: We take as dissimilarity measures between these points, the usual Euclidean distance and we carry out the AHC algorithm according to the three agglomerative criteria, The results are reported in the following

Data

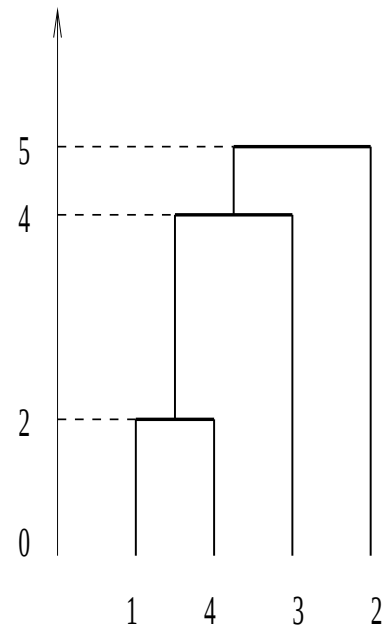


D_{min}

	1	2	3	4
1	0			
2	11	0		
3	6	5	0	
4	2	9	4	0

	{1,4}	2	3
{1,4}	0		
2	9	0	
3	4	5	0

	{1,4,3}	2
{1,4,3}	0	
2	5	0

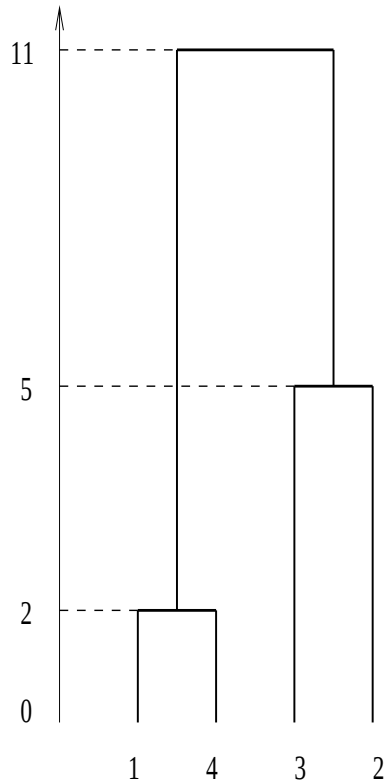


$D_{\max}$

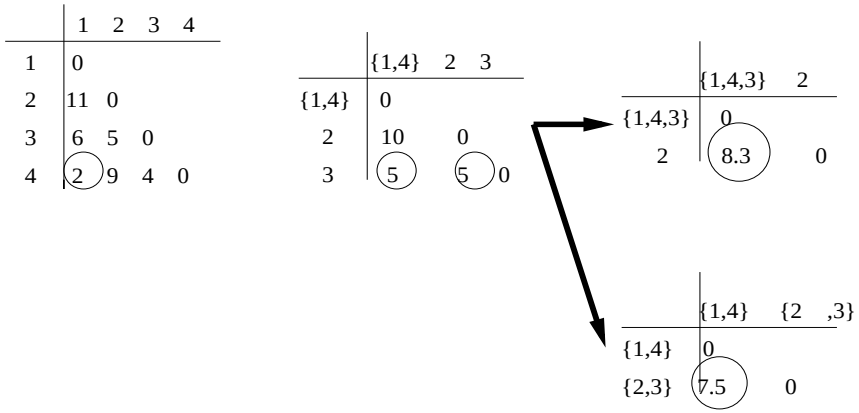
	1	2	3	4
1	0			
2	11	0		
3	6	5	0	
4	2	9	4	0

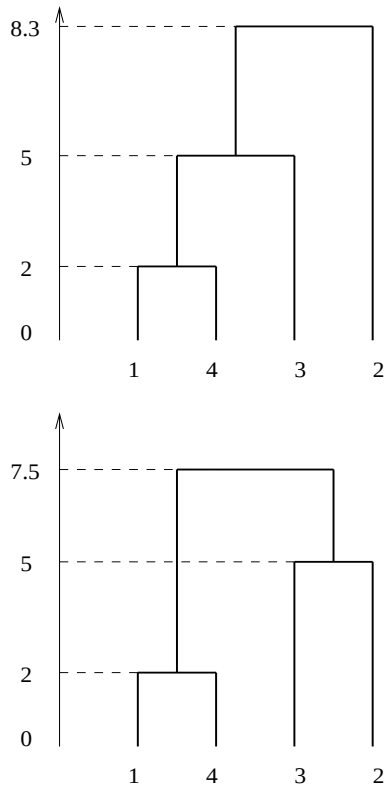
	{1,4}	2	3
{1,4}	0		
2	11	0	
3	6	5	0

	{1,4}	{2,3}
{1,4}	0	
{2,3}	11	0



Daverage





Note that in the last case, we can obtain two different solutions whether we choose to merge the clusters {1,4} and {3} or the clusters {2} and {4}.

Ward Method or Minimum variance approach

- Unlike the criteria previously described, the Ward criterion (Ward, 1963) requires raw data than the dissimilarity between objects. When the set Ω to classify is associated to a cloud of points in $\mathbb{R}^p$ provided by a weight equal to $\frac{1}{n}$ for each element and the Euclidean metric, the criterion takes the following form

$$D(A,B) = \frac{n_A n_B}{n_A + n_B} d^2(\mu_A, \mu_B)$$

where μ_E represents the center of the set E. The associated AHC algorithm is often called the Ward method (Ward, 1963). There also exists in this case a recurrence formula:

$$D(A, B \cup C) = \frac{(n_A + n_B) \times D(A,B) + (n_A + n_C) \times D(A,C) - n_A \times D(B,C)}{n_A + n_B + n_C},$$

which can deduced from the general recurrence formula,

$$D(A, B \cup C) = \alpha_1 D(A,B) + \alpha_2 D(A,C) + \beta D(B,C) + \gamma |D(A,B) - D(A,C)|,$$

with $\alpha_1 = \frac{n_A + n_B}{n_A + n_B + n_C}$, $\alpha_2 = \frac{n_A + n_C}{n_A + n_B + n_C}$, $\beta = -\frac{n_A}{n_A + n_B + n_C}$ and $\gamma = 0$

Analysis of Flying Mileages Between Ten U.S. Cities

Atlanta	Chicago	Denver	Houston	LA	Miami	NewYork	SanFrancisco	Seattle	WashingtonD.C
0									
587	0								
1212	920	0							
701	940	879	0						
1936	1745	831	1374	0					
604	1188	1726	968	2339	0				
748	713	1631	1420	2451	1092	0			
2139	1858	949	1645	347	2594	2571	0		
2182	1737	1021	1891	959	2734	2408	678	0	
543	597	1494	1220	2300	923	205	2442	2329	0

Applications

- Single linkage
- Complete linkage
- Average linkage
- Ward linkage

Introduction

- We have seen that the concept of (H, i) is equivalent to the concept of ultrametric. The AHC algorithm transforms then an initial dissimilarity measure d into a new dissimilarity measure δ having the property of an ultrametric. The aim of hierarchical clustering could be then posed in these terms: Find the closest ultrametric δ to the dissimilarity measure d.

- It remains to provide a distance to the space of dissimilarity measures on Ω. We can use, for example

$$\Delta(d, \delta) = \sum_{i,i' \in \Omega} (d(i, i') - \delta(i, i'))^2$$

or

$$\Delta(d, \delta) = \sum_{i,i' \in \Omega} |d(i, i') - \delta(i, i')|.$$

Unfortunately, it is a difficult problem and we now will study the optimality properties of the various algorithms previously described.

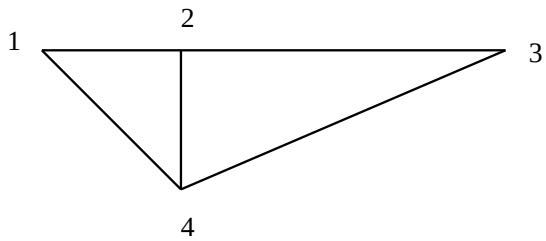
Hierarchy of single linkage

- Let U be the set of all ultrametrics smaller than the initial dissimilarity measure d ($\delta \in U \Leftrightarrow \forall i, i' \in \Omega, \delta(i, i') \leq d(i, i')$)
- Note that δ_m is a higher envelope of U, i.e the mapping from $\Omega \times \Omega$ to $\mathbb{R}$ verifying $\forall i, i' \in \Omega, \delta_m(i, i') = \sup\{\delta(i, i'), \delta \in U\}$
- We can then show that δ_m remains an ulrametric and this one is the ultrametric obtained by the AHC algorithm with the single linkage criterion and moreover it is, among all utrametrics less than d, the closest of d with respect to Δ
- This ultrametric is commonly called *sub-dominant* ultrametric

Property of single linkage

- Link between this hierarchical clustering and the determining of the minimum spanning tree, a well-known problem in graph theory. We consider the definite complete graph on Ω.
- Each edge (a, b) of this graph is valued by the distance $d(a, b)$. We can show that the search for this tree is equivalent to the search for the sub-dominant ultrametric

- To find the tree associated to the single linkage, it is possible to use the algorithms which were developed to find a minimum spanning tree for a connected weighted graph, for instance the Prim's and Kruskal's algorithms (Prim, 1957) and (Kruskal, 1956). We can illustrate this on a small example



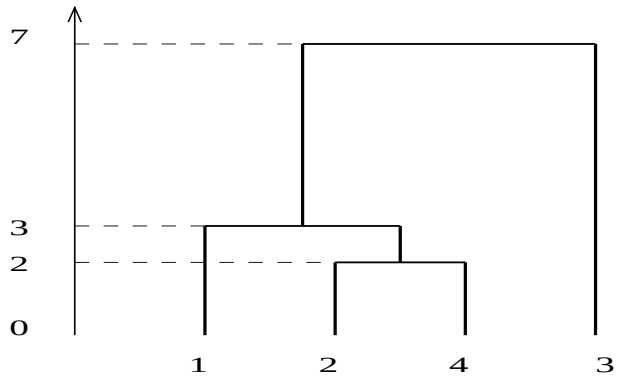
- For the Kruskal's algorithm see for instance in *Wikipedia* a description illustrated by a simple example

Construction of the sub-dominant ultrametric

	1	2	3	4
1	0			
2	3	0		
3	10	7	0	
4	3.6	2	7.3	0

	1	{2,4}	3
1	0		
{2,4}	3	0	
3	10	7	0

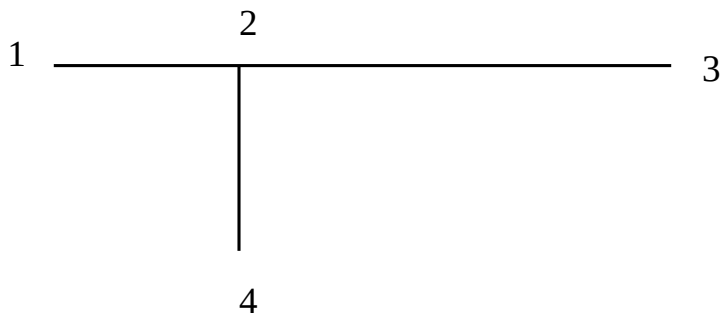
	{1,2,4}	3
{1,2,4}	0	
3	7	0



	1	2	3	4
1	0			
2	3	0		
3	7	7	0	
4	3	2	7	0

By retaining from the initial complete graph only 3 edges having participating in the algorithm, the edge (2,4) of length 2, the edge (1,2) of length 3 and the edge (2,3) of length 7, we obtain the minimum spanning tree

Minimum spanning tree



Hierarchy of complete Linkage

- The ultrametric is higher than the dissimilarity measure d. Unfortunately, the properties of the ultrametric provided by the AHC algorithm are not also interesting as those of the sub-dominant ultrametric.
- In particular, there is not necessarily an uniqueness of dendrogram. For example, we can obtain different results if we change the order of agglomerations of clusters due to the presence of two or several smallest indexes in the intermediate dissimilarity matrices
- This point is also observed in the previous example for the average criterion
- As before we attempt proceed to define the closest ultrametric higher than d. As for the sub-dominant ultrametric defined from the higher envelope of the ultrametrics smaller than d, we consider this time the lower envelope of the ultrametrics greater than d.
- Unfortunately this envelope is not necessarily an ultrametric

counterexample: d and ultrametrics δ_1 et δ_2

	a	b	c
a	0		
b	1	0	
c	2	1	0

	a	b	c
a	0		
b	1	0	
c	2	2	0

	a	b	c
a	0		
b	2	0	
c	2	1	0

We can verify that δ_1 (second table) and δ_2 (third table) are two ultrametrics higher to d (first table) defined with the 3 points a,b, c and that the lower envelope of these ultrametrics is just d. Hence, this envelope of the all ultrametrics higher to d is necessarily d which is not an ultrametric

Hierarchy of Average linkage

- The average criterion does not check any problem of optimality, but in practice it has been showed that is close to the ultrametric minimizing

$$\sum_{i,i' \in \Omega} (d(i,i') - \delta(i,i'))^2$$

The Ward method

- Let $\mathbf{z} = (z_1, \ldots, z_K)$ be a partition and $\mathbf{z}'$ the partition obtained from $\mathbf{z}$ by merging the clusters $\mathbf{z}_k$ and $\mathbf{z}_\ell$. We can show the following equality:

$$W(\mathbf{z}') - W(\mathbf{z}) = \frac{n_k n_\ell}{n_k + n_\ell} d^2(\overline{x}_k, \overline{x}_\ell)$$

- The merge of two classes increases necessarily the within-cluster variance. It is then possible to propose an AHC algorithm which merges at each stage the two classes increasing the least possible the within-cluster variance, i.e. minimizing the following expression:

$$D(A,B) = \frac{n_k n_\ell}{n_k + n_\ell} d^2(\overline{x}_k, \overline{x}_\ell),$$

  and we find the Ward criterion

- The AHC algorithm has then a local optimum: at each stage, we seeks to minimize the within-cluster variance

Distance

- Although used for many years, the first difficulties for hierarchical clustering are the choice of the dissimilarity measure on Ω and of the agglomerative criterion. Indeed, the quality of a partition and its interpretation highly depend on them.
- The choice of distance or dissimilarity used on Ω plays an important role, for example the measurement unit can affect the clustering, then a simple standardization of original data is often necessary
- Different kind of transformations depending on the type of variables (continuous, binary, nominal or ordinal etc.) are available in the literature
- In the absence of information permitting to employ the appropriate distance, the squared distance is the most used distance for continuous data.

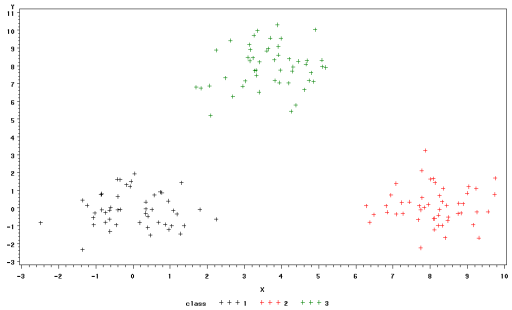
Criterion

- Once this difficulty is overcomed it remains the problem of the choice of agglomerative criterion (Choice of D), we have seen that exist several criteria and the dendrograms obtained from them can be not identical, then which criterion is more adapted to clustering ?
- This question is crucial and has been extensively studied under different approaches for hierarchical and nonhierachical methods. We retain that the sizes, shapes and presence of outliers influence on the obtained results
- we provide some practical guidelines to select the appropriate criterion

Remarks

- When there is clearly a clustering structure without outliers and the clusters are well separated, the different criteria can give the same dendrogram

Example



- From the theoretical point of view, the single linkage satisfies a certain number of desirable mathematical properties. But in practice, it outperforms other studied criteria only when the clusters are elongated or irregulated and there is not any chain of points between the clusters, it is very prone to chaining effects
- The complete linkage allows one to integrate the outliers in the process of training clusters and avoids therefore the clusters with a single outlying objects
- Even if it has not any theoretical property, this average criterion tends to produce clusters that reflect accurately the structure present in data but can require a great deal of computation. It tends to join clusters with small variances, and it is slightly biased toward producing clusters with the same variance

- When the variables are continuous, the advised criterion is the within-cluster variance even if it deals with great computation. The results are then jointly used with those of PCA because the Ward is the more adapted. Taking $\beta < 0$ in the recurrence formula, we can avoid the problem of outliers. However, it tends to give spherical clusters of nearly equal sizes, this remark will be commented under the mixture approach.
- It is often necessary to have some tools facilitating the interpretation and some tools allowing to decrease the number of levels of the dendrogram such as the classical *semi-partial R-square* (SPRQ)

$$SPRSQ = \frac{W(z_k \cup z_\ell) - W(z_k) - W(z_\ell)}{I},$$

  which expresses the loss of homogeneity when the clusters z_k and z_ℓ are agglomerated. This loss decreases with the number of cluster then a scree plot showing one or several *elbows* can be used to propose a cut of the dendrogram. Note
- Note that the problems arising from the time complexity of AHC algorithms which depend on the linkage chosen, is solved in practice by using more effective algorithms (De Rham, 1980) based on the construction of nearest neighbor chains and carrying out agglomerations whenever reciprocal nearest are encountered.

Drawbacks of methods and new algorithms

- By their and tree structure, the hierarchical methods have a great success. Unfortunately, because their complexity, AHC is not adapted for large data. In addition, in the merge process, once a cluster is formed, it does not undo what was previously, then no modification of clusters or permutation of objects are possibles
- Finally, the AHC algorithm with the fourth criteria studied gives generally convex clusters and are fragile in the presence of outliers

New algorithms

- Other approaches can be used to correct these weaknesses such as
  1. CURE (Clustering using Representatives) by Guha et al.
  2. CHAMELEON based on k-nearest neighbor graph
  3. BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) by Zhang et al.
- The two latter are particularly interesting for large data sets. In this context, we will see that the nonhierarchical methods alone or combined to hierarchical methods are more preferable. Next, we will focus on these kind of methods.

Papers to read

Density Linkage

- The phrase density linkage is used here to refer to a class of clustering methods using nonparametric probability density estimates (for example, Hartigan 1975; Wong 1982; Wong and Lane 1983). Density linkage consists of two steps:
  1. A new dissimilarity measure, d^*, based on density estimates and adjacencies is computed. If x_i and x_j are adjacent (the definition of adjacency depends on the method of density estimation), then $d^*(x_i, x_j)$ is the reciprocal of an estimate of the density midway between x_i and x_j; otherwise, $d^*(x_i, x_j)$ is infinite.
  2. A single linkage cluster analysis is performed using d^*

Uniform-Kernel Method

- The uniform-kernel method uses uniform-kernel density estimates. Consider a closed sphere centered at point x with radius r. The estimated density at x, $f(x)$ is the proportion of observations within the sphere divided by the volume of the sphere. The new dissimilarity measure is computed as

$$\begin{cases} d^*(x_i, x_j) = \frac{1}{2}\left(\frac{1}{f(x_i)} + \frac{1}{f(x_j)}\right) \text{ if } d(x_i, x_j) \leq r, \\ d^*(x_i, x_j) = \infty \text{ otherwise} \end{cases}$$

kth-Nearest Neighbor Method

- The kth-nearest-neighbor method (Wong and Lane 1983) uses kth-nearest neighbor density estimates. Let $r_k(x)$ be the distance from point x to the kth-nearest observation. Consider a closed sphere centered at x with radius $r_k(x)$. The estimated density at x, $f(x)$ is the proportion of observations within the sphere divided by the volume of the sphere. The new dissimilarity measure is computed as

$$\begin{cases} d^*(x_i, x_j) = \frac{1}{2}(\frac{1}{f(x_i)} + \frac{1}{f(x_j)}) \text{ if } d(x_i, x_j) \leq max(r_k(x_i), r_k(x_j)), \\ d^*(x_i, x_j) = \infty \text{ otherwise} \end{cases}$$

Wong and Lane (1983) show that kth-nearest-neighbor density linkage is strongly set consistent for high-density (density-contour) clusters if k is chosen such that $k/n \longrightarrow 0$ and $k/\ln(n) \longrightarrow \infty$ when $n \longrightarrow \infty$

Comments on density linkage

- Density linkage applies no constraints to the shapes of the clusters and, unlike most other hierarchical clustering methods, is capable of recovering clusters with elongated or irregular shapes
- Problem of choice the values of smoothing parameters k and r (Hybrid method, Wong 1982)

Outline

Dissimilarity data matrix

	a	b	c	d	e
a	0				
b	0.2	0			
c	1	1.05	0		
d	0.7	0.75	0.3	0	
e	1	0.8	1.5	1.3	0

Criteria

- Single criterion
- Complete criterion
- Average criterion

Results

- Dendrogram
- Ultrametric
- What is the appropriate dendrogram ?

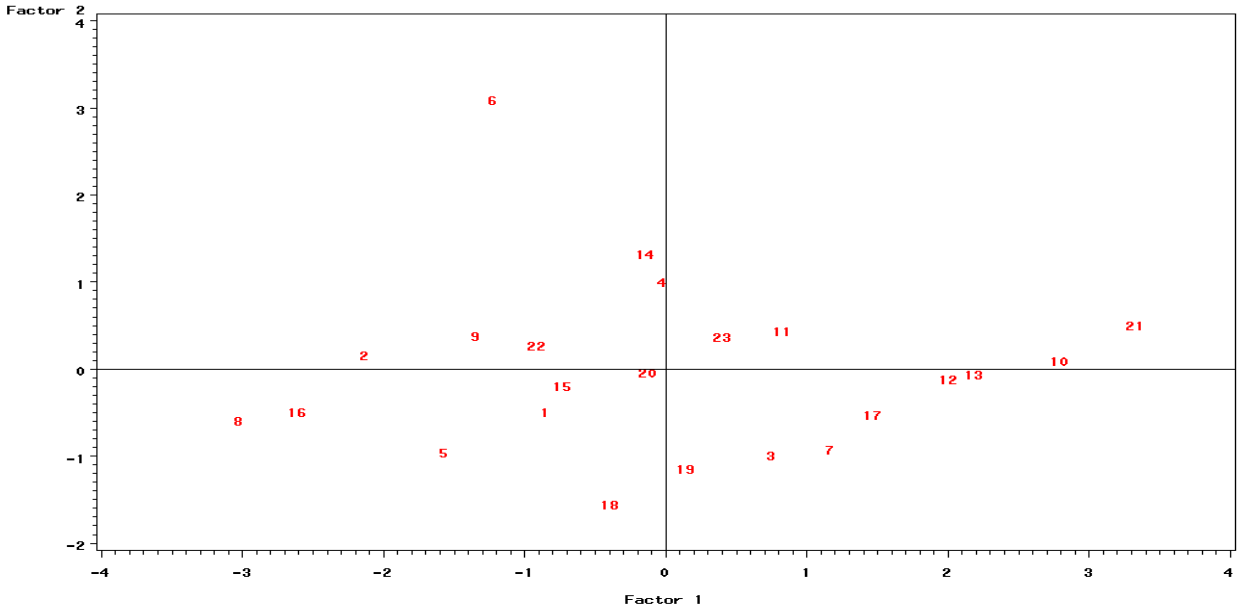
Measures on 23 Butterflies

num	Z1	Z2	Z3	Z4
1	22	35	24	19
2	24	31	21	22
3	27	36	25	15
4	27	36	24	23
5	21	33	23	18
6	26	35	23	32
7	27	37	26	15
8	22	30	19	20
9	25	33	22	22
10	30	41	28	17
11	24	39	27	21
12	29	39	27	17
13	29	40	27	17
14	28	36	23	24
15	22	36	24	20
16	23	30	20	20
17	28	38	26	16
18	25	34	23	14
19	26	35	24	15
20	23	37	25	20
21	31	42	29	18
22	26	34	22	21
23	24	38	26	21

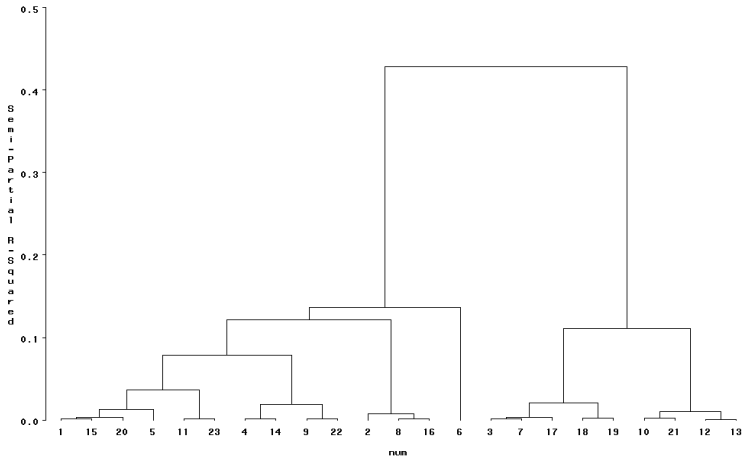
Measures on 23 Butterflies

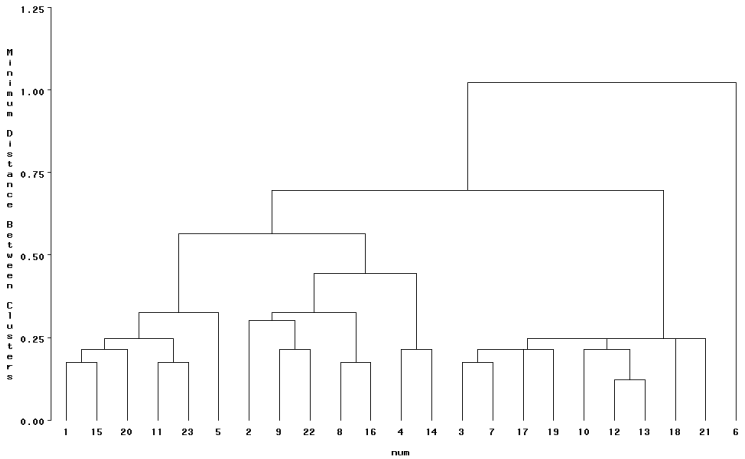
- Problem of clustering
- Use just basic tools of visualization (SAS/Insight)

Principal components analysis



Dendrograms with ward and single criterion





Outline

Clustering of variables

- When the variables are continuous, the principal aim is often to give clusters where each cluster include variables correlated then the correlation matrix can be converted to a dissimilarity matrix by replacing each correlation $\rho_{jj'}$ between two variables j and j' by

$$1 - \rho_{jj'}$$

  which is an Euclidean distance

- However, unless that all correlations are positive, this distance is not appropriate, indeed two variables highly negatively correlated may also be considered to be very similar. Then in this case

$$1 - |\rho_{jj'}|$$

  appears more adapted. In the other hand, since the correlation has an interpretation as the cosine of the angle between two vectors corresponding to two variables, the dissimilarity expressed as

$$\arccos(|\rho_{jj'}|)$$

  can be used. Once the dissimilarity matrix defined the AHC method with different criteria previously listed can be employed.

- Furthermore, there is an other approach commonly used based on the factor analysis. The different steps described below lead with a Descendent Hierarchical algorithm.

Conclusion

Advantages

- Simple methods
- Give readable results
- Complementary to PCA or MDS
- Extension to contingency tables or categorical data by using the correspondence analysis
- Can be applied (in certain cases) for the variables (see course 5)
- Methods available in Statistic and data mining Software (See **R)**

Disadvantages

- Complexity
- Depend on the shape of clusters

Course 2

- Nonhierarchical methods