University of Ferrara

Statistics for Economics and Business

Stefano Bonnini & Valentina Mini

Cluster Analysis: hierarchical methods

Lecture 10 – 22th of March 2019

CA hierarchical methods

- Hierarchical methods provide a family of partitions of the statistical units with a number ***g*** of groups which varies from ***n* to *1:***
  - o Trivial starting partition: ***g=n*** groups of *1* unit
  - o Intermediate partitions: ***1 < g < n***
  - o Final partition: ***g=1*** group of *n* units

Example: wine survey on Passito

- oTrivial starting partition: g= 386 (each customer is one group)
- oIntermediate partitions: number of groups varies from *385 to 2*
- oFinal partition: *g=1 (*all *386* customers represent one group)

CA hierarchical methods

Methods which use the $n \times n$ matrix of distances (or of proximities) D:

1. The two nearest units (with minimum distance or maximum proximity) are grouped

2. A new $(n-1) \times (n-1)$ D matrix is computed, which represents the distances (or proximities) between the *n-1* clusters obtained in the previous step (*n-2* clusters with *1* unit and *1* cluster with *2* units)

3. In the new D matrix the minimum distance (or maximum proximity) is detected and the two corresponding clusters are grouped

4. Previous steps are repeated, according to an iterated procedure, where at step *t* we have g=*n-t+1* groups and a $(n-t+1) \times (n-t+1)$ *D* matrix, and the two nearest clusters are grouped, with *t=1,…,n*

5. At the end of the procedure (*t=n*) we have *1* group with all the *n* units

Ex:

Sptep 1

$$D^{(1)} = \begin{array}{c|ccccc} & A & B & C & D & E \\ A & 0 & 1 & 5 & 6 & 8 \\ B & 1 & 0 & 3 & 8 & 7 \\ C & 5 & 3 & 0 & 4 & 6 \\ D & 6 & 8 & 4 & 0 & 2 \\ E & 8 & 7 & 6 & 2 & 0 \end{array}$$

Lower distance within our statistical units

After the first step we cluster A and B together and we treat it as a single entity.
Now we re-compute the distance matrix among pairs of our ***4 elements (AB, C, D, E)***

Step 2

$$D^{(2)} = \begin{array}{c|cccc} & AB & C & D & E \\ AB & 0 & 3 & 6 & 7 \\ C & 3 & 0 & 4 & 6 \\ D & 6 & 4 & 0 & 2 \\ E & 7 & 6 & 2 & 0 \end{array}$$

Lower distance within our statistical units

After the second step we cluster D and E together and we treat it as a single entity.
Now we re-compute the distance matrix among pairs of our ***3 elements (AB, C, DE)***
And so on

....

CA hierarchical methods

Criteria for computing the distance between two clusters (groups):

Let C_1 and C_2 be two clusters witn n_1 and n_2 units respectively

- **Single linkage** or **nearest neighbour** method:

$$d(C_1,C_2)=min(d_{iu})\ i \in C_1\ ,\ u \in C_2$$

- **Complete linkage** or **farthest neighbour** method:

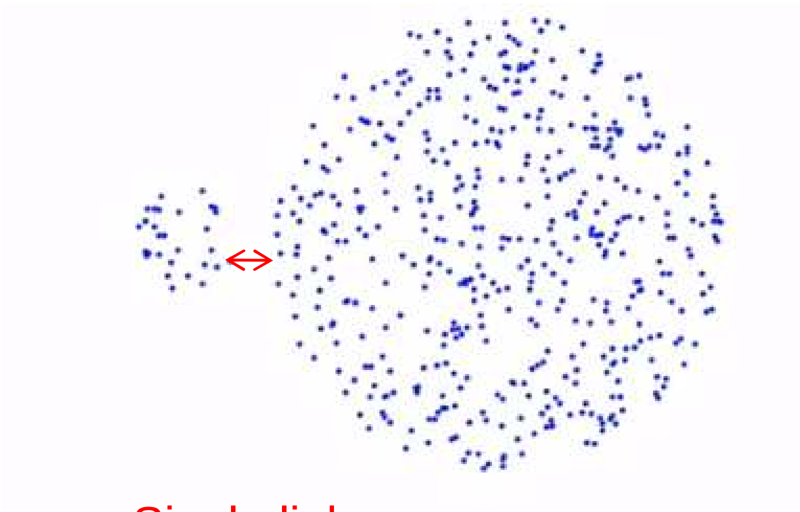
$$d(C_1,C_2)=max(d_{iu})\ i \in C_1\ ,\ u \in C_2$$

- **Average linkage between groups** method or **UPGMA** (Unweighted Pair-Group Method Using arithmetic Averages):

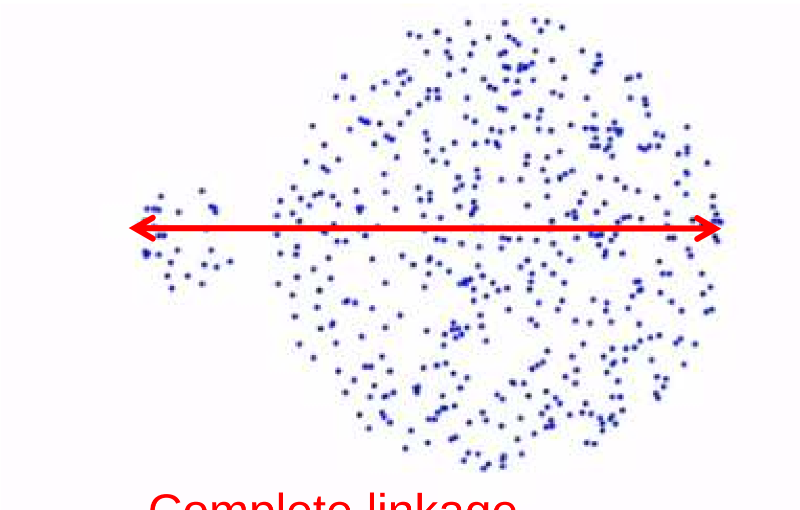
$$d(C_1,C_2)= \Sigma_{i,u}\ d_{iu} / (n_1 n_2)\ ,\ i \in C_1\ ,\ u \in C_2$$

- **Average linkage within groups** method (arithmetic average of the distances between all the $m=n_1+n_2$ units of the two clusters joined together):

$$d(C_1,C_2)= \Sigma_{i>u}\ d_{iu} / [m(m-1)/2]\ ,\ i,u \in C_1 \cup C_2$$

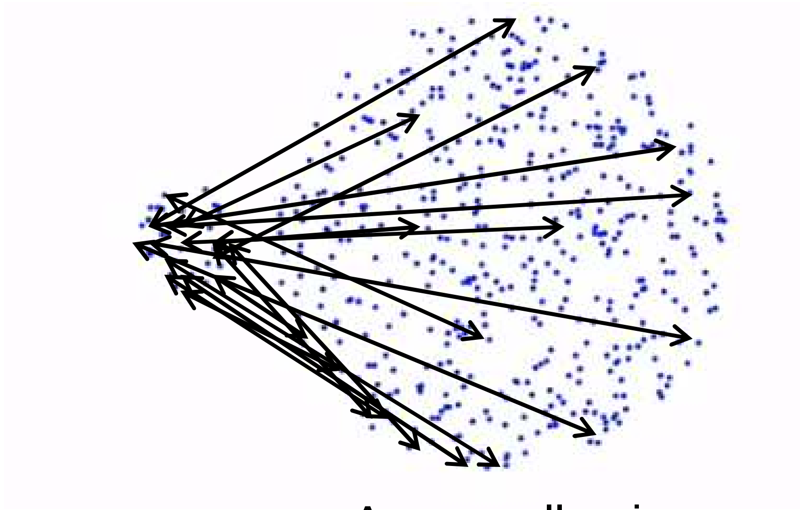


Single linkage

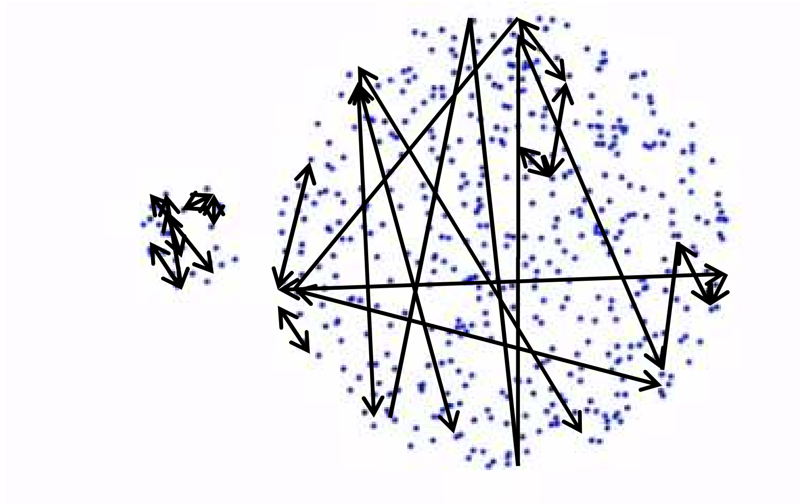


Complete linkage

Average linkages:



…. Among all pairs



CA hierarchical methods

Remarks:

- With the nearest neighbour method (SINGLE LINKAGE) we can have the «**chain effect**»:
    - two far units can be joined into the same cluster in the presence of a sequence of intermediate points

- With the farthest neighbour method (COMPLETE LINKAGE) we can have compact groups but with **an approximately hyperspherical shape**

- Average linkage method **can be a good compromise** to have internal cohesion and external separation between the groups

CA hierarchical methods

Hierarchical methods which also use the original matrix of observed data:

- Centroid method:

$$d(C_{1,}, C_2) = d(\bar{\boldsymbol{x}}_{1,}, \bar{\boldsymbol{x}}_2)$$

the distance between two clusters is equal to the distance between the two *k*-dimensional vectors of means computed on the n_1 units of C_1 and the n_2 units of C_2

CA hierarchical methods

Hierarchical methods which also use the original matrix of observed data:

- Ward method or least deviance method.

Uses the breakdown of the total deviance:

$$TD = \sum_{j=1}^{k} \sum_{i=1}^{n} \left(x_{ij} - \bar{x}_j\right)^2$$

$$WD = \sum_{l=1}^{g} \left[\sum_{j=1}^{k} \sum_{i=1}^{n} \left(x_{ij} - \bar{x}_{j,l}\right)^2\right] = \sum_{l=1}^{g} DW_l$$

$$BD = \sum_{j=1}^{k} \sum_{l=1}^{g} n_l \left(\bar{x}_{j,l} - \bar{x}_j\right)^2$$

$$TD = WD + BD$$

$\bar{x}_j$: sample mean of *j*-th variable

$\bar{x}_{j,l}$: sample mean of *j*-th variable in cluster *l*

At each step of the procedure, the aggregation which causes the least increasing of *DW* is chosen

CA hierarchical methods

Criteria for evaluating the partitioning:

Let C_1 and C_2 be two clusters witn n_1 and n_2 units respectively

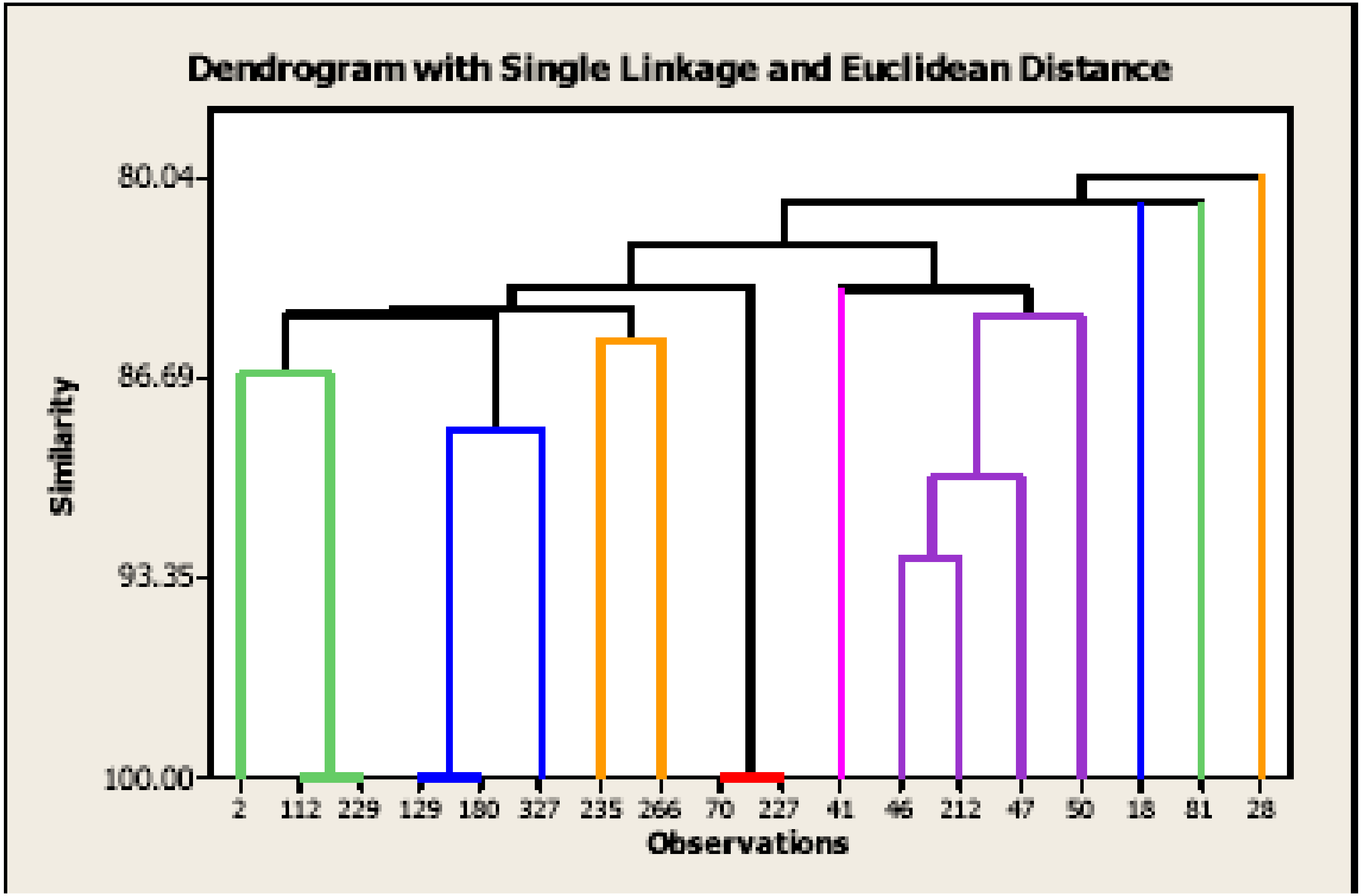
- Given a partition of the units in *g* groups, the proportion of global variability explained by this partition is:

$$R^2 = 1 - WD / TD = BD / TD$$

This index takes values between 0 and 1 and the **smaller the number *g* (of groups) the smaller the index value**

CA hierarchical methods

Dendogram



The state of the art

CA: aim = identify the lower number of clusters such that

The units belonging the same cluster are more similar than …→ The units belonging different clusters

- High within-cluster similarity
- Low within-cluster variance

- Low between-cluster similarity
- High between-cluster variance

To identify clusters we should define

Distance or similarity

Distance:

-Euclidean

-Manhattan

-Minkosky

-Chebichev

Similarity:
1. case of dichotomous var.
2. case of categorical var.
:

*Ind. of co-presences (Russel&Rao; Jaccart)

*Ind. Co-presences and co-absences (Sokal & Michener)

Grouping's rule

Hierarchical methods

Non Hier. methods

Divisive:

-Edwards & Cavalli Sforza (trace of the deviance matrix)

-Friedman &Rubin (min. the deviance matrix determinant)

Agglomerative:

-Single linkage
-Complete linkage
-Average linkage

•Centroide method
• Ward method