Ipsos Strategic Marketing

CLUSTER ANALYIS

an introduction for marketing, media and public affairs analyst

FOCUS
FOOD CONSUMER SCIENCE IN THE BALKANS



Concept

- Cluster analysis is as set of tehniques and algorithms aiming to group observations into homogenous classes - *clusters, and observations should be*
  - HOMOGENOUS **within clusters**: observations in the same cluster should be *similar* (*not dissimilar*)
  - HETEROGENOUS **between clusters**: observations from different clusters should be *quite distinct* (*dissimilar*)
- This means that we are interested in determining groups of observations internally characterized by a high level of cohesion.

Data – the way they should look

- We are looking at a set of n objects (observations)

  $S=\{e_1, e_2, \ldots e_n\}$

  on which we measure values of p variables/indicators:

  $X = (X_1, X_2, \ldots X_p)$

  resulting in a matrix X: $\mathbf{X} = [x_{ij}]_{nxp}$

- This is, in fact, a set of data organized in n rows and p columns. The element x_{ij} shows the value of indicator X_j for the object e_i.
- Data for cluster analysis don't have to comply with all strict conditions for statistical testing. They only have to be defined on a numerical scale (rational, ordinal or even dichotomous)

→ **CLUSTER ANALYSIS IS EASILY ADMINISTERED, PRIMARILY AS A DESCRIPTIVE TECHNIQUE**

Origins

- In the textbooks cluster analysis appears under a number of names, but since the 90ies the term CLUSTER dominates. Previously, it went under names:
  - Taxonomy or taxometry
  - Numerical or automatic classification
  - Botrology
  - Cluster Analysis

- Until the 80ies the development of this technique was slow, the origin stemming from different areas (biology, anthropology, economics, political sciences, language studies, etc.)

Cluster vs Discra

- In cluster analysis classes are IDENTIFIED, while in discriminative analysis ("Discra") class borders are being defined
- The answer to the coloquial question "how is something to be classified" is provided by both techniques:
  - Cluster analysis classifies the S set members (observations) into classes that are mutually similar based on X variables
  - Discriminative analysis starts from the *apriori* known class membership trying to find out the best distinction between the known classes. This distinction is defined by a function of vector X – this function applied to the values $X(e_j)$ in order to recognise which class e_i belongs to.

Clustering of clustering metods

- In cluster analysis there are several basic families of algorythms
- Hierarchical – Nonhierarchical
  - Hierarchical
    - Divisive...
    - Aglomerative
  - Nonhierarchical
    - K-min / realocational ...
    - Partition – by S i X simultaneously....
- With relational constraints..
- Grouping of OBJECTS or VARIABLES or BOTH AT THE SAME TIME
- Fuzzy clustering ...

Clustering objects

- Objects that are closer together based on pairwise multivariate distances or pairwise correlations are assigned to the same cluster, whereas those further apart or having low pairwise correlations are assigned to different clusters.

Object	Variable 1	Variable 2	Variable *p*
1	X_{11}	X_{21}	X_{p1}
2	X_{12}	X_{22}	X_{p2}
⋮	⋮	⋮	⋮
N	X_{1n}	X_{2n}	X_{pn}

Clustering variables

- Variables that have high pairwise correlations are assigned to the same cluster, whereas those having low pairwise correlations are assigned to different clusters.

Object	Variable 1	Variable 2	Variable *p*
1	X_{11}	X_{21}	X_{p1}
2	X_{12}	X_{22}	X_{p2}
⋮	⋮	⋮	⋮
N	X_{1n}	X_{2n}	X_{pn}

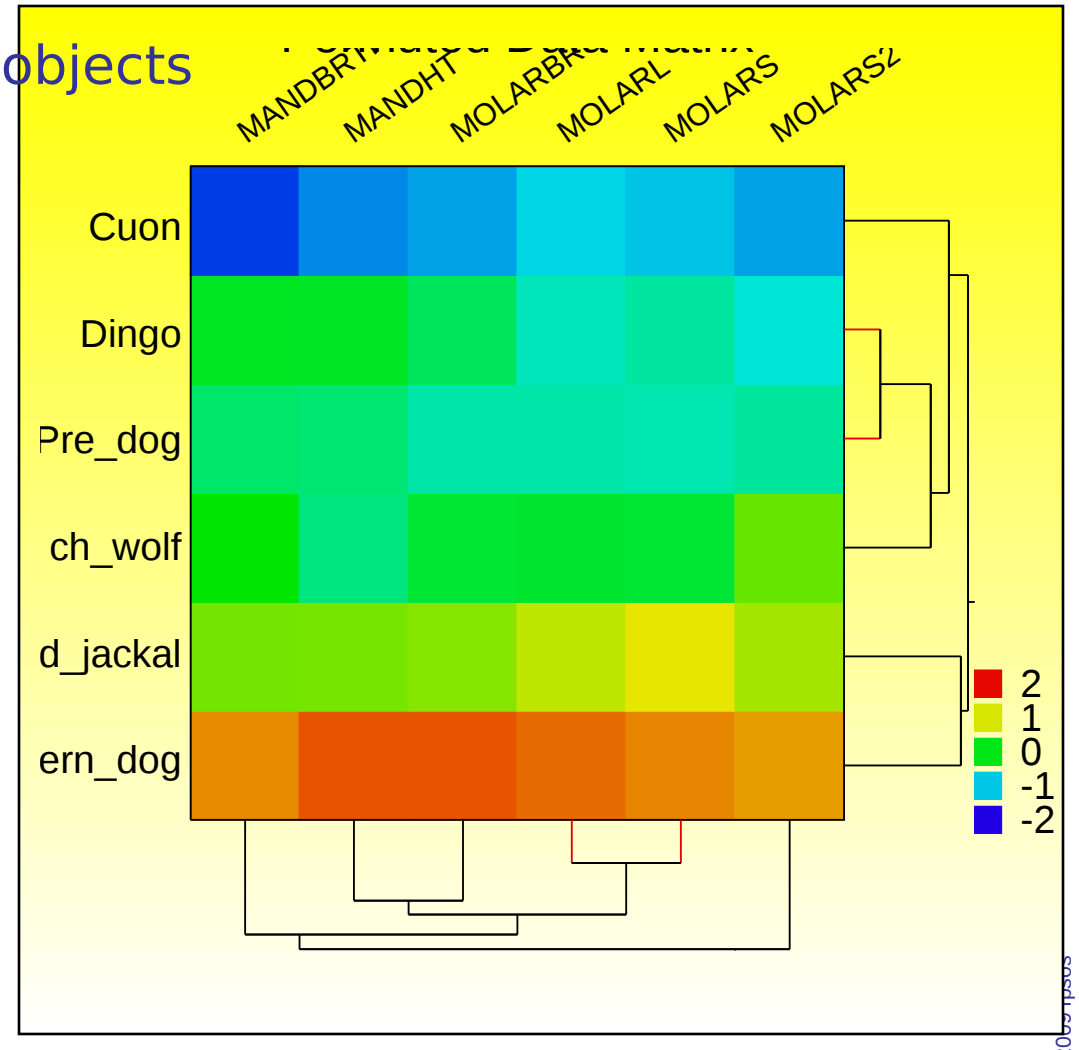
Clustering objects and variables

- Object/variable combinations are classified into discrete categories determined by the magnitude of the corresponding entries in the original data matrix
- Allows for easier visualization of object/variable clusters.

Object	Variable 1	Variable 2	Variable *p*
1	X_{11}	X_{21}	X_{p1}
2	X_{12}	X_{22}	X_{p2}
⋮	⋮	⋮	⋮
N	X_{1n}	X_{2n}	X_{pn}

Hierarchical clustering of objects and variables

- Standardized data matrix is used to produce a two-dimensional colour/shading graph with colour codes/shading intensities determined by the magnitude of the values in the original data matrix…
- …which allows one to pick out “similar” objects and variables at a glance.



Focusing on...

1. .... hierarchical clustering “bottom to top”, i.e. aglomerative h.c. – in which once the two objects are assigned to the same cluster, they remain together from that point onwards.

2. ... nonhierarchical clustering – in which objects, once assigned to a cluster, may leave that cluster to join another one. The center of a cluster (centroid) is being chosen first and all objects that lie within a certain distance (threshold) are assigned to that cluster....

Distance measures

Cluster analysis is typically based on *dissimilarity* between observations and, also, between groups (of observations).

In the context of cluster analysis, any measure of the dissimilarity (distance measure) between two observations, say the *i*-th and the *k*-th, satisfies the following:

$$d_{i,k} \geq 0 \qquad \text{for all } i,k$$

$$d_{i,i} = 0$$

$$d_{i,k} = d_{k,i}$$

$d_{ik} = 0$ does not mean that two cases are identical. This only means that they are not dissimilar *with respect to the particular context under analysis*

Measures of dissimilarity and similarity (1)

- Are used to group similar objects into the same cluster, and dissimilar objects into different clusters
- Often used measures are:
  1. Distance measures,
  2. Corelation coefficients, and
  3. Coefficients of association (contingencies)i.
- **Distance measures**: the most popular is Eucledian distance, d_{ij}:

$$d_{ij}^2 = \sum_{m=1}^{p} (x_{im} - x_{jm})^2$$

where x_{im} and x_{jm} are standardised values of m-th atribute of objects i and i

Measures of dissimilarity and similarity (2)

- The main shortcoming is that it is applied after all variables are standardised (for comparison, units of measurement)
- But this procedure also leads to lower visibility of differentiation between clusters and variables
- Outliers should be removed...

- There is a set of other measured ... e.g. *Manhattan metric*

$$d_{ij} = \sum_{m=1}^{p} \left| x_{im} - x_{jm} \right|$$

- **or..** Minkowski metric, I-distance, weighted distance ...

Measures of dissimilarity and similarity (3)

- **Correlation coefficients**
  - Major problem is their sensitivity to the pattern of ups and downs across the variables at the expense of the magnitude of difference between the variables.
- **Association coefficients**
  - Are used to establish similarity between objects when binary variables are used.

$$d_{ij} = \mathbf{D} = \begin{bmatrix} 0 & d_{12} & d_{13} \\ d_{21} & 0 & d_{23} \\ d_{31} & d_{32} & 0 \end{bmatrix}$$

$$\mathbf{r} = \begin{bmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{bmatrix}$$

Some clustering distances

Distance metric	Description	Data type
Gamma	Computed using $1 - \gamma$ correlation	Ordinal, rank order
Pearson	1- *r* for each pair of objects	quantitative
R^2	$1 - r^2$ for each pair of objects	quantitative
Euclidean	Normalized Euclidean distance	quantitative
Minkowski	*p*th root of mean *p*th powered distance	quantitative
χ^2	χ^2 measure of independence of rows and columns on 2 X *N* frequency tables	counts
MW	Increment in SS_{within} if object moved into a particular cluster	quantitative

Scale considerations

- In general, correlation measures are not influenced by differences in scale, but distance measures (e.g. Euclidean distance) are affected.
- So, use distance measures when variables are measured on common scales, or compute distance measures based on standardized values when variables are not on the same scale.

Dissimilarity matrix

Dissimilarities between observations are arranged in a square ($n \times n$) matrix, where n is the number of observations.

Cases \ Cases				
	d_{11}	d_{12}	….	d_{1n}
	d_{21}	d_{22}	….	d_{2n}
	….	….	….	….
	d_{n1}	d_{n2}	….	d_{nn}

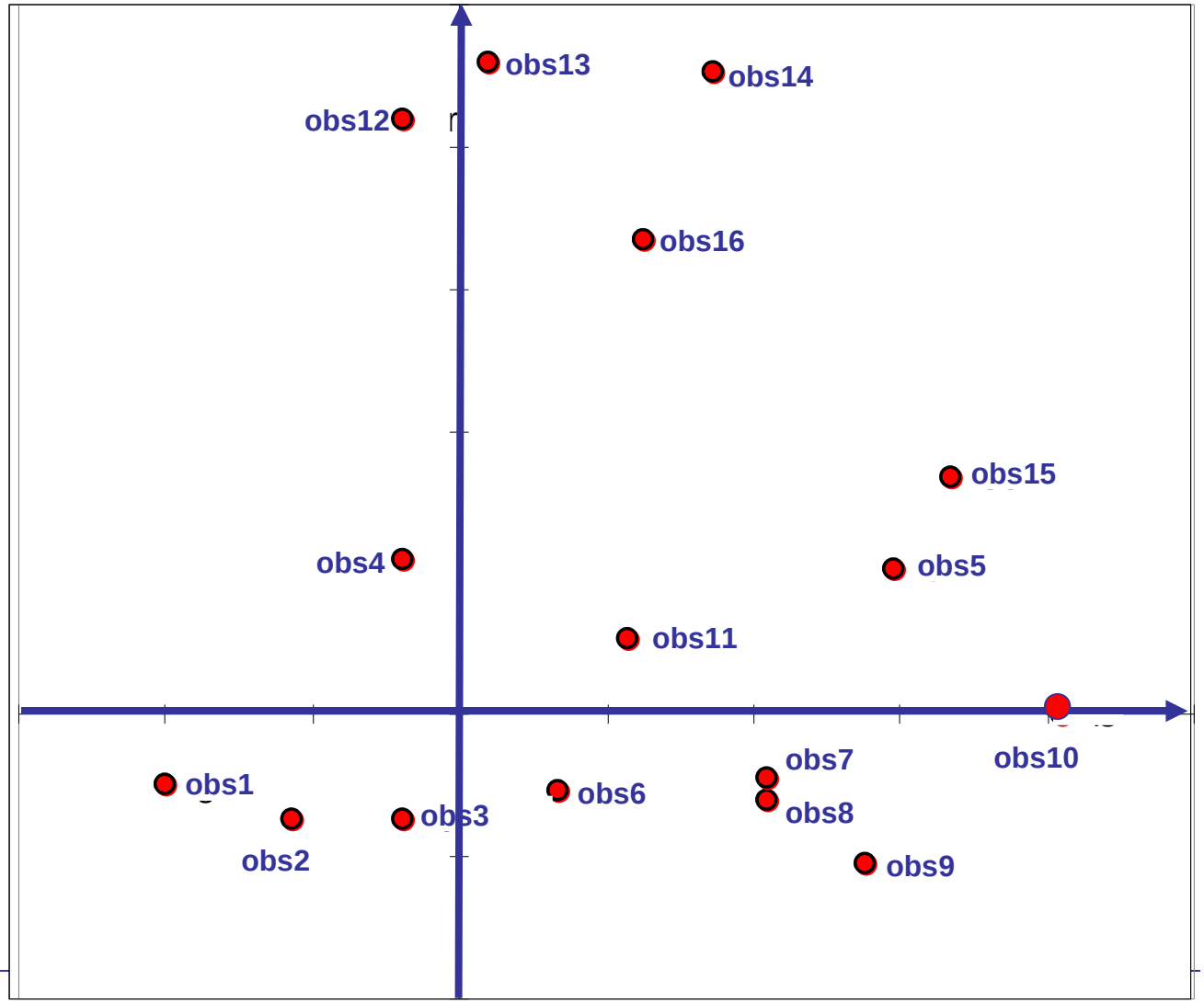
The (i,k)-th element of the matrix is the dissimilarity between the i-th and the k-th case. The matrix is symmetric, since we assumed that $d_{i,k} = d_{k,i}$

In some applications the dissimilarity matrix is obtained by taking into account **measurements on a set of variables**. Different measures of dissimilarities have been introduced in literature depending on the characteristics of the involved variables. Hence, **different dissimilarity matrices** can be obtained.

In other situations, the dissimilarity matrix may contain other kind of information, for example *judgements* about the dissimilarity between cases. In this case, the **dissimilarity matrix is given**.

An example (borrowed from the Internet)

Example (synthetic data). Simple example, 2 dimensions – graphical analysis



2 groups are clearly identifiable

But maybe also 3 groups may be considered. Which cluster should obs11 and obs6 be assigned to?

An example: hierarchy vs nonhierarchy

Hierarchical (agglomerative) algorithms

Sequential procedures.

At the first step, each observation constitutes a cluster. At each step, the two closest clusters are joined to form a new cluster. Thus, the groups at each step are nested with respect to the groups obtained at the previous step.

Once an object has been assigned to a group it is never *removed* from the group later on in the clustering process.

The hierarchical method produce a complete sequence of cluster solutions beginning with *n* clusters and ending with one clusters containing all the *n* observations.

In some application the set of nested clusters is the required solution whereas in other applications only one of the cluster solutions is selected as the solution, i.e., the proper number of clusters has to be selected.

Partitioning algorithms

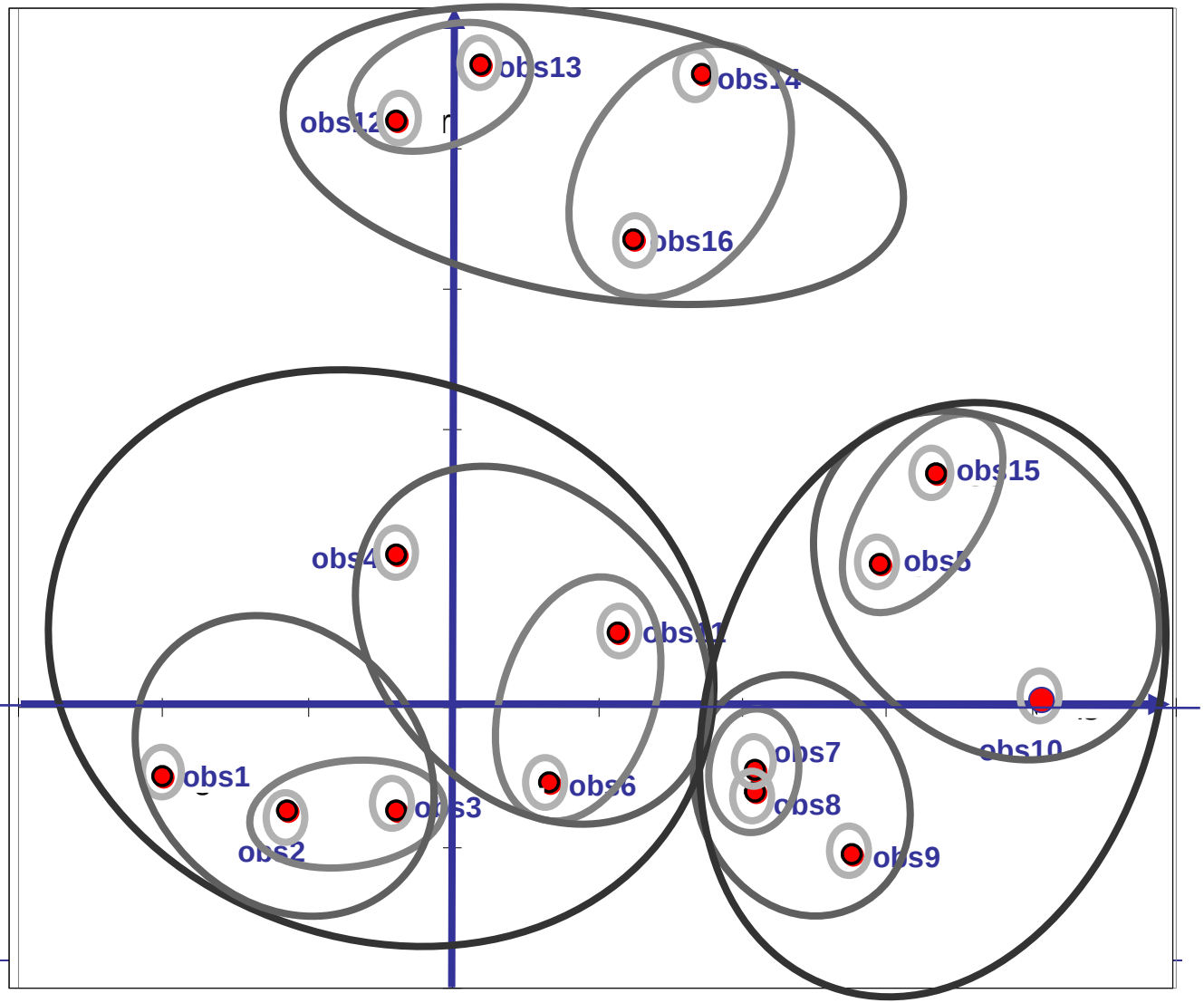
Iterative procedures

In these methods, the aim is to partition cases into a given number of clusters, say ***G***. The algorithm usually begins with an initial solution (partition). Observations are then reallocated to cluster so as to maximize a pre-specified objective function.

Objects possibly change cluster membership throughout the cluster formation process.

An example: hierarchical approach

Initial solution: ***n clusters*** (one for each observation)



At each step: the two closest (lowest dissimilarity) clusters are joined to form a new cluster

Hierarchical agglomerative algorithms

At each step, we should join the two closest clusters.

Our starting point is the dissimilarity matrix. It is almost easy to determine which are the two closest observations.

Nevertheless, now a problem arises: how do we calculate the dissimilarity between one observation and one cluster or between two clusters?

Definition of criteria to measure the

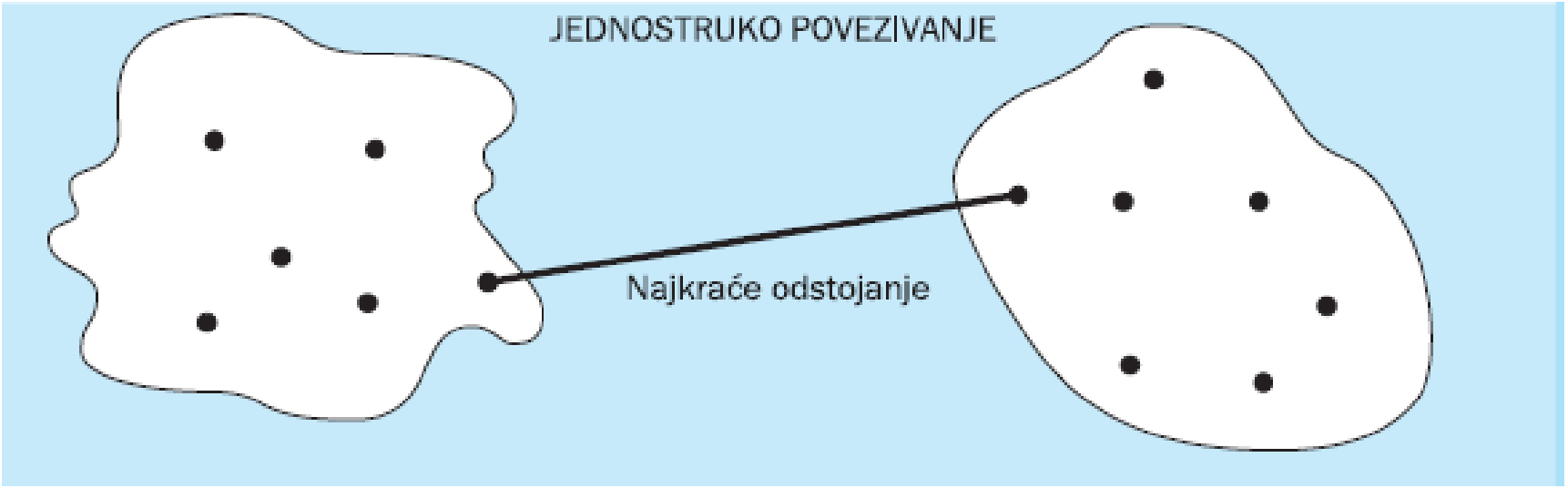
Dissimilarity between groups of observations (clusters)

Hierarchical clustering

1. Single linkage method
2. Complete linkage method
3. Average linkage method
4. Ward's method
5. Centroid method

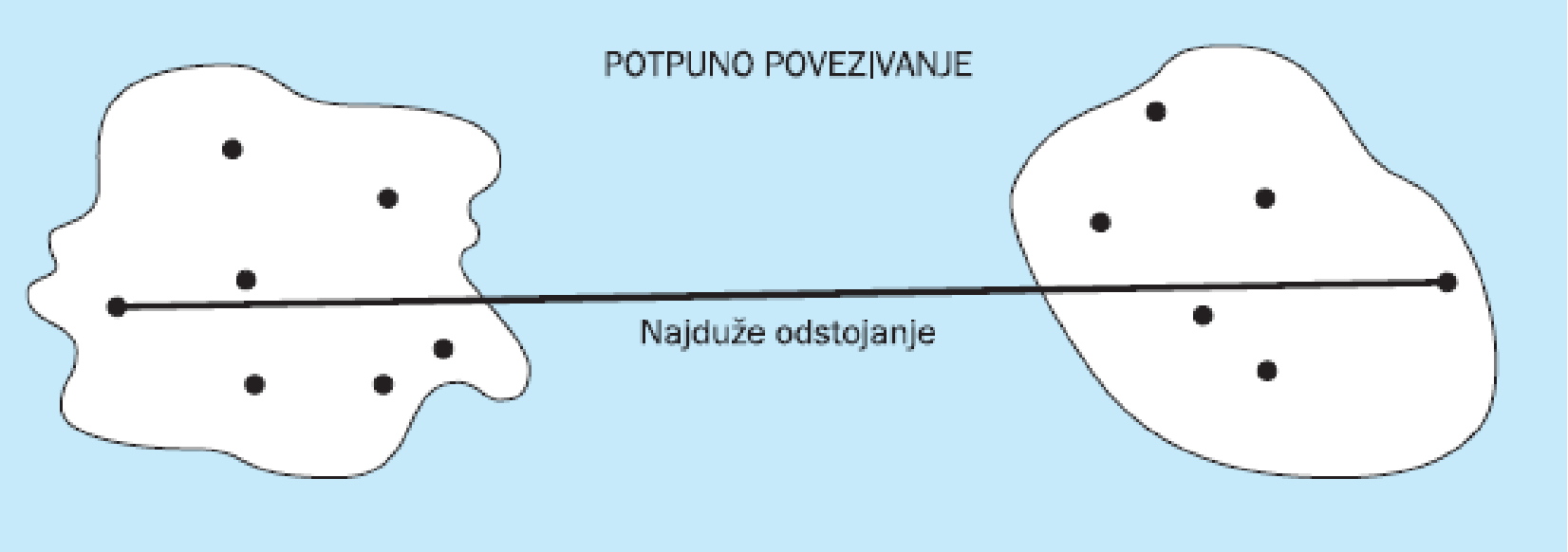
1. Single linkage

- Is based on the shortest distance and is also referred to as the *nearest-neighbour approach*
  - First, it finds the two individuals (objects) separated by the shortest distance and placese them in the first cluster (*minmin*)
  - The next shortest distance is found and either a third individual joins the first two to form a cluster or a new two-individual cluster is formed...



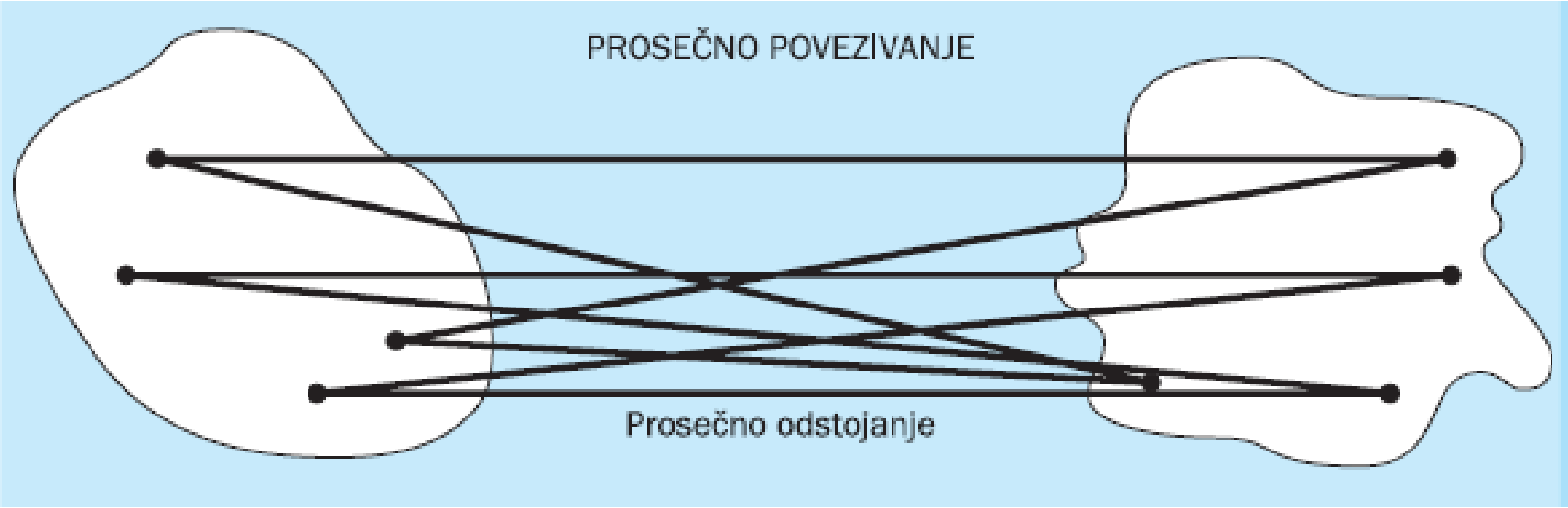
2. Complete linkage

- Is based on the longest distance between the objects an is referred to as the *farthest-neighbour approach*
  - Two objects that are at the longest distance are assigned to two separate clusters (*minmax*)



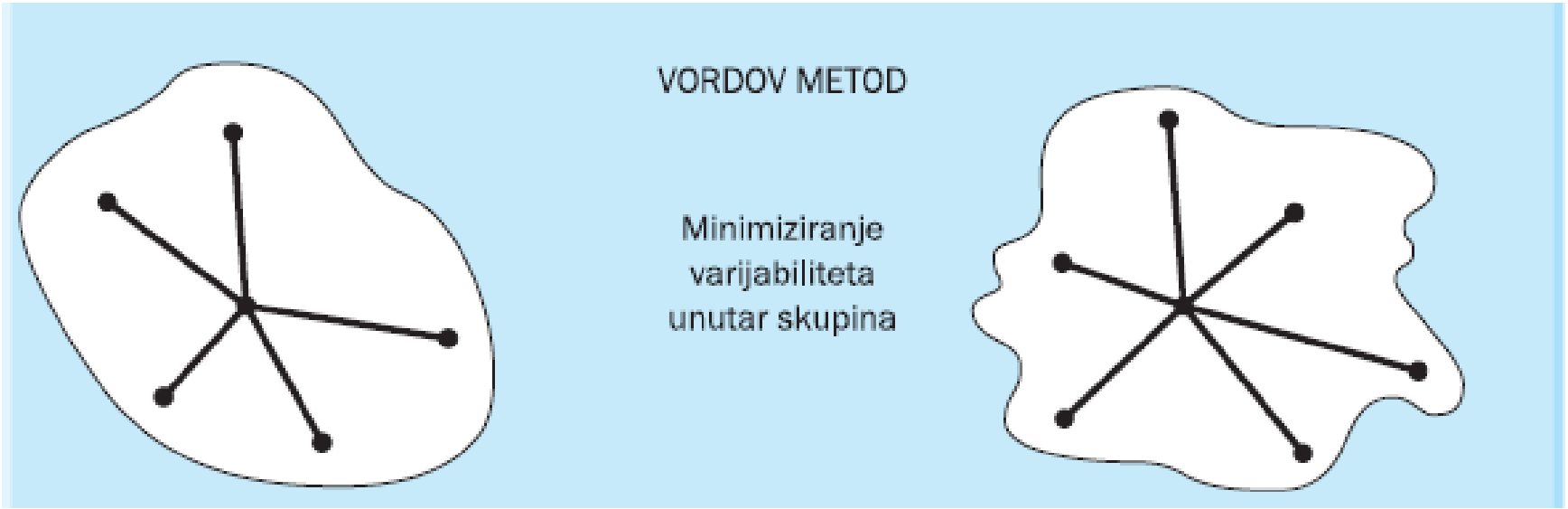
3. Average linkage

- The clustering criterion is the average distance from objects in one cluster to objects in another.
  - It is based on all members of the clusters rather than on a single pair of extreme member



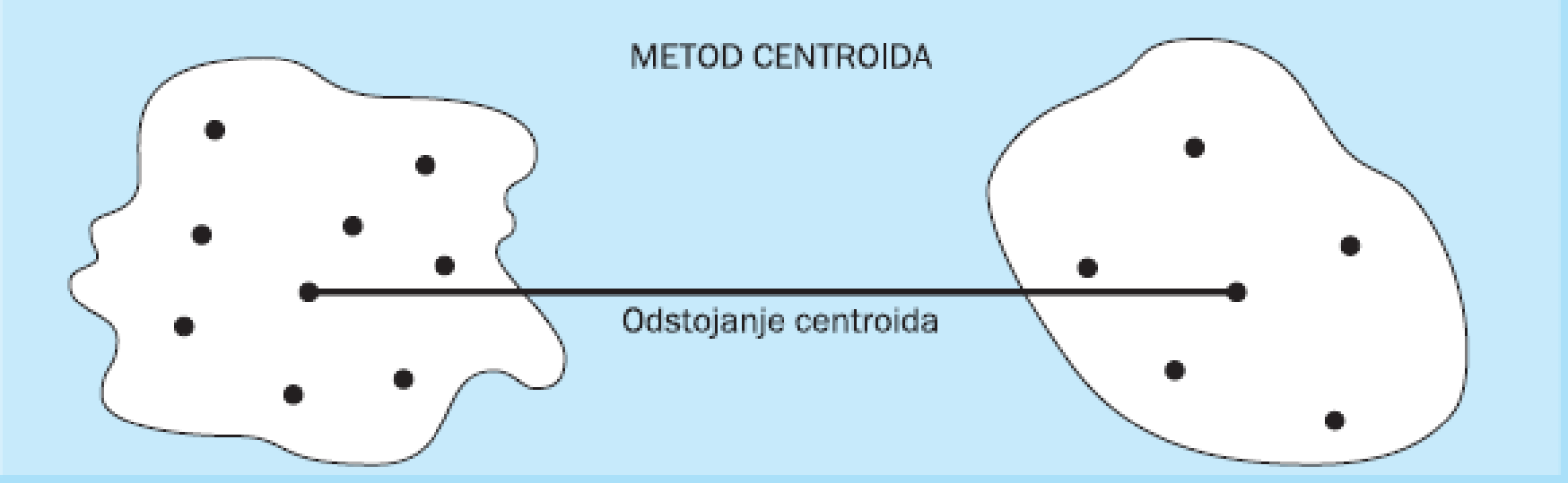
4. Ward's method (metod minimalne sume kvadrata)

- Uses measurement of total sum of squared deviations of every object from the mean of the cluster to which the object is assigned
  - At each stage the error sum of squares is minimised over all partitions (clusters) obtainable by combining two clusters from the previous stage



5. Centroid method

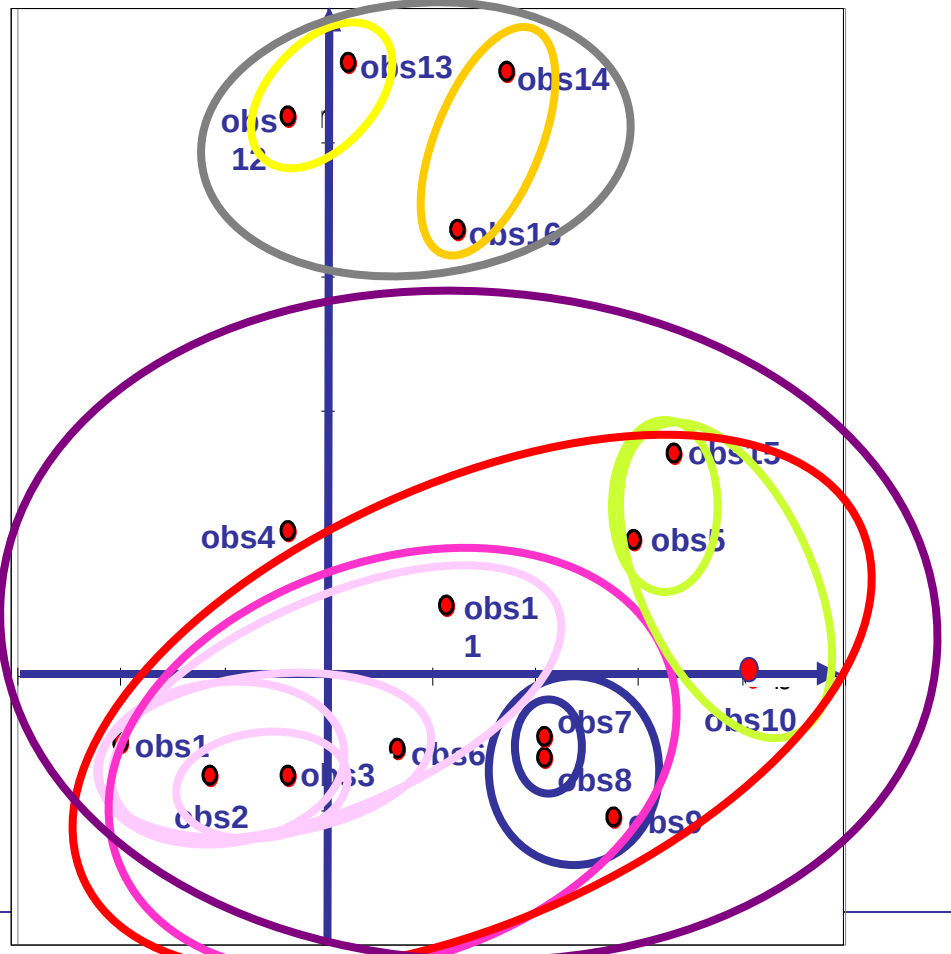
- Measures the distance between group (cluster) centroids
  - The process continues by combining the groups according to the distance between their centroids, the groups with the shortest distance being combined first



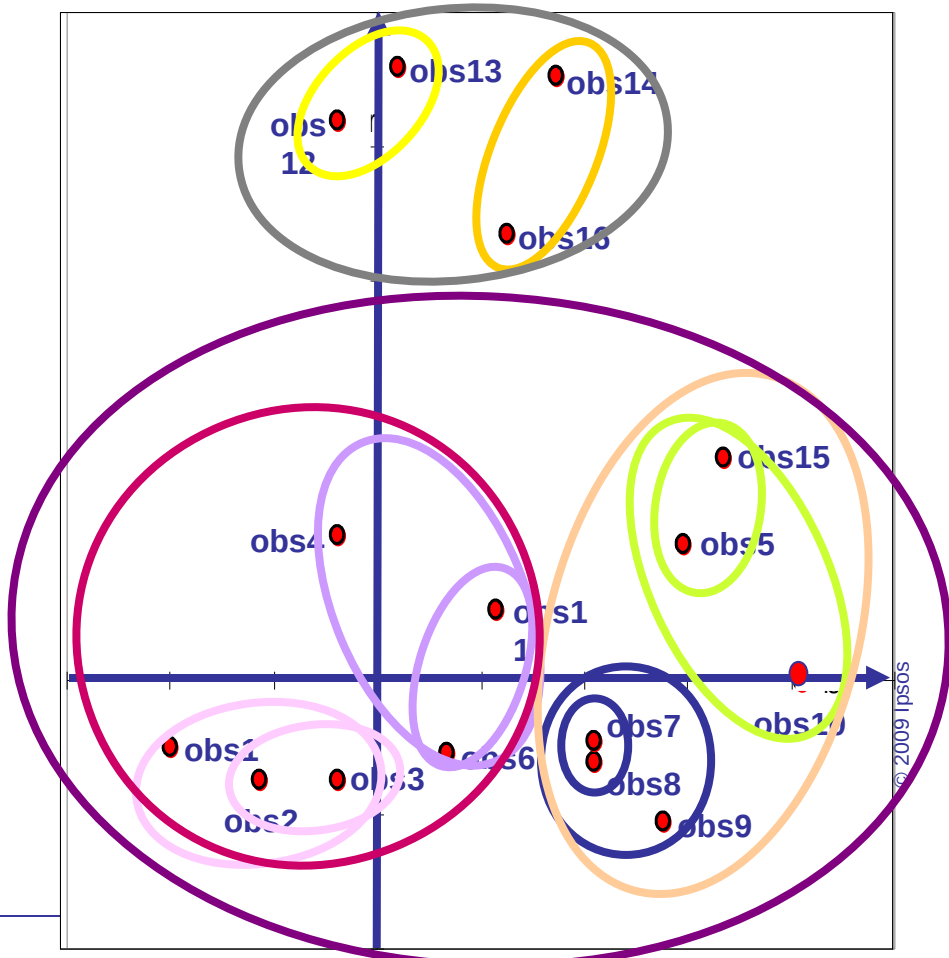
An example: two hierarchy algorythms

Given a dissimilarity matrix, based on a certain measure of the dissimilarity between cases, there are different methods to measure the dissimilarity between *clusters.* These criteria often lead to different partitions.

ingle Linkage Cluster Analysis

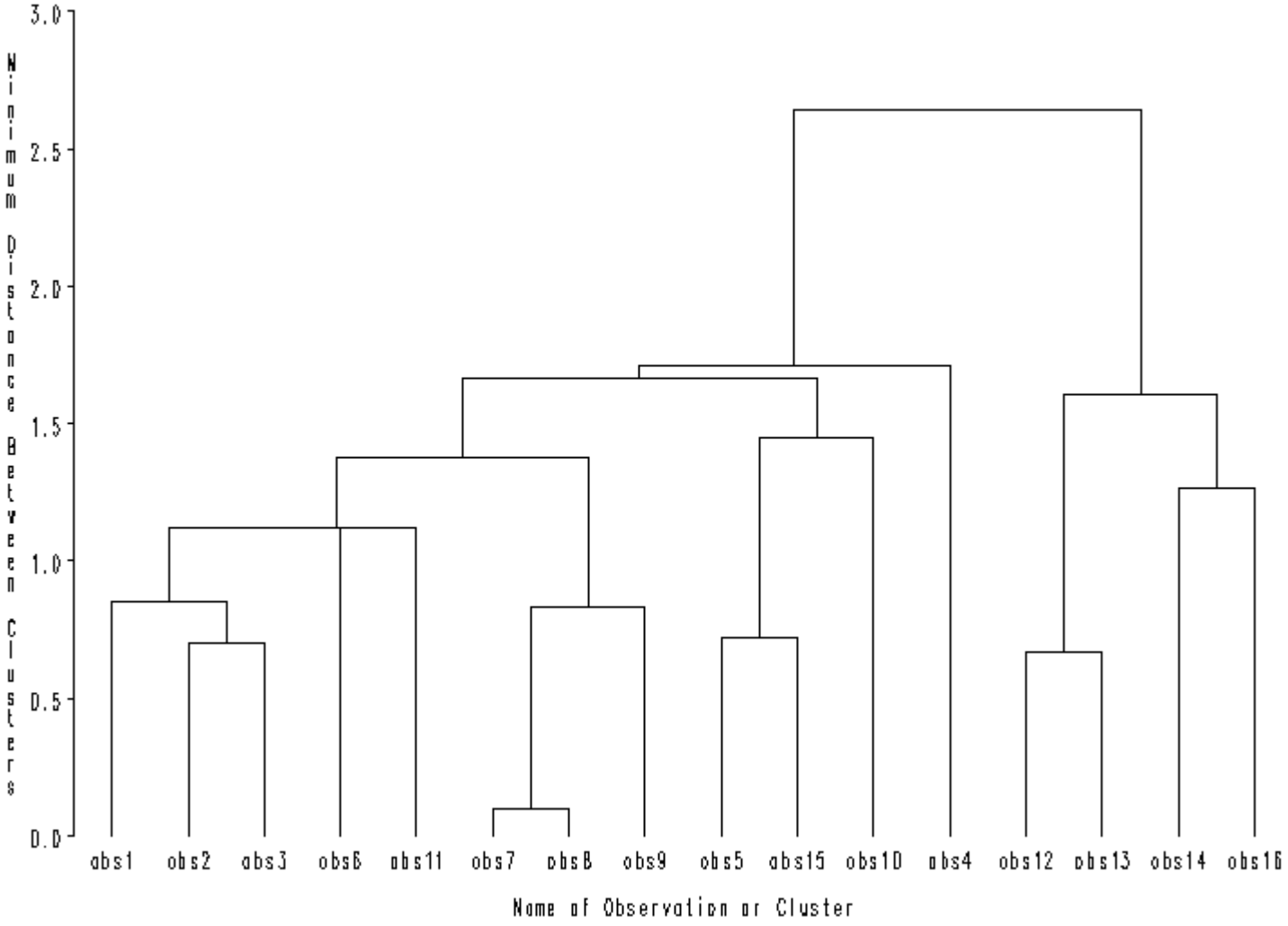


Complete Linkage Cluster Analysis

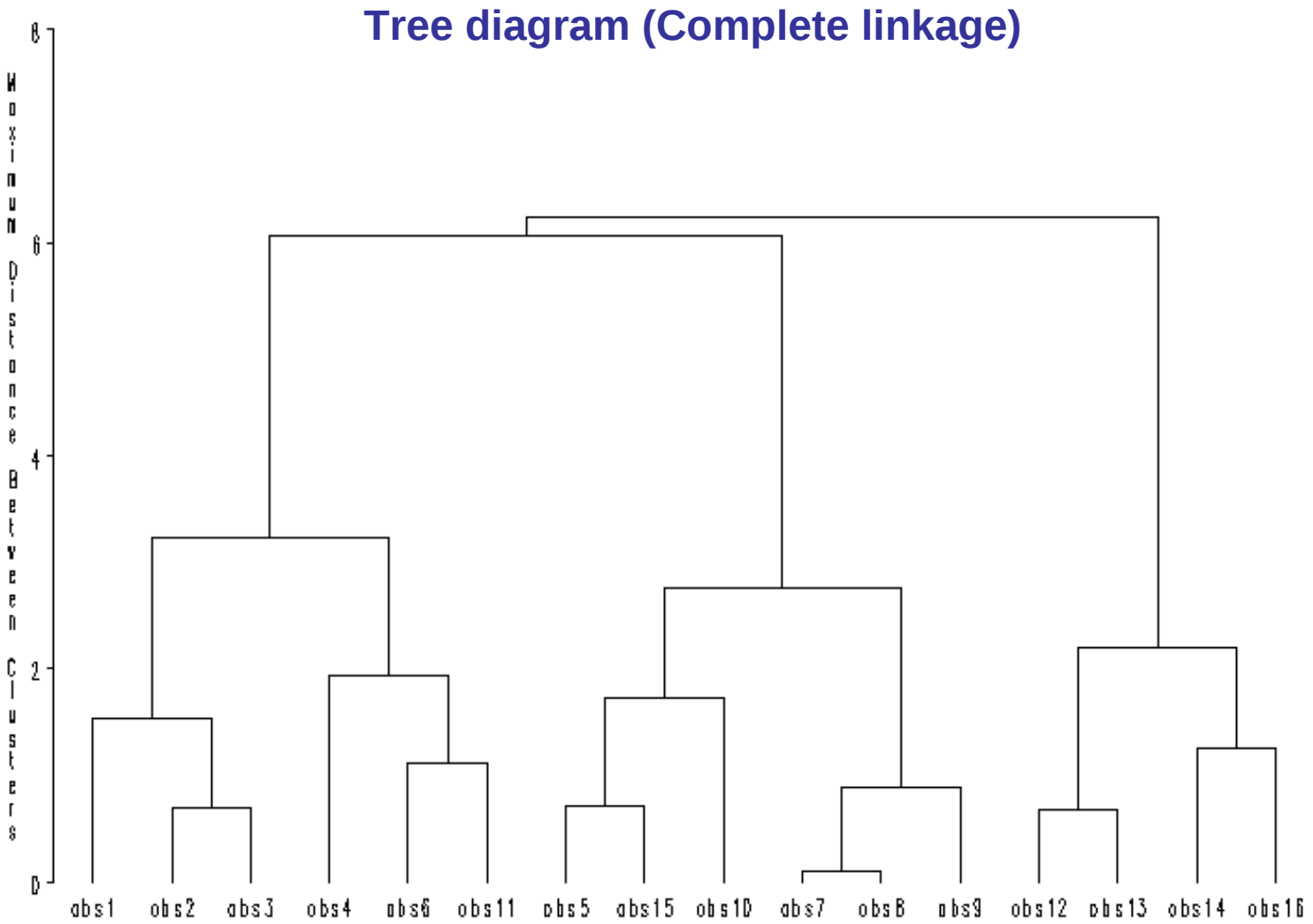


Cluster analysis: dendrogram

Tree diagram (Single linkage)



Cluster analysis: dendrogram



Cluster analysis: hierarchical algorithms

To apply a hierarchical agglomerative algorithm we have to:

1. **Obtain the dissimilarity matrix** containing the dissimilarities between all the possible pairs of observations **(**as we will see later, different criteria may be referred to)
2. **Choose a method to measure the dissimilarity between clusters**

These choices have an impact on the *sequence of nested partitions* obtained as an output. So we usually have **different sequences of nested partitions**.

But, also, for a <u>given sequence of nested partitions</u> the following problem arises:

How should we select a suitable number of clusters?

Cluster analysis: the number of clusters

We consider first the problem of **choosing one out of the clusters solutions** obtained with one hierarchical clustering process.

At this aim, the agglomeration process is monitored as the number of clusters declines from *n* to 1, and some quality of clustering criteria are evaluated.

1. **Internal criteria.** The simplest approach to cluster choice consists in the evaluation of the ***dissimilarity between the two clusters joined at each step.*** In the first steps of the procedure, similar cases/groups will be joined to form new clusters. At subsequent steps, we can expect an increasing of this dissimilarity, and this increase will tend to grow exponentially in the last aggregation phases, i.e. when very dissimilar clusters are joined.
2. **External criteria.** Another possibility consists in the evaluation of some statistics – not related to the criterion used to measure the dissimilarity between clusters – which are solely based upon the R^2, the within and the between sum of squares characterizing partition of different degree (different number of clusters)

Nonhierarchical clustering – partitioning algorithms

In partitioning algorithms, the number of clusters has to be specified. The algorithm usually starts with an initial allocation of the objects into *G* groups. Then observations are placed in the cluster they are closest to. Alternatively, observations are assigned to one cluster so as to maximize an objective function. The procedure iterates until all objects belong to the closest group (the objective function is maximized) or until a convergence criterion is satisfied .

Usually (SAS) partitioning methods are based upon measurements on a set of variables rather than on a dissimilarity, and on Euclidean distances.

One of the most important partitioning algorithms is the ***k-means* algorithm.** In this algorithm, **the distance from one observation to a cluster is measured as the distance between the observation and the centroid of the cluster**.

It can be easily shown that in this case the algorithms attempts to find the partition characterized by the minimum **Within SS**, i.e., by the maximum R^2.

In this sense, Ward's and the *k*-means algorithms are two R^2-maximizing algorithms. The former is based upon a hierarchical solution to the optimization problem. The latter is instead based on an iterative search of the optimum.

Nonhierarchical clustering (1)

- Also known as iterative partitioning
- Objects may leave one cluster and join another one, if that would be an improvement as defined by the clustering criterion

1. **Sequential threshold**
   - Cluster center is selected and all objects within a prespecified threshold value are grouped
   - Then a new cluster center is calculated and the procedure repeated for unclustered objects
   - Once objects enter a cluster they are removed from further procedure.

Nonhierarchical clustering (2)

1. **Paralel threshold**
   - Is similar to the preceeding method, except that several cluster centers are selected simultaneously and objects within the threshold level are assigned to the nearest center
   - The threshold level can then be adjusted to admit fewer or more objects to the cluster
2. **Optimising**
   - Modification of previous two procedures.
   - Objects may later be reassigned to another cluster by optimising some overall criterion measure (e.g. average within-cluster distance)

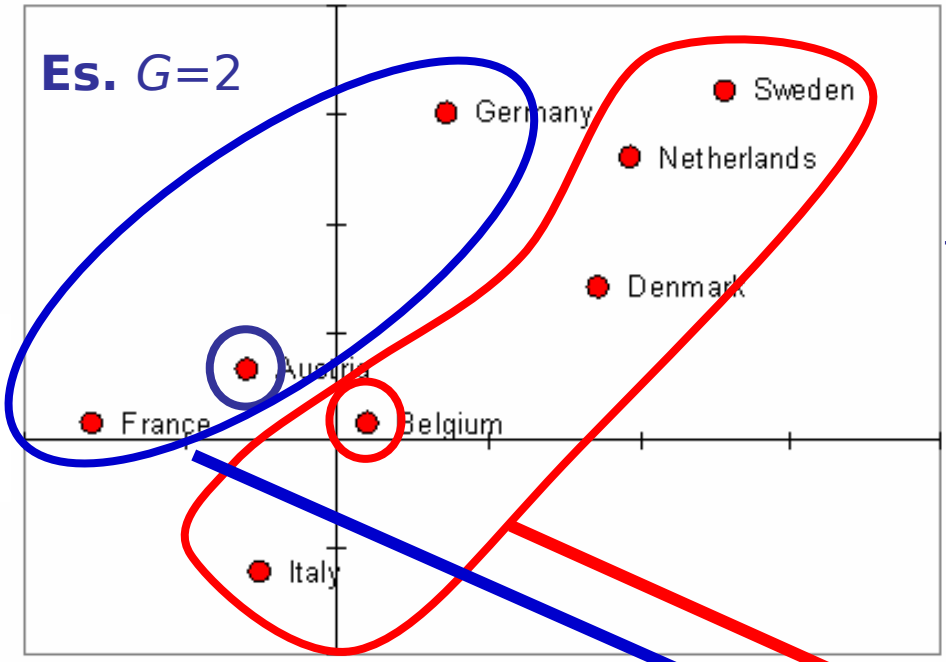
Determining the number of clusters

1. May be predefined (for theoretical, logical and practical reasons)
2. May be defined based on certain criteria for clustering
3. May be determined from the pattern of clusters the program calculates for various cluster numbers
4. The ratio of total within-group variance to between-group variance is plotted relative to the number of clusters. The point at which an elbow or a sharp bend occurs indicates the appropriate number of clusters.

Evaluating and profiling the clusters

- Describing the newly-formed clusters
- A frequently used measure is the centroid, especially if the data provided are defined on an interval scale and the clustering is performed in the original variables space
- The mean scores should help to describe or profile the clusters.

Cluster analysis – partitioning algorithms



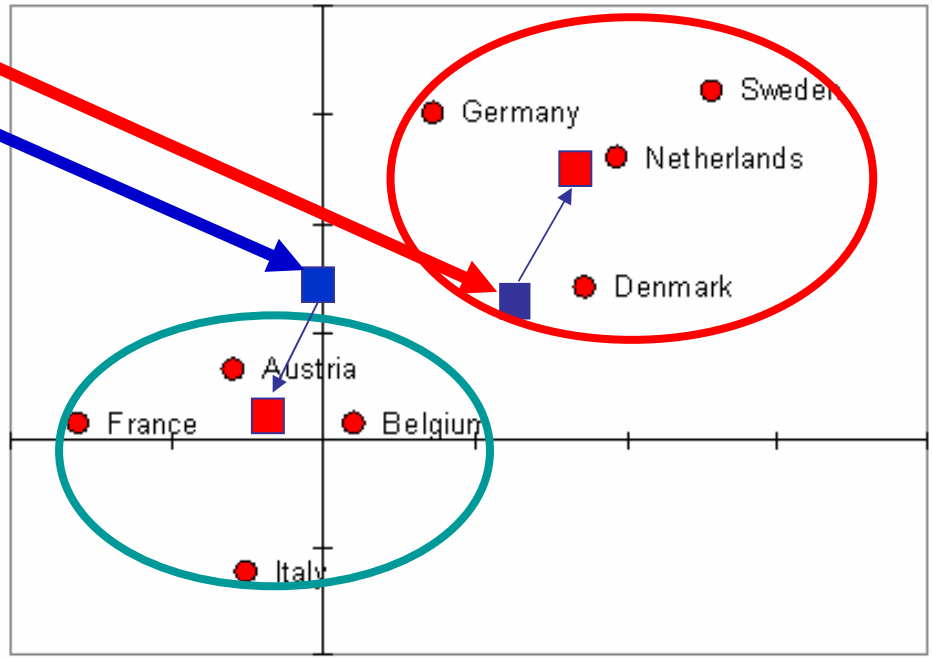
Step 1: Select the initial partition

(this partition may also be defined on the basis of a preliminary cluster analysis / hierarchical procedure)

Usually, ***G seeds*** are selected

Step 2: Allocation

Each case is allocated to the closest cluster (closest centroid)



Step 3: Seeds update

Seeds are updated: centroid of the obtained clusters

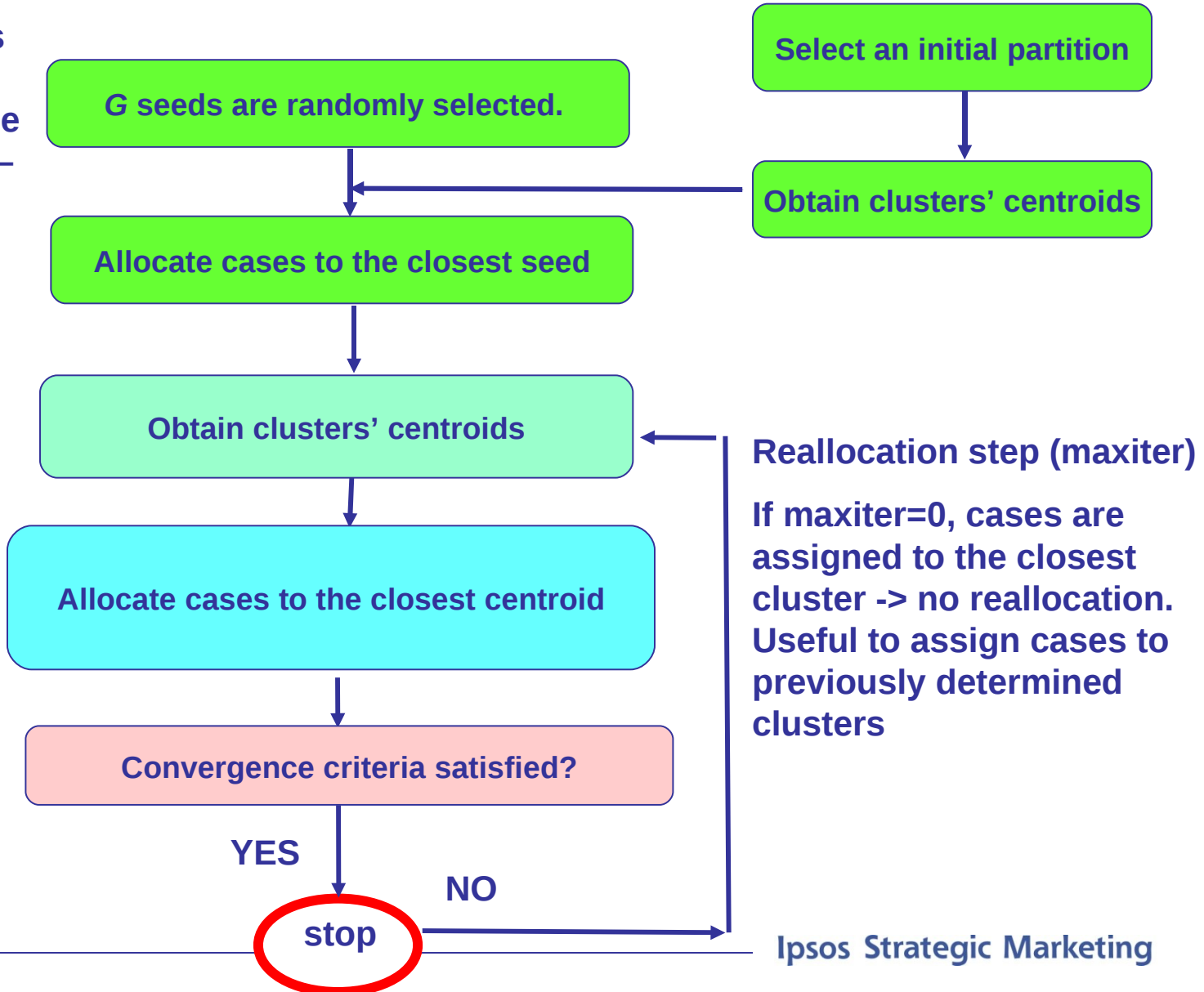
Step 2 and 3 are iterated until convergence:

2. Re-allocation
3. Seeds update

Cluster analysis – partitioning algorithms

k-means algorithm

SAS – subroutines defined to select seeds not too close one to each other – higher group separation



Reallocation step (maxiter)

If maxiter=0, cases are assigned to the closest cluster -> no reallocation. Useful to assign cases to previously determined clusters

Clustering procedure

- Consists of the following steps:

1. Defining the problem
2. Finding appropriate dissimilarity or similarity measure
3. Deciding on the grouping algorythm
4. Choosing the appropriate number of clusters
5. Describing and profiling the newly-formed clusters.

Advantages and disadvantages of the two clustering approaches

- Hierarchical clustering
  - Advantages: it has a logical structure, is easy to read and interpret.
  - Disadvantages: it is relatively unstable and unreliable – the first combination or separation of objects, which may be based on a small difference in the criterion, will constrain the rest of the analysis
  - It is sound practice to split the sample into at least two groups and do two independent clustering runs...
- Nonhierarchical clustering
  - Advantages: more reliable approach (similar results for split-sample results), objects may leave one enter another cluster to improve criterion
  - Disadvantages: the series of clusters is usually a mess and difficult to interpret; the number of clusters must be predefined.

Advantages and disadvantages of the two clustering approaches

- The both approaches could be used consequtively.
- Hierarchical approach could be used to identify the number of clusters, remove the “outliers” and to obtain group centers
- On the remaining objects the nonhierarchical approach is used. The inputs are the number of clusters, and the cluster centers obtained from the hierarchical approach.
- This combines the merits of both approaches...

The final word: Cautions

- Cluster analysis (as we described it) is a descriptive technique. The solution is not unique and it strongly depends upon the analyst's choices. We described how it is possible to combine different results in order to obtain stable clusters, not depending too much on the criteria selected to analyze data.
- Cluster analysis always provides groups, even if there is no group structure. When applying a cluster analysis we are *hypothesizing* that the groups exist. But this assumption may be false or weak.
- Cluster analysis results' should not be generalized**.** Cases in the same cluster are (hopefully) *similar* only with respect to the information cluster analysis was based on (i.e., dimensions/variables inducing the considered dissimilarities).