UNISS
UNIVERSITÀ
DEGLI STUDI
DI SASSARI

UNIVERSITY OF
GEORGIA

CLUSTER ANALYSIS

CORRADO DIMAURO, ALBERTO CESARANI, NICOLO' P.P. MACCIOTTA

ATHENS, JANUARY 2025

Let’s consider the usual matrix of data X: *n* = objects and *q* = variables

$$\mathbf{X} = \begin{pmatrix} x_{11} & \cdot & \cdot & x_{1q} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ x_{n1} & \cdot & \cdot & x_{nq} \end{pmatrix}$$

In many areas of research, there is often the need to divide the *n* objects, on which the *q* variables are measured, in different groups

Human beings

Economic status

Lower class

Middle class

Upper class

Alchol consumption

Lower class

Middle class

Upper class

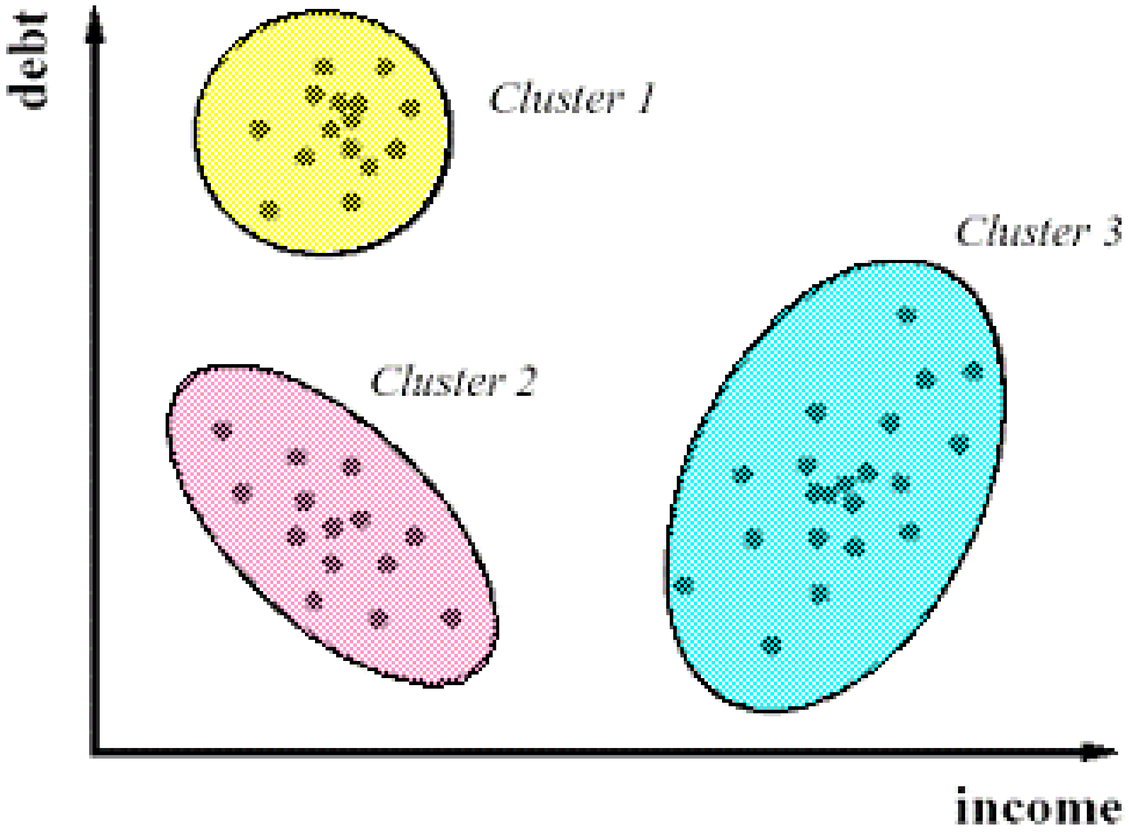
Variables used to classify with respect the economic status or the annual consumption of alcohol are different

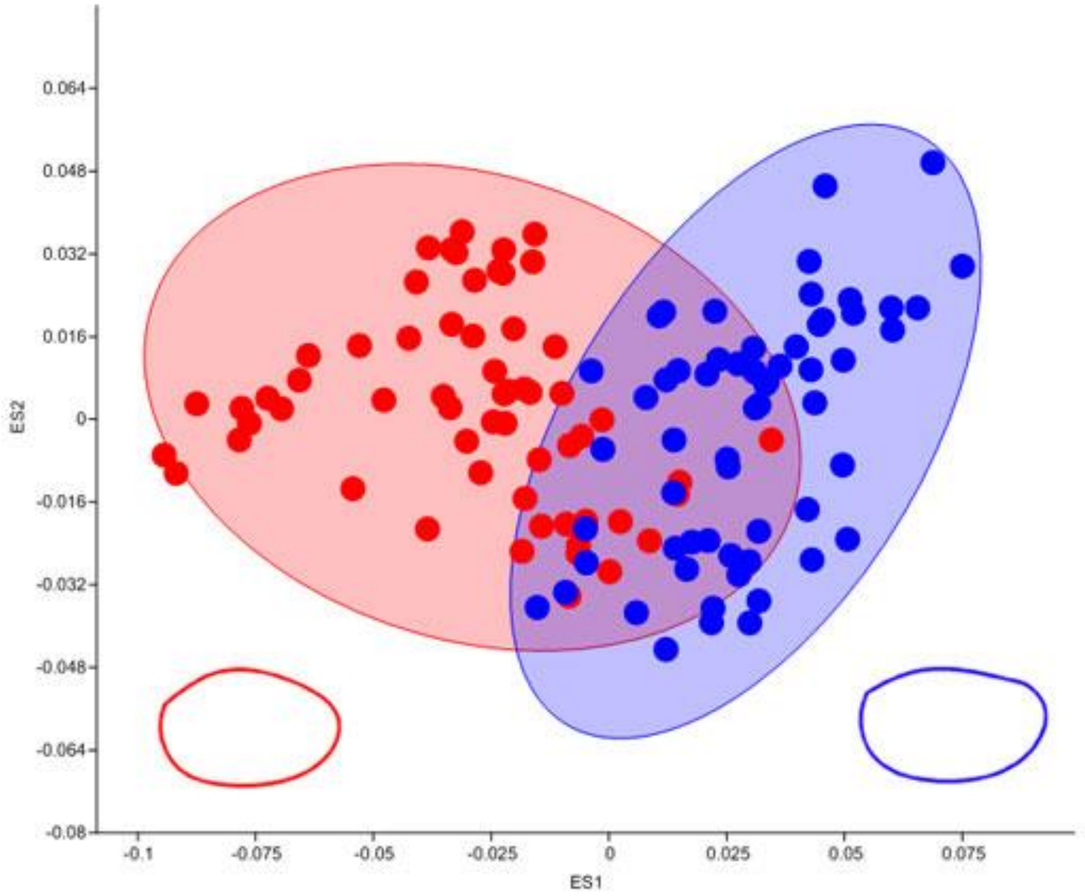
In the example of human beings we already know that three groups exist. We will use the DA

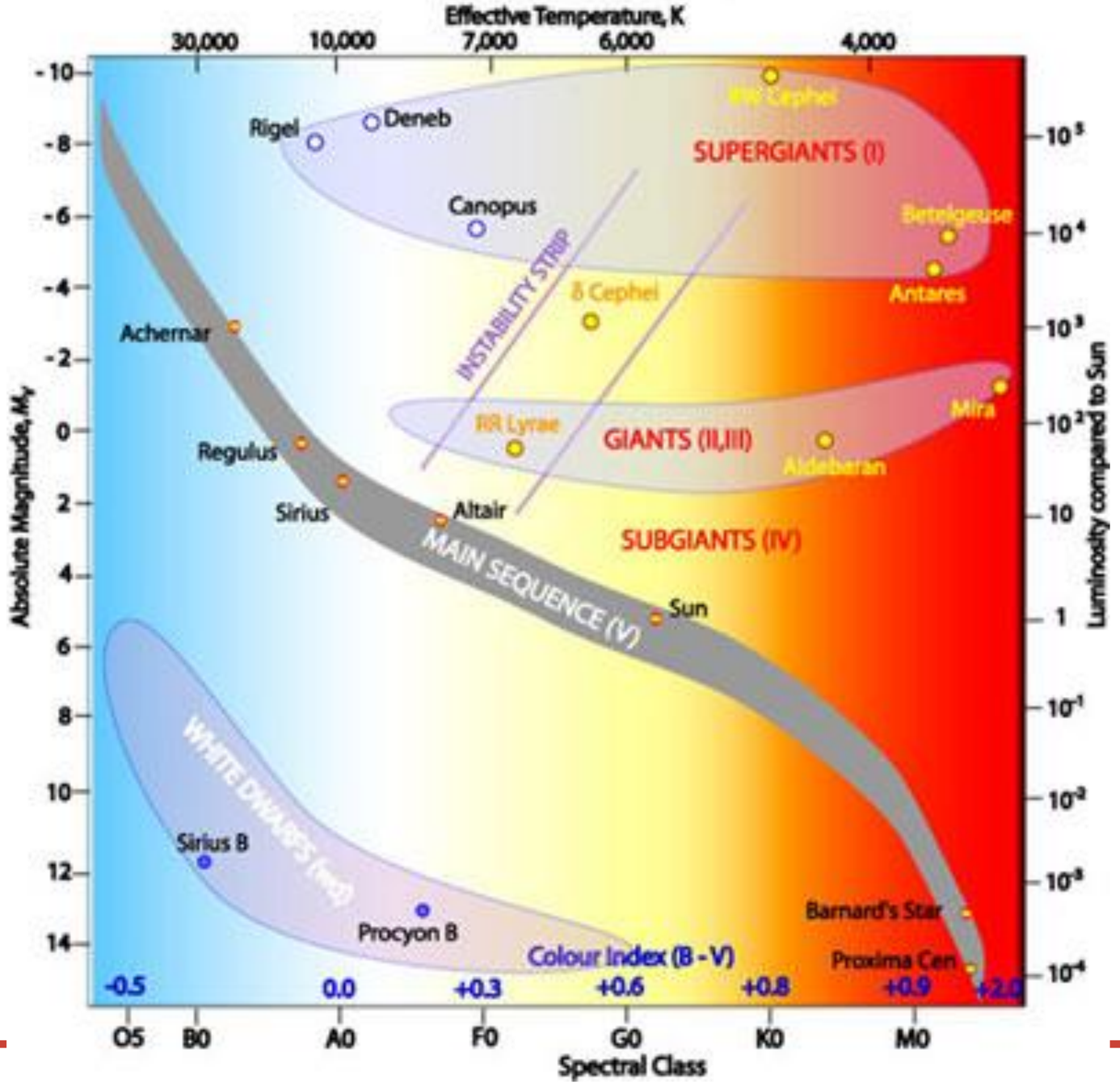
Cluster analysis is used when we think that, based on the involved variables objects form different groups.

When we use CA, we only suspect that some objects are more like each other than others.

If q=2 , a graphic rappresentation is possible






Hertzsprung-Russell Diagram
Effective Temperature, K
30,000
10,000
7,000
6,000
4,000
Absolute Magnitude, M_V
Luminosity compared to Sun
Rigel
Deneb
Canopus
SUPERGIANTS (I)
Betelgeuse
Antares
INSTABILITY STRIP
δ Cephei
Achernar
Mira
RR Lyrae
GIANTS (II,III)
Aldebaran
Regulus
Sirius
Altair
SUBGIANTS (IV)
MAIN SEQUENCE (V)
Sun
WHITE DWARFS (VII)
Sirius B
Procyon B
Barnard's Star
Proxima Cen
Colour Index (B - V)
-0.5
0.0
+0.3
+0.6
+0.8
+0.9
O5
B0
A0
F0
G0
K0
M0
Spectral Class

AIM:

Identify, if they exist, natural clusters, i.e. truly distinct groups

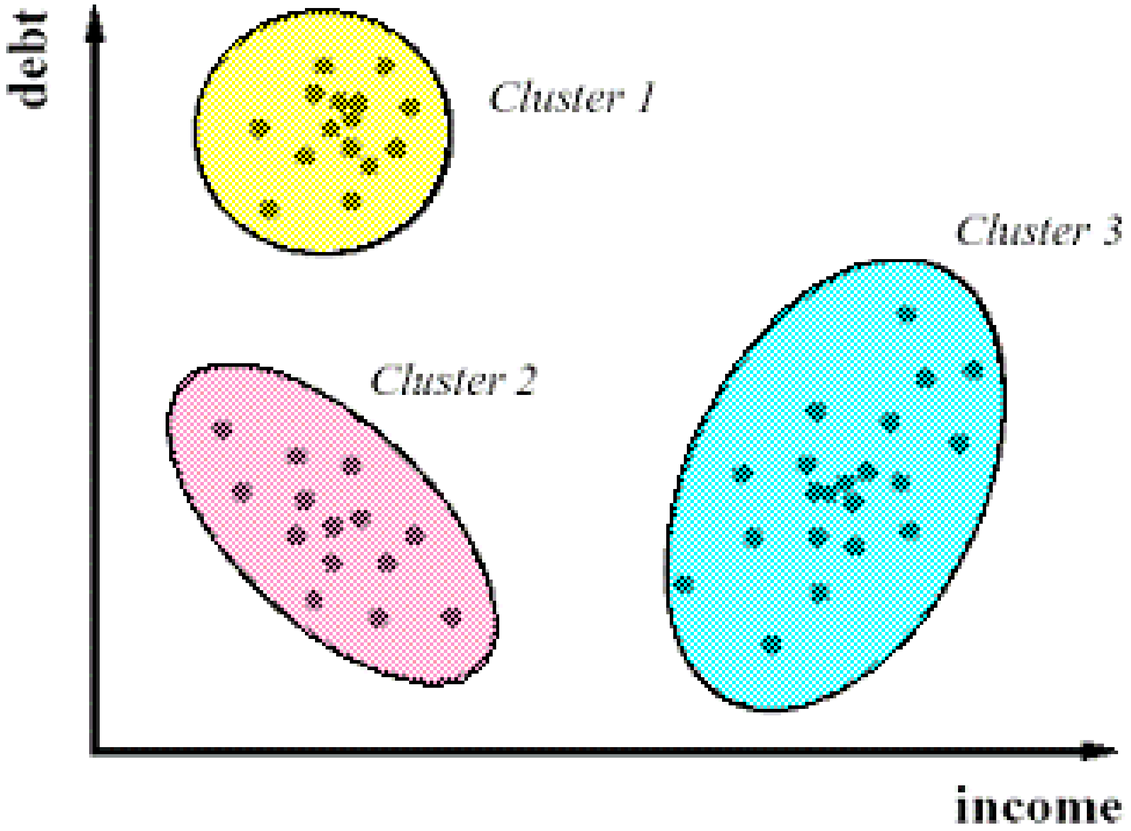
Cluster analysis poses a very difficult task:

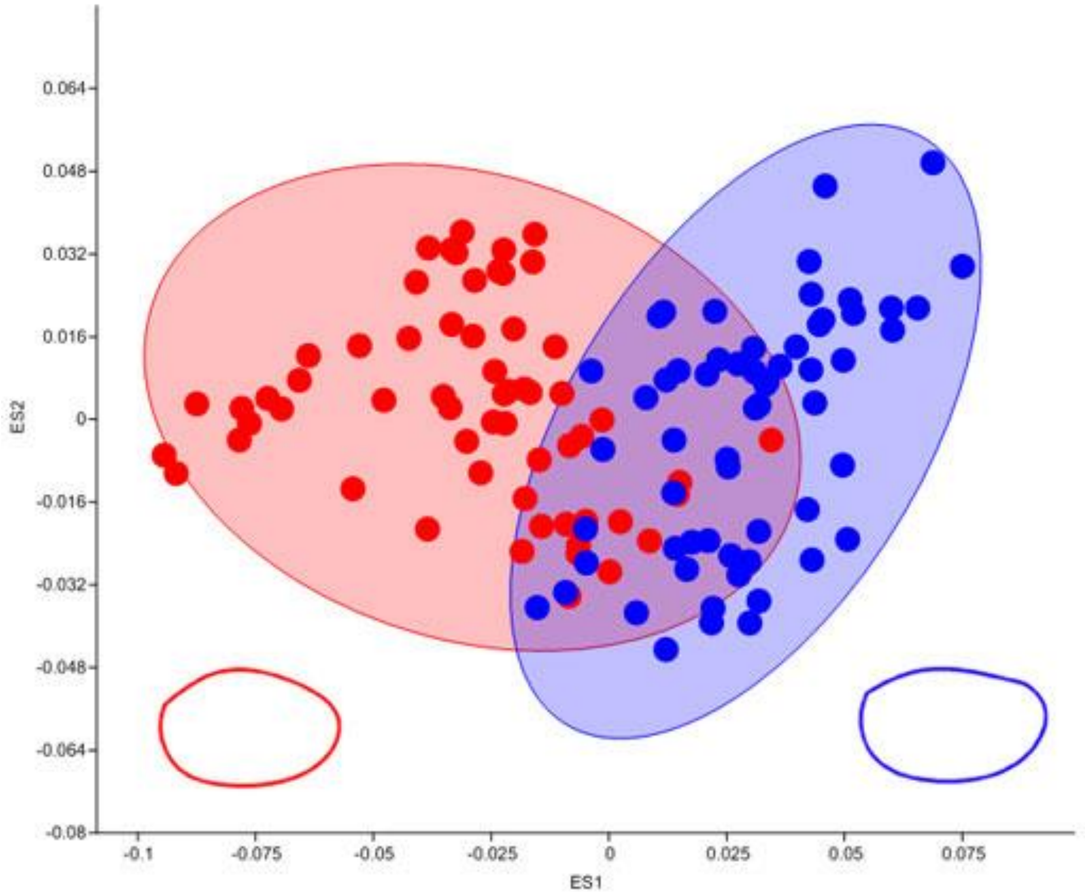
We want to discover whether distinct groups exist,

Using characters that we don’t know if they are discriminating

The choice of which variables should be included in a cluster analysis is a very crucial point

If q=2 , a graphic rappresentation is possible

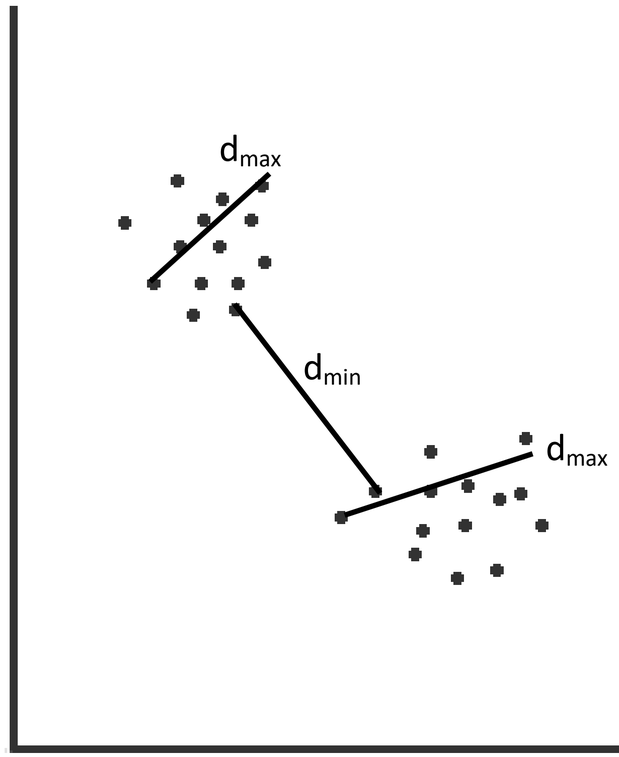




If q > 2, only an analytical representation is possible

THE RATIONALE OF CLUSTER ANALYSIS:

- **minimize the ‘dissimilarity’ or distance within each group**
- **massimize ‘dissimilarity’ or distance between different groups**



Several proximity measures exsist. They are known as METRICS

However , they must all respect the following rules:

1) $d(\mathbf{x}_i, \mathbf{x}_k) = 0$ **if and only if** $i = k$

2) $d(\mathbf{x}_i, \mathbf{x}_k) > 0$ **If** $i \neq k$

3) $d(\mathbf{x}_i, \mathbf{x}_k) = d(\mathbf{x}_k, \mathbf{x}_i)$

4) $d(\mathbf{x}_i, \mathbf{x}_k) + d(\mathbf{x}_l, \mathbf{x}_k) \geq d(\mathbf{x}_i, \mathbf{x}_l)$ se $i \neq k$, $k \neq l$, $l \neq i$;

A metric which follows those rules is the DISTANCE

$$d_p(\mathrm{x}_i, \mathrm{x}_k) = \sqrt[p]{\sum_{ij}^{q} \left| x_{ij} - x_{kj} \right|^p}$$

If p=1 we have the MANHATTAN distance

$$d_1(\mathrm{x}_i, \mathrm{x}_k) = \sum_{ij}^{q} \left| x_{ij} - x_{kj} \right|$$

If p=2 we have the EUCLIDEAN distance

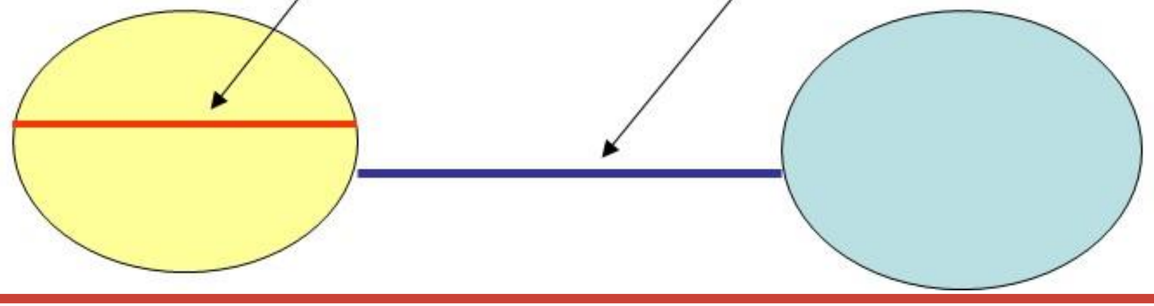
$$d_2(\mathrm{x}_i, \mathrm{x}_k) = \sqrt{\sum_{ij}^{q} \left(x_{ij} - x_{kj} \right)^2}$$

Suppose we have a set A divided in A_1, ……, A_n subsets

1) $A_1 + A_2 + ... + A_r = A$ 2) $A_i \cap A_j = 0$

Let's $d(x_i, x_k)$ a METRIC between the *i* and *k* objects

$$\max_{i,h}\left[d(x_i, x_h)\right] < \min_{i,k}\left[d(x_i, x_k)\right]$$



In practice, the starting point is the matrix of the distances D_{nxn}

$$D = \begin{bmatrix} d(x_1,x_1) & d(x_1,x_2) & . & d(x_1,x_n) \\ d(x_2,x_1) & d(x_2,x_2) & . & d(x_2,x_n) \\ . & . & . & . \\ d(x_n,x_1) & d(x_n,x_2) & . & d(x_n,x_n) \end{bmatrix}$$

$$D = \begin{bmatrix} 0 & d(x_1,x_2) & . & d(x_1,x_n) \\ d(x_2,x_1) & 0 & . & d(x_2,x_n) \\ . & . & . & . \\ d(x_n,x_1) & d(x_n,x_2) & . & 0 \end{bmatrix}$$

$$\mathbf{D} = \begin{bmatrix} 0 & d(x_1,x_2) & . & d(x_1,x_n) \\ & 0 & . & d(x_2,x_n) \\ . & . & . & . \\ & & . & 0 \end{bmatrix}$$

The number of distances to calculate is

$$\frac{n(n-1)}{2}$$

Hierarchical clustering

The starting point is the matrix of the distances (nxn) $\mathbf{D} = \begin{bmatrix} 0 & d(x_1, x_2) & . & d(x_1, x_n) \\ & 0 & . & d(x_2, x_n) \\ . & . & . & . \\ & & . & 0 \end{bmatrix}$

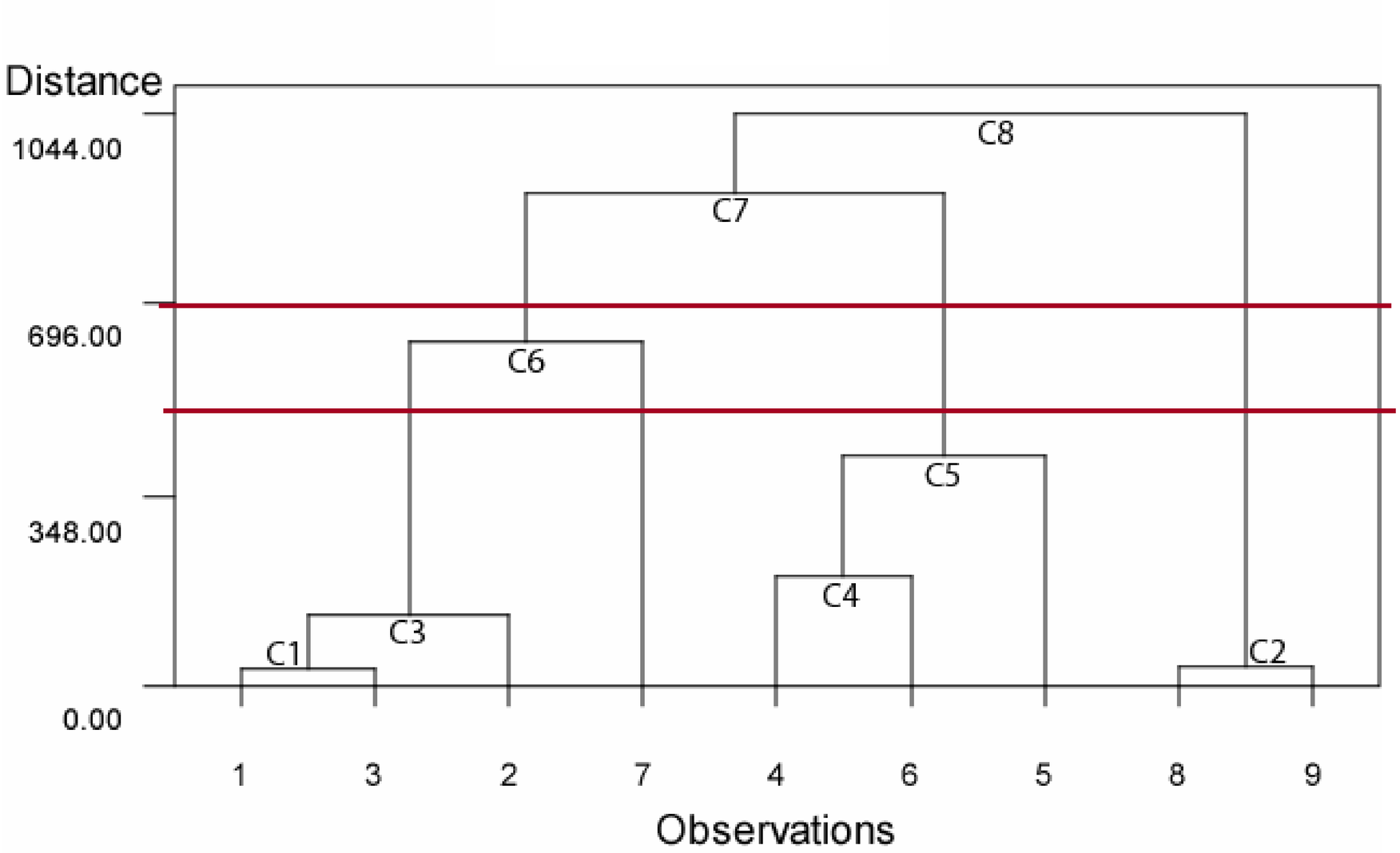
Step 1: n clusters

Step 2: the 2 data points with the lower distance are joined to form 1 cluster (n-1 clusters remain)

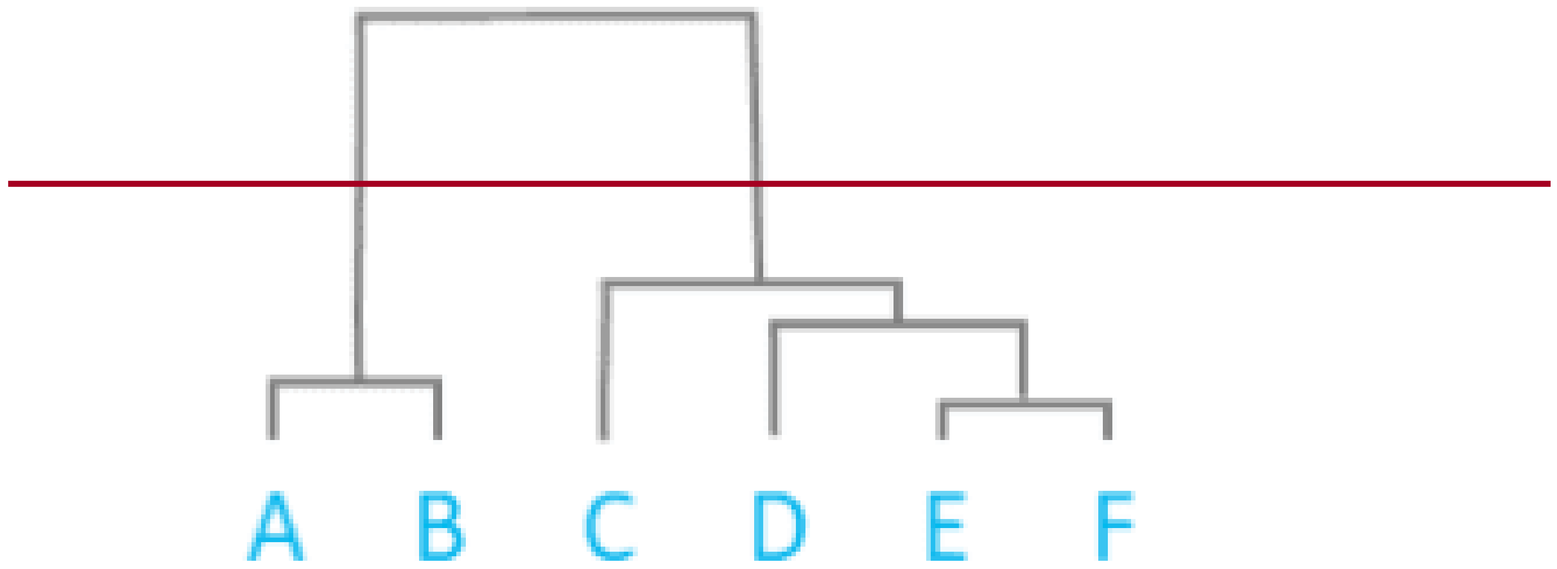
Step 3: distances are recalculated

Step 4: as in step2, the 2 clusters with the smallest distance are joined and n-2 clusters remain

The dendrogram



Dendrogram



Unlike DA, CA contains numerous sources of uncertainty.

1) How we calculate the distance between objects

Euclidean, Manhattan, Pearson, Mahalanobis and so on

2) How we chose the number of clusters in the dendrogram

3) The method of linkage we use

Linkage methods

Single linkage

$$d(x_g, x_t) = \min_{i,k}\left[d(x_i, x_k)\right]$$

Complete linkage

$$d(x_g, x_t) = \max_{i,k}\left[d(x_i, x_k)\right]$$

Average linkage

$$d(x_g, x_t) = \frac{1}{n_t n_g}\sum_{i}^{n_g}\sum_{k}^{n_t} d(x_i, x_k)$$

Centroids

$$d(x_g, x_t) = d(\bar{x}_g, \bar{x}_t)$$

Single linkage

$$d(x_g, x_t) = \min_{i,k}\left[d(x_i, x_k)\right]$$

$$\mathbf{D}_1 = \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{pmatrix} 0.0 & & & & \\ 2.0 & 0.0 & & & \\ 6.0 & 5.0 & 0.0 & & \\ 10.0 & 9.0 & 4.0 & 0.0 & \\ 9.0 & 8.0 & 5.0 & 3.0 & 0.0 \end{pmatrix}.$$

Cluster 1= 1,2

$$d_{(12)3} = \min(d_{13}, d_{23}) = d_{23} = 5.0$$

$$d_{(12)4} = \min(d_{14}, d_{24}) = d_{24} = 9.0$$

$$d_{(12)5} = \min(d_{15}, d_{25}) = d_{25} = 8.0.$$

$$\mathbf{D}_2 = \begin{array}{c} (12) \\ 3 \\ 4 \\ 5 \end{array} \begin{pmatrix} 0.0 & & & \\ 5.0 & 0.0 & & \\ 9.0 & 4.0 & 0.0 & \\ 8.0 & 5.0 & 3.0 & 0.0 \end{pmatrix}.$$

Non-hierarchical methods:

Basic idea

The number of clusters is specified in advance, *k* clusters

The *n*-objects are assigned to the *k*-clusters previously considered

Objects are then moved from one cluster to another in order to obtain the best subdivision

Several algorithms are at disposal.

The k-means algorithm

1) The number of clusters, *k*, each with one object, is fixed

2) The *n-k* objects are assigned to the nearest cluster basing on the distance from the centroids of the clusters

3) Distances from each object to each centroid are re-calculated and observations that are not in the cluster that they are closest are moved

4) Step 3) is repeated till distances of all objects from their own clusters are lower than distances from the other clusters

5) At each run, cluster centroids are recalculated.

Non-hierarchical methods

Are often used when large data sets are involved

It is sometimes preferred because it allows subjects to move from one cluster to another

Two disadvantages of non-hierarchical cluster analysis are:

You need to know the number of clusters in advance

Hierarchical clustering can be very sensitive to the choice of initial cluster centres