University of Ferrara

DIPARTIMENTO
DI ECONOMIA
E MANAGEMENT

Stefano Bonnini

ClusterAnalysis

Summary

- Introduction to Cluster Analysis (CA)
- Distances and Similarity Indices
- CA hierarchical methods
- CA non hierarchical methods

Introduction to Cluster Analysis (CA)

- Available information: data about k variables observed on n statistical units

- Table of data: $n \times k$ matrix $\boldsymbol{X}=[x_{ij}]$
  - x_{ij} = value of X_j observed on unit i
  - $i=1,\ldots,n$
  - $j=1,\ldots,k$

- Goal of the analysis: classification of the n units into homogeneous groups, according to predefined criteria of diversity or similarity, with the intent of getting a small number of categories or classes

Introduction to Cluster Analysis (CA)

Example: A marketing survey on the demand of the wine «Passito» has been performed.

A sample of n=386 people has been interviewed. The questionnaire includes several questions about their preferences and behaviors related to drinking wine

- Age: ____ - Sex: M ❍ F ❍ - Province of Residence: ____________

- Do you like drinking wine?	not at all	1	2	3	4	5	6	7	very much
		❍	❍	❍	❍	❍	❍	❍	

- How often do you drink wine…	never	rarely	sometimes	often	regularly
…at home with meals?	❍	❍	❍	❍	❍
…in bars or pubs?	❍	❍	❍	❍	❍
…at restaurants with meals?	❍	❍	❍	❍	❍

- Do you know the wine Passito?	not at all	1	2	3	4	5	6	7	very well
		❍	❍	❍	❍	❍	❍	❍	

Introduction to Cluster Analysis (CA)

The variables:

Label	Description	Coding
ID	Personal ID of the interviewed	Increasing integer number
AgeClass	Age of the person	Age (years)
AGE_CLASS	Age class of the person	1-6
SEX	Sex of the person	M or F
PROV	Province where the interviewed lives	Province code
LIKE_WINE	How much do you like drinking wine?	Integrer number from 1 to 7
FREQ_HOME	How often do you drink wine at home with meals?	Integrer number from 1 to 5
FREQ_BAR	How often do you drink wine in bars/pubs?	Integrer number from 1 to 5
FREQ_REST	How often do you drink wine at restaurants with meals?	Integrer number from 1 to 5
KNOW_PAS	Do you know the wine Passito?	Integrer number from 1 to 7
FREQ_PAS	How often do you drink Passito?	Integrer number from 1 to 5
FREQ_P_HOL	How often do you drink Passito on holidays and celebrations?	Integrer number from 1 to 5
FREQ_P_ALO	How often do you drink Passito when you are alone?	Integrer number from 1 to 5
FREQ_P_MEA	How often do you drink Passito at the end of meals?	Integrer number from 1 to 5
FREQ_P_OFF	How often do you drink Passito offered by someone?	Integrer number from 1 to 5
HOW_MUCH	How much wine do you drink in one year?	Integrer number from 1 to 4
LIKE_PAS	How much do you like drinking Passito?	Integrer number from 1 to 7
LIKE_AROMA	How much do you like aroma and smell of Passito?	Integrer number from 1 to 7
LIKE_SWEET	How much do you like the sweetness of Passito?	Integrer number from 1 to 7
LIKE_ALCOHOL	How much do you like the alcohol content of Passito?	Integrer number from 1 to 7
LIKE_TASTE	How much do you like the intensity of taste of Passito?	Integrer number from 1 to 7
PRICE	How much could you pay for one bottle of Passito? (0.5 litre)	Integrer number from 1 to 5

Introduction to Cluster Analysis (CA)

The dataset:

ID	AGE	AGE_CLAS	SEX	PROV	LIKE_WINE	FREQ_HOME	FREQ_BAR	FREQ_REST	KNOW_PAS	…
1	26	1	M	PD	6	2	4	4	4	
2	43	3	M	PD	7	3	1	4	6	
3	32	2	M	VR	6	4	3	3	6	
4	53	4	F	PD	6	4	2	5	5	
5	30	2	M	PD	4	2	3	4	2	
6	23	1	F	VR	5	3	2	4	5	
7	46	3	M	VE	5	2	3	6		
8	26	1	M	PD	6	3	2	5	5	
9	25	1	M	BL	6	3	4	4	7	
10	22	1	M	VE	5	3	4	4	5	
11	24	1	M	VE	4	1	3	3	3	
12	22	1	M	VE	7	5	4	5	7	
13	23	1	M	VI	7	3	5	5	7	
14	23	1	M	VE	7	4	4	4	4	
…		…	…	…	…	…	…	…	…	

Introduction to Cluster Analysis (CA)

- The Group Analysis or Cluster Analysis is a typical explorative method for the identification of clusters of similar units according to the *n* *k*-dimensional observations. Before the analysis there is no certainty that such groups exist

- Example: segmentation of the market of wine drinkers by the identification of homogeneous groups of customers

- Final result: reduction of the dimension of the table of data from the point of view of the statistical units (number of rows) $\rightarrow$ from *n* observed statistical units to *g* homogeneous groups ($g \ll n$)

Introduction to Cluster Analysis (CA)

Choices in CA:

- Which informative variables must be considered?
- Which distance or index of similarity must be used?
- Which method for the definition of the groups must be applied?
  - General criterium: internal cohesion and external separation
  - Methods:
    - Hierarchical method: progressive aggregation of units
    - Non hierarchical method: unique partition given the number g of groups
- How to evaluate the final partitions and to choose the optimal one?

Summary

- Introduction to Cluster Analysis (CA)
- Distances and Similarity Indices
- CA hierarchical methods
- CA non hierarchical methods

Distances and Similarity Indices

- Let's denote with $\boldsymbol{x}_i=(x_{i1},x_{i2},\ldots,x_{ik})'$ *and* $\boldsymbol{x}_u=(x_{u1},x_{u2},\ldots,x_{uk})'$ the k-dimensional vectors of two statistical units (i-th and u-th row of the dataset)

- *Proximity:* resemblance, non diversity, … between two statistical units measured through the index $PI_{iu}=f(\boldsymbol{x}_i,\boldsymbol{x}_u)$

Distances and Similarity Indices

- *Proximity Indices:*

    - *For numeric variables*
        - ✓ *Distances*
        - ✓ Distance indices
        - ✓ Dissimilarity indices

    - For categorical variables
        - ✓ Similarity indices

Distances and Similarity Indices

• *Distance (metrics)* between units *i* and u is a function $d_{iu}=d(\mathbf{x}_i, \mathbf{x}_u)$ such that:

1. $d(\mathbf{x}_i, \mathbf{x}_u) \geq 0$ (non negativity)

2. $d(\mathbf{x}_i, \mathbf{x}_u) = 0 \Leftrightarrow \mathbf{x}_i = \mathbf{x}_u$ (identity)

3. $d(\mathbf{x}_i, \mathbf{x}_u) = d(\mathbf{x}_u, \mathbf{x}_i)$ (symmetry)

4. $d(\mathbf{x}_i, \mathbf{x}_u) \leq d(\mathbf{x}_i, \mathbf{x}_s) + d(\mathbf{x}_s, \mathbf{x}_u) \quad \forall \mathbf{x}_i, \mathbf{x}_s, \mathbf{x}_u \in \Re^k$
(triangular inequality)

Distances and Similarity Indices

- Euclidean distance: $_2d_{iu} = \|\boldsymbol{x}_i - \boldsymbol{x}_u\| = \left[\sum_{j=1}^{k}\left(x_{ij} - x_{uj}\right)^2\right]^{1/2}$
- Manhattan distance: $_1d_{iu} = \sum_{j=1}^{k}\left|x_{ij} - x_{uj}\right|$
- Minkowski distance: $_md_{iu} = \left[\sum_{j=1}^{k}\left|x_{ij} - x_{uj}\right|^m\right]^{1/m}$
- Chebichev distance: $_\infty d_{iu} = \lim_{m\to\infty} {}_md_{iu} = \max_{j=1,\cdots,k}\left|x_{ij} - x_{uj}\right|$
  (Lagrange distance)

Distances and Similarity Indices

- Properties:
  - *P1:* euclidean distance $_2d_{iu}$ is affected more strongly than Manhattan distance by great differences between pairs of values
  - *P2:* Minkowski distance $_m d_{iu}$ is non increasing function of parameter *m*: $_1d_{iu} \geq {}_2 d_{iu} \geq \cdots \geq{}_\infty d_{iu}$

Distances and Similarity Indices

- Properties:

  - *P3:* Minkowski distance ${}_m d_{iu}$ is invariant respect to variable translation ${}_m d(\boldsymbol{x}_i + \boldsymbol{c}, \boldsymbol{x}_u + \boldsymbol{c}) = {}_m d(\boldsymbol{x}_i, \boldsymbol{x}_u)$, with $\boldsymbol{c} = (c_1, \cdots, c_k)' \epsilon \, \Re^k$ but not respect to linear transformations of one or more variables such as $a_j x_{ij} + c_j$, $i = 1, \cdots, n, j = 1, \cdots, k$. Hence a change of the scale or the measurement unit determines a change of the distance

  - *P4:* euclidean distance ${}_2 d_{iu}$ is invariant respect to ortogonal transformations (rotations), that is ${}_2 d(\boldsymbol{T}\boldsymbol{x}_i, \boldsymbol{T}\boldsymbol{x}_u) = {}_2 d(\boldsymbol{x}_i, \boldsymbol{x}_u)$ with $\boldsymbol{T}$ $k \times k$ matrix such that $\boldsymbol{T}'\boldsymbol{T} = \boldsymbol{I}$

Distances and Similarity Indices

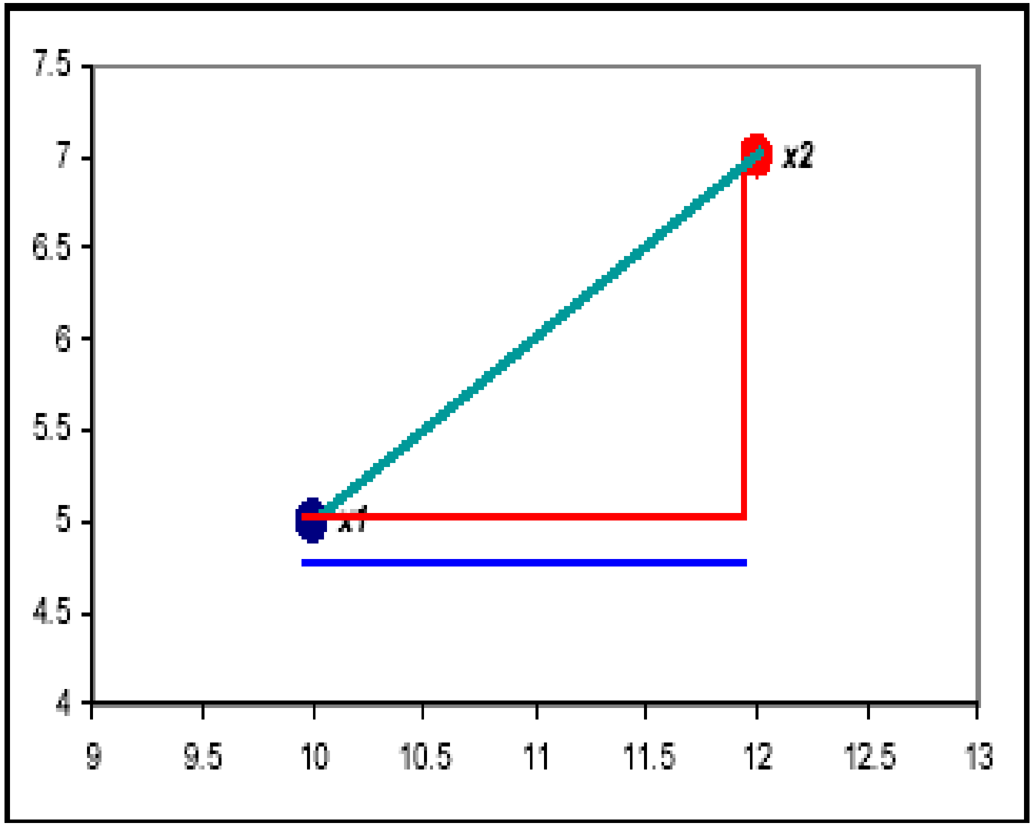
Example 1.

n=2 and k=2
$\mathbf{x}_1$=(10;5)'
$\mathbf{x}_2$=(12;7)'

${}_1d_{12} = 4$

${}_2d_{12} = 2.83$

${}_\infty d_{12} = 2$



Distances and Similarity Indices

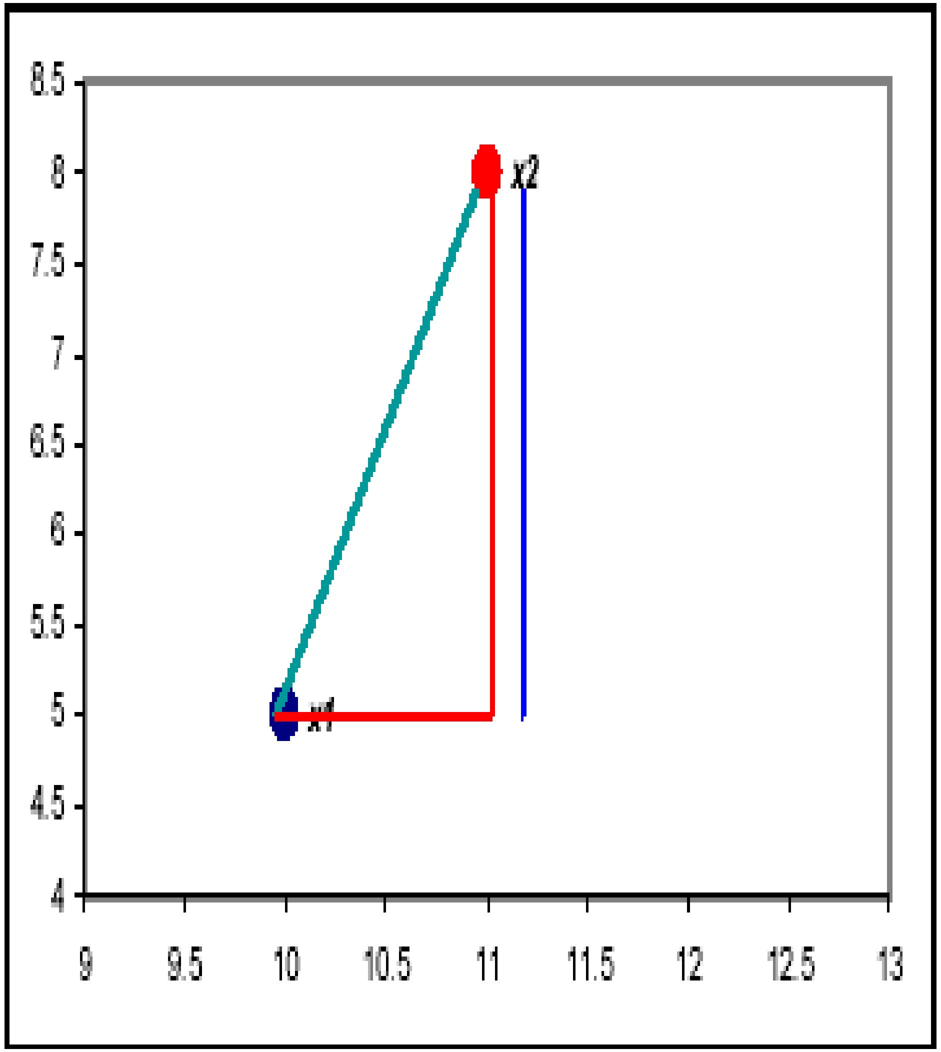
Example 2.

n=2 and k=2
$\mathbf{x}_1$=(10;5)'
$\mathbf{x}_2$=(11;8)'

${}_1\mathbf{d}_{12} = 4$

${}_2\mathbf{d}_{12} = 3.16$

${}_\infty\mathbf{d}_{12} = 3$



Distances and Similarity Indices

Starting point of hierarchical methods: $n \times n$ matrix of distances

$$\mathbf{D} = \left[d_{ij} \right] = \begin{bmatrix} 0 & d_{12} & d_{13} & \dots & d_{1n} \\ & 0 & d_{23} & \dots & d_{2n} \\ & & 0 & \dots & d_{3n} \\ & & & \dots & \dots \\ & & & & 0 \end{bmatrix}$$

Distances and Similarity Indices

- *Distance index* between units *i* and u is a function $DI_{iu}=DI(\boldsymbol{x}_i, \boldsymbol{x}_u)$ such that:

1. $DI(\boldsymbol{x}_i, \boldsymbol{x}_u) \geq 0$ (non negativity)

2. $DI(\boldsymbol{x}_i, \boldsymbol{x}_u) = 0 \Leftrightarrow \boldsymbol{x}_i = \boldsymbol{x}_u$ (identity)

3. $d(\boldsymbol{x}_i, \boldsymbol{x}_u) = d(\boldsymbol{x}_u, \boldsymbol{x}_i)$ (symmetry)

Example: ${}_2d_{iu}^2 = \|\boldsymbol{x}_i - \boldsymbol{x}_u\|^2$ statisfies the additivity property, that is:

$${}_2d_{iu}^2 = \sum_{j=1}^{k_1} (x_{ij} - x_{uj})^2 + \sum_{j=k_1+1}^{k} (x_{ij} - x_{uj})^2$$

Distances and Similarity Indices

• Dissimilarity index (or diversity index according to Leti) between units *i* and u is a function $DS_{iu}=DS(\mathbf{x}_i ,\mathbf{x}_u)$ such that:

1. $DS(\mathbf{x}_i ,\mathbf{x}_u)\geq 0$ (non negativity)

2. $DS(\mathbf{x}_i ,\mathbf{x}_u)= 0 \Leftarrow \mathbf{x}_i = \mathbf{x}_u$

3. $DS(\mathbf{x}_i ,\mathbf{x}_u)= DS(\mathbf{x}_u ,\mathbf{x}_i)$ (symmetry)

Distances and Similarity Indices

• Similarity index (with categorical variables) between units i and u is a function $S_{iu}=S(\mathbf{x}_i, \mathbf{x}_u)$ such that:

1. $S(\mathbf{x}_i, \mathbf{x}_u) \geq 0$ (non negativity)
2. $S(\mathbf{x}_i, \mathbf{x}_i) = 1 \;\; \forall i$ (normalization)
3. $S(\mathbf{x}_i, \mathbf{x}_u) = S(\mathbf{x}_u, \mathbf{x}_i)$ (symmetry)

Distances and Similarity Indices

<u>Case of *k* dichotomous variables</u>:

- Each variable takes two possible levels
  - 1=presence of a given characteristic
  - 0= absence of the characteristic

- For each couple of units *(i,u)* we compute:
  - f_{11} = frequency of characteristics jointly present in *i* and *u* $\rightarrow \sum_{j=1}^{k} x_{ij}\, x_{uj}$
  - f_{10} = frequency of of characteristics present in *i* but not in *u* $\rightarrow \sum_{j=1}^{k} x_{ij}\,(1 - x_{uj})$
  - f_{01} = frequency of of characteristics present in *u* but not in *i* $\rightarrow \sum_{j=1}^{k} (1 - x_{ij}) x_{uj}$
  - f_{00} = frequency of characteristics jointly absent in *i* and in *u* $\rightarrow \sum_{j=1}^{k} (1 - x_{ij})(1 - x_{uj})$

Distances and Similarity Indices

Indices based on co-presences:

- Index of Russel and Rao: $${}_1S_{iu} = \frac{f_{11}}{k}$$

- Index of Jaccart: $${}_2S_{iu} = \frac{f_{11}}{f_{11}+f_{10}+f_{01}}$$

Indices based on co-presences and co-absences:

- Index of Sokal and Michener: $${}_3S_{iu} = \frac{f_{11}+f_{00}}{k}$$
  (index of simple correspondence)

We have: ${}_1S_{iu} \leq {}_2S_{iu} \leq {}_3S_{iu} =$

Distances and Similarity Indices

<u>Case of *k* categorical (not all dichotomous) variables</u>:

- Some of the k variables can take more than two levels (categories)
- Variable X_j can take r_j categories and $\sum_{j=1}^{k} r_j = R$
- Each variable can be represented by r_j dichotomous variables *(j=1,…,k)*
- An index based on co-presences applied to the *R* dichotomous variables can be considered

Summary

- Introduction to Cluster Analysis (CA)
- Distances and Similarity Indices
- CA hierarchical methods
- CA non hierarchical methods

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*
  - Trivial starting partition: $g=n$ groups of *1* unit
  - Intermediate partitions: $1 < g < n$
  - Final partition: $g=1$ group of *n* units

- Example: wine survey on Passito
  - Trivial starting partition: each customer is one group
  - Intermediate partitions: number of groups varies from *385 to 2*
  - Final partition: 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 $t=1,\ldots,n$
5. At the end of the procedure ($t=n$) we have 1 group with all the n units

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):

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

CA hierarchical methods

Remarks:

- With the nearest neighbour method 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 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

<u>Criteria for evaluating the partitioning</u>:

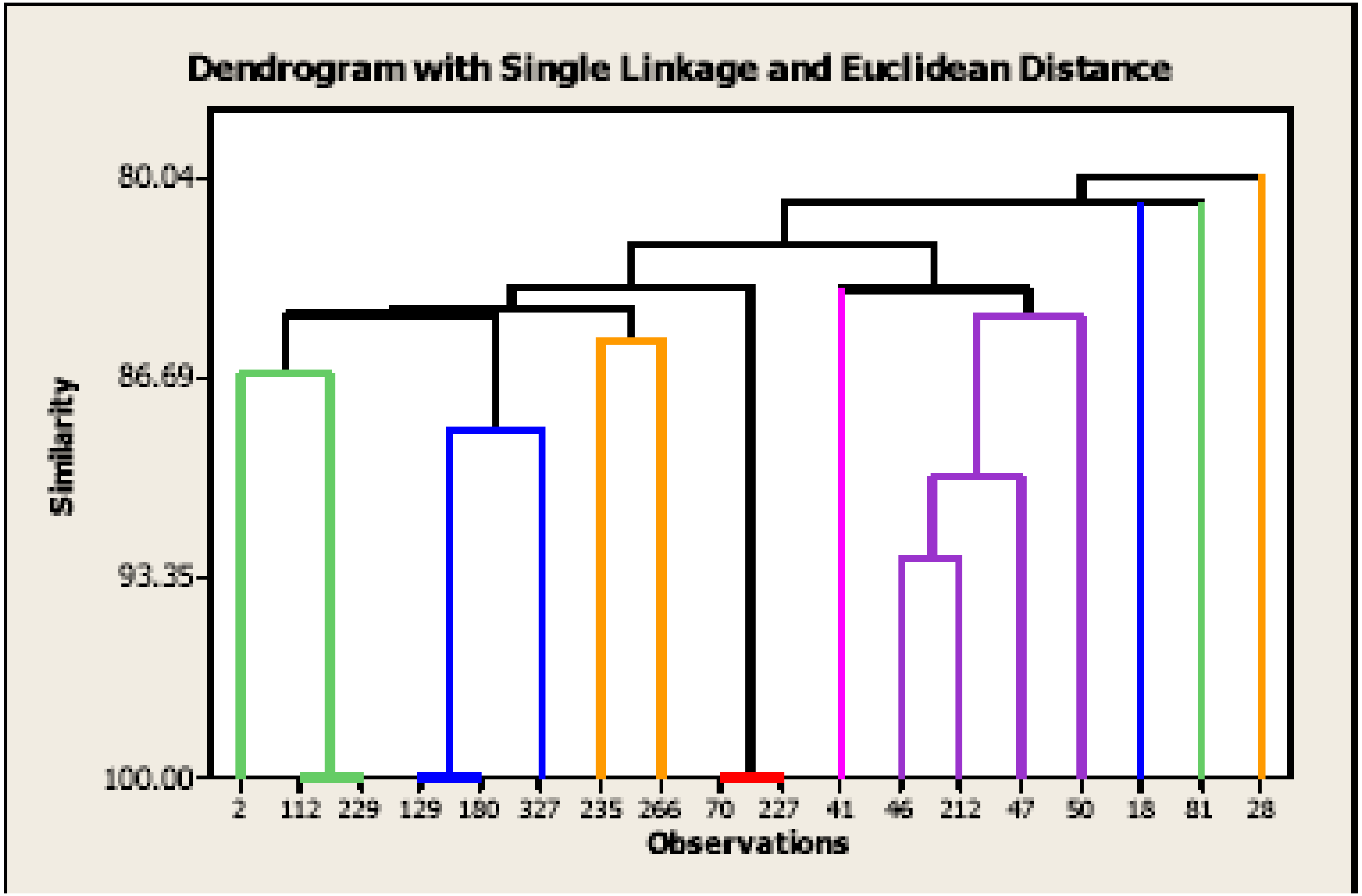
- Given a partition of the units in *g* groups, the proportion of global variability explained by this partition is:

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

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



Summary

- Introduction to Cluster Analysis (CA)
- Distances and Similarity Indices
- CA hierarchical methods
- CA non hierarchical methods

CA non hierarchical methods

With the non hierarchical methods we get just one partition of the *n* units into *g* clusters, for a pre-determined number *g* of clusters

- The rule for the allocation of the units into the clusters considers an objective function, usually based on the breakdown of the total deviance *TD*
- Example. Wine survey on Passito:
    - Starting partition: the customers are classified into *g* groups
    - Intermediate partitions: the customers are reallocated in the groups and, for each reallocation, the corresponding value of the objective function is computed
    - Each customer is assigned to a new group when this assignement provides the greatest improvement of internal cohesion
    - The reallocations are repeated until a given stopping rule is satisfied
- Weaknesses of the procedure: (1) The choice of the number g of groups is arbitrary; (2) the starting partition affects the final result

CA non hierarchical methods

K-means method

1. Chose *g* starting seeds or poles as centroids of the starting partition and assign each unit to the cluster with the nearest centroid
2. Compute the centroids of the *g* new clusters created at step (1)
3. Assign each unit to the new cluster with the nearest centroid
4. Repeat step (2) and step (3) until one of the following convergence rules is satisfied:
   I. R^2 variation is less than a given treshold
   II. The changes of the centroid positions are less than a given treshold
   III. The number of iterations reaches a certain predetermined value
   IV. …

Remark: with euclidean distance we always have convergence of the algorithm

R exercises

Problem 1 - Passito

- Perform a hierarchical CA on the 17 response variables of the questionnaire which represent habits, behaviors and preferences of wine drinkers (from variable LIKE_WINE to variable PRICE) to detect homogeneous market segments of wine drinkers
- Perform a k-means CA on the 17 response variables of the questionnaire to detect 4 homogeneous market segments of wine drinkers

R exercises

Problem 2 - Students

- Perform a FA on the 5 observed response variables to detect new *q<5* variables which «explain» data
- Perform a PCA on the 5 response variables with the same goal

R exercises

Problem 3 – Eating Habits

- Perform a FA on the 12 observed response variables (from *Alcoholic.Beverages* to *Milk*) to detect new *q<12* variables which «explain» data
- Perform a PCA on the 12 response variables with the same goal