*Università degli studi di Ferrara*
*Dipartimento di Matematica*
*A.A. 2019/2020 – I semestre*

# STATISTICA MULTIVARIATA

SSD MAT/06

## LEZIONE 15– Cluster analysis

Docente: Valentina MINI

valentina.mini@unife.it

RICEVIMENTO: LUNEDI POMERIGGIO,
appuntamento previa mail

# INTRODUZIONE

La cluster analysis e' una tecnica di analisi multivariata attraverso la quale e' possibile raggruppare le unità statistiche, in modo da minimizzare la "lontananza logica" interna a ciascun gruppo e di massimizzare quella tra i gruppi.

La "lontananza logica" viene quantificata per mezzo di misure di similarità/dissimilarità definite tra le unità statistiche.

# FOCUS DELL'ANALISI: DEFINIZIONI DI DISTANZE

**Date le seguenti proprietà:**

1. separabilità $d(i,h) = 0$ se e solo se $x_i = x_h$.
2. simmetria $d(i,h) = d(h,i)$
3. disuguaglianza triangolare:

$$d(i,h) \leq d(i,e) + d(e,h) \quad \forall i, e, h$$

4. condizione di Krassner:

$$d(i,h) \leq \sup(d(i,e); d(e,h)) \quad \forall i, e, h$$

**Un'applicazione d che associa un valore positivo o nullo a ciascuna coppia (i,h) si definisce:**

a) indice di dissimilarità se soddisfa le proprietà 1 e 2;
b) metrica o distanza se soddisfa 1,2 e 3;
c) ultrametrica se soddisfa 1,2 e 4.

# FOCUS DELL'ANALISI: DEFINIZIONI DI DISTANZE

La scelta tra indici di dissimilarità e metrica e' legata al tipo di dati che si hanno a disposizione.
Per dati di tipo numerico (quantitativi) possiamo utilizzare delle misure di distanza, ovvero delle metriche.
Per dati di tipo qualitativo bisogna utilizzare misure *matching-type*, cioe' di associazione (similarità o dissimilarità).

**Come si effettua una cluster analysis?**

Si parte dalla matrice dei dati X di dimensione *nxp* e la si trasforma in una matrice *nxn* di dissimilarità o di distanze tra le n coppie di osservazioni (vettori di p elementi).
Si sceglie poi un algoritmo che definisca le regole su come raggruppare le unità in sottogruppi sulla base delle loro similarità.
Lo scopo e' di identificare un minor numero di gruppi tali che gli elementi appartenenti ad un gruppo siano – in qualche senso – piu' simili tra loro che non agli elementi appartenenti ad altri gruppi.

- Cluster Analysis is a collective term for **various methods** to find **group structures in data**

- The groups are called **CLUSTERS** and are usually not known *a priori*

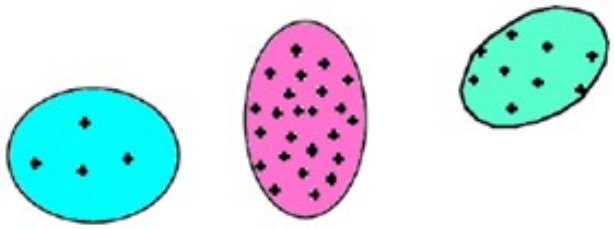
- **The aim = identification of a minimum number of groups** such that:
  - we minimize the "distance" among statistical units **within** the same cluster;
  - We maximize the "distance" **between** different clusters

- Basic concepts:
  - **Distance for quantitative** data
  - **Similarity (association) for qualitative** data

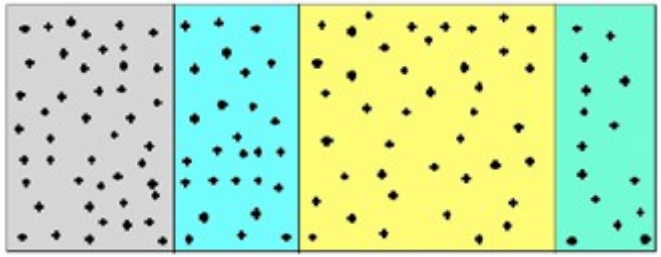
Natural or arbitrary cluster?

*Kruskal (1977):" ... We call clusters natural if the membership is determined fairly well in a natural way by the data, and we call the clusters arbitrary if there is a substantial arbitrary element in the assignment process".*

Natural clusters



Arbitrary clusters



## INDIVIDUAZIONE DI GRUPPI

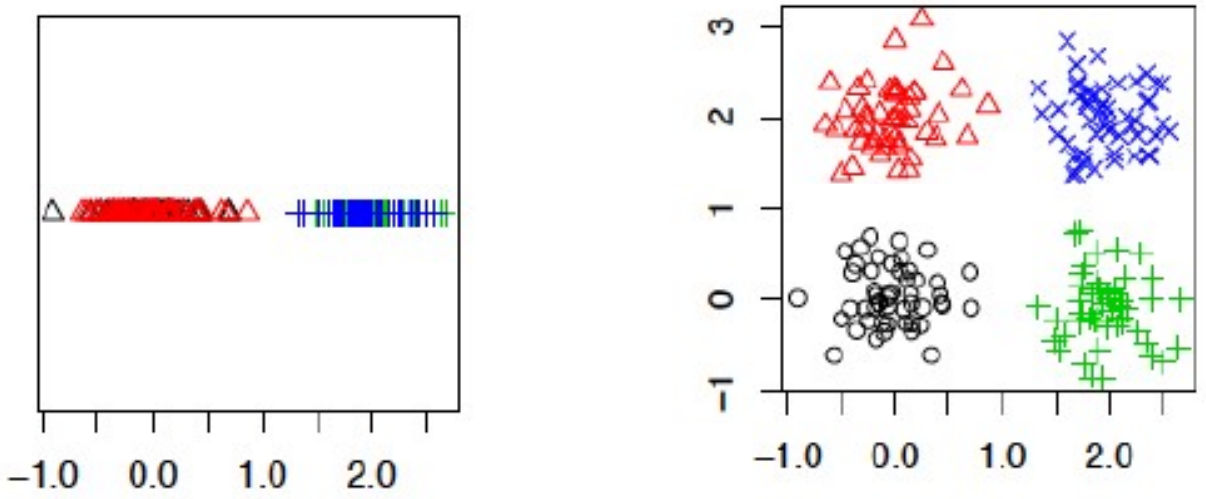
How many clusters?

**Sebbene non sia semplice immaginare una tecnica con basi così labili ed evanescenti è altrettanto complicato trovare una tecnica di analisi così tanto diffusa**

# INTRODUZIONE

## INDIVIDUAZIONE DI GRUPPI

La cluster analysis è una procedura multivariata. La scelta di quante e quali variabili adoperare è fondamentale.




I quattro raggruppamenti sono evidenti nel grafico bivariato a destra, ma si riesce a percepirne solo due nel grafico univariato a sinistra

- The cluster analysis is the scientific procedure **to identify clusters**

- In cluster analysis, **the group membership** of the individual observations is determined such that:
  - the **groups are as heterogeneous as possible** and
  - the observations **within a group are as homogeneous** as possible

- Whether a resulting cluster solution makes sense depends in practice on the **interpretability** of the identified clusters

- There are **many different algorithms** which result in different numbers and sizes of clusters: in the following, we focus on the most common procedures

Clustering methods can be divided into two large groups:

1. **partitioning** methods and
2. **hierarchical** clustering methods

1. **partitioning**

**Partitioning methods** are characterized by the fact that the number of resulting clusters *kk* must **be specified beforehand**

As already mentioned, however, the number of **clusters is usually not known**, which is why this property is often viewed as a disadvantage

Depending on the method used, the **algorithm iteratively seeks the optimum**

1. **partitioning**

The **famous k-means algorithm** belongs to the partitioning cluster method:

- *kk* cluster **centers are chosen randomly** and then
- the **sum of the squared distances** of the observations to the nearest cluster center **is minimized**
- The cluster centers are **then re-determined by averaging** and
- the observations **reassigned to the nearest clusters**
- This happens until **the assignment of observations does not change** anymore

2 - **Hierarchical** clustering methods are divided into

**2.A - agglomerative and**

**2.B - divisive methods.**

2.A - **Agglomerative** means nothing more than:

* initially treating each observation as a separate cluster.
* next, the two clusters that are closest to each other are clustered and
* the distances between all clusters are calculated again.
* this happens until **all observations are finally grouped** into a cluster.

2.B - In the **divisive method**, it is exactly the other way round:

* the starting point is one single cluster that contains **all observations**
* this cluster divided into more and more clusters during the subsequent steps
* at the end each observation will be a single cluster.

Hierarchical cluster methods can be represented graphically by a **DENDOGRAM**

A dendrogram shows at **which distance** observations are summarized (agglomerative) or separated (divisive).

2 - **Hierarchical** clustering methods are subdivided into

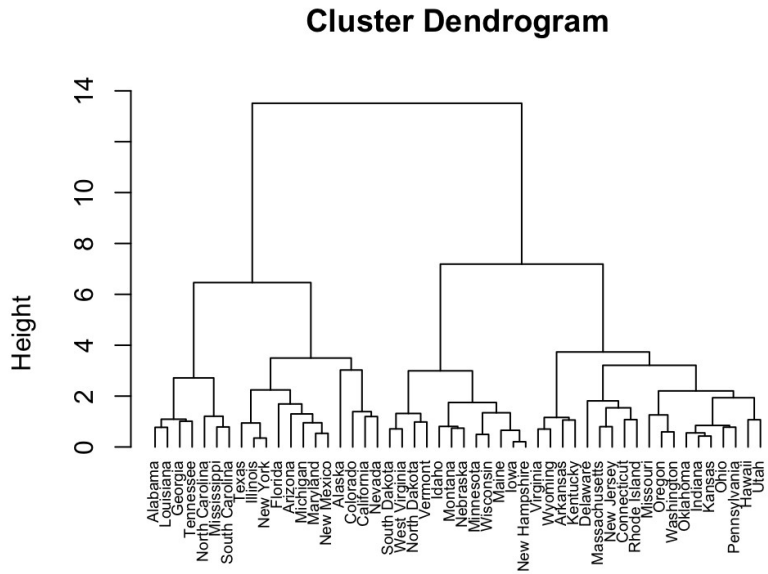
**2.A - agglomerative and**

**2.B - divisive methods.**

Ex. Of agglomerative/divisive procedure

Divisive method applied to vehicles considering the number of wheels

Ex. Of a dendogram




Dendogram constructed on a countries' food behavior dataset

## Punto di partenza

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

- Table of data: $n \times k$ (n by 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**

## CA come analisi esplorativa

- The Group Analysis or Cluster Analysis is a typical **explorative method** for the identification of clusters of similar units according to the $n_x k$-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 data table from the point of view **of the statistical units** (number of rows) → **from *n* observed statistical units to *g* homogeneous groups ($g<n$)**

## Domande centrali per eseguire una CA

Choices in CA:

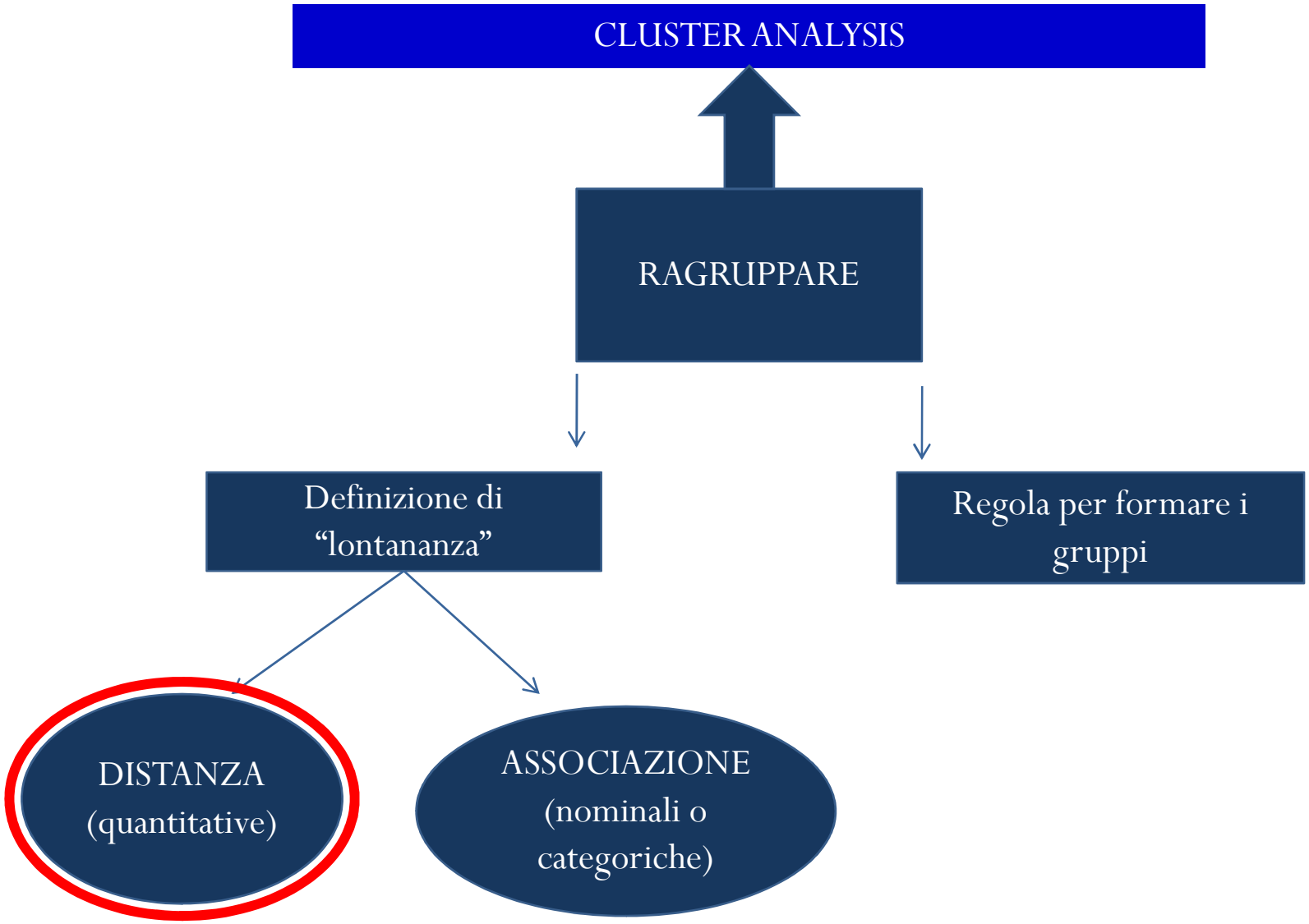
1. Which **informative variables** must be considered?

2. Which **distance or index of similarity** must be used?

3. Which **method** for the groups' definition must be applied?
   a) General criterium: internal cohesion and external separation
   b) Methods:
      - **Hierarchical method**: progressive aggregation of units
      - **Non hierarchical method**: unique partition given the number $g$ of groups

4. How to **evaluate the final partitions** and to **choose** the optimal one?

# Procedimento di definizione dei gruppi



## Misure di distanza

Partendo dalla matrice dei dati, ricordiamo che i suoi (n) vettori riga rappresentano le n unità statistiche. Ciascuna unità statistica e' quindi un vettore di p-elementi, contenenti i valori da essa assunti sulla prima, la seconda, la j-esima e la p-esima variabile.
Supponiamo che tali valori siano numeri e non attributi, ovvero supponiamo che le p variabili siano quantitative.
Possiamo definire la distanza tra due unità statistiche, i ed h, in diversi modi.

Il punto di partenza fondamentale e' la definizione di una misura di similarità o di distanza tra gli oggetti (cioe' tra le righe della matrice dei dati).
L'altro punto fondamentale e' la regola in base alla quale si formano i gruppi.

A seconda del tipo di dati, si hanno misure diverse. Per dati quantitativi si hanno misure di distanza; per dati qualitativi si hanno misure di associazione.

• 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

$$PI_{iu}=\mathrm{f}(\boldsymbol{x}_i,\boldsymbol{x}_u)$$

- *Proximity Indices:*
  - *For numeric variables*
    - ✓ *Distances*
      - ✓ Distance indices
  - *For categorical variables*
    - ✓ Similarity indices

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

1. $d(\boldsymbol{x}_i, \boldsymbol{x}_u) \geq 0$ (non negativity)
2. $d(\boldsymbol{x}_i, \boldsymbol{x}_u) = 0 \Leftrightarrow \boldsymbol{x}_i = \boldsymbol{x}_u$ (identity)
3. $d(\boldsymbol{x}_i, \boldsymbol{x}_u) = d(\boldsymbol{x}_u, \boldsymbol{x}_i)$ (symmetry)
4. $d(\boldsymbol{x}_i, \boldsymbol{x}_u) \leq d(\boldsymbol{x}_i, \boldsymbol{x}_s) + d(\boldsymbol{x}_s, \boldsymbol{x}_u)$ $\forall\ \boldsymbol{x}_i, \boldsymbol{x}_s, \boldsymbol{x}_u \in \Re^k$ (triangular inequality)

- 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)

- Properties:
  - *P1:* euclidean distance $_2d_{iu}$ is affected more strongly than Manhattan distance by great differences between pairs of values
  - *P2:* Minkowski distance $_md_{iu}$ is non increasing function of parameter *m*: $_1d_{iu} \geq_2 d_{iu} \geq \cdots \geq_\infty d_{iu}$

- 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 $\boldsymbol{T}'\boldsymbol{T} = \boldsymbol{I}$

## Example

n=2 and k=2

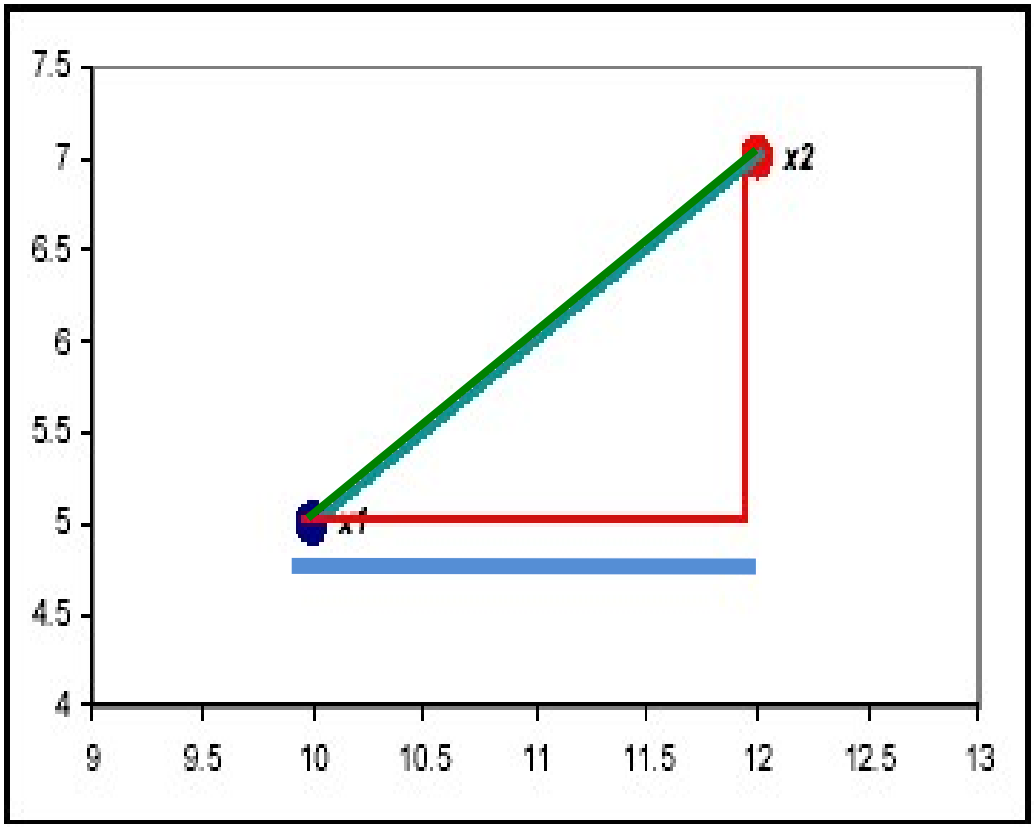
$\mathbf{x}_1=(10;5)'$

$\mathbf{x}_2=(12;7)'$

$_1d_{12} = 4$ (Manhattan)

$_2d_{12} = 2.83$ (Euclidean)

$_\infty d_{12} = 2$ (Chebichev)



Starting point of hierarchical methods:

$n \times n$ matrix of distances

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

- *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:

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

# Procedimento di definizione dei gruppi: misure di distanza

In questo caso la dissomiglianza tra unità coincide con la distanza tra le stesse. Diverse sono le forme di distanze che vengono considerate nella pratica. Sia $X$ una matrice dati $n \times k$, $X_i$ il vettore $k$-dimensionale della $i$-esima osservazione ed $x_{ih}$ il suo elemento generico. Sia inoltre $S^{-2}$ l'inversa della matrice di varianze e covarianze campionarie.

*Distanza city-block o di Manhattan*

$$d_{ij} = \sum_{h=1}^{k} |x_{ih} - x_{jh}|$$

*Distanza euclidea*

$$d_{ij} = \left[ \sum_{h=1}^{k} (x_{ih} - x_{jh})^2 \right]^{\frac{1}{2}}$$

*Distanza di Minkowsky*

$$d_{ij} = \left[ \sum_{h=1}^{k} |x_{ih} - x_{jh}|^r \right]^{\frac{1}{r}}$$

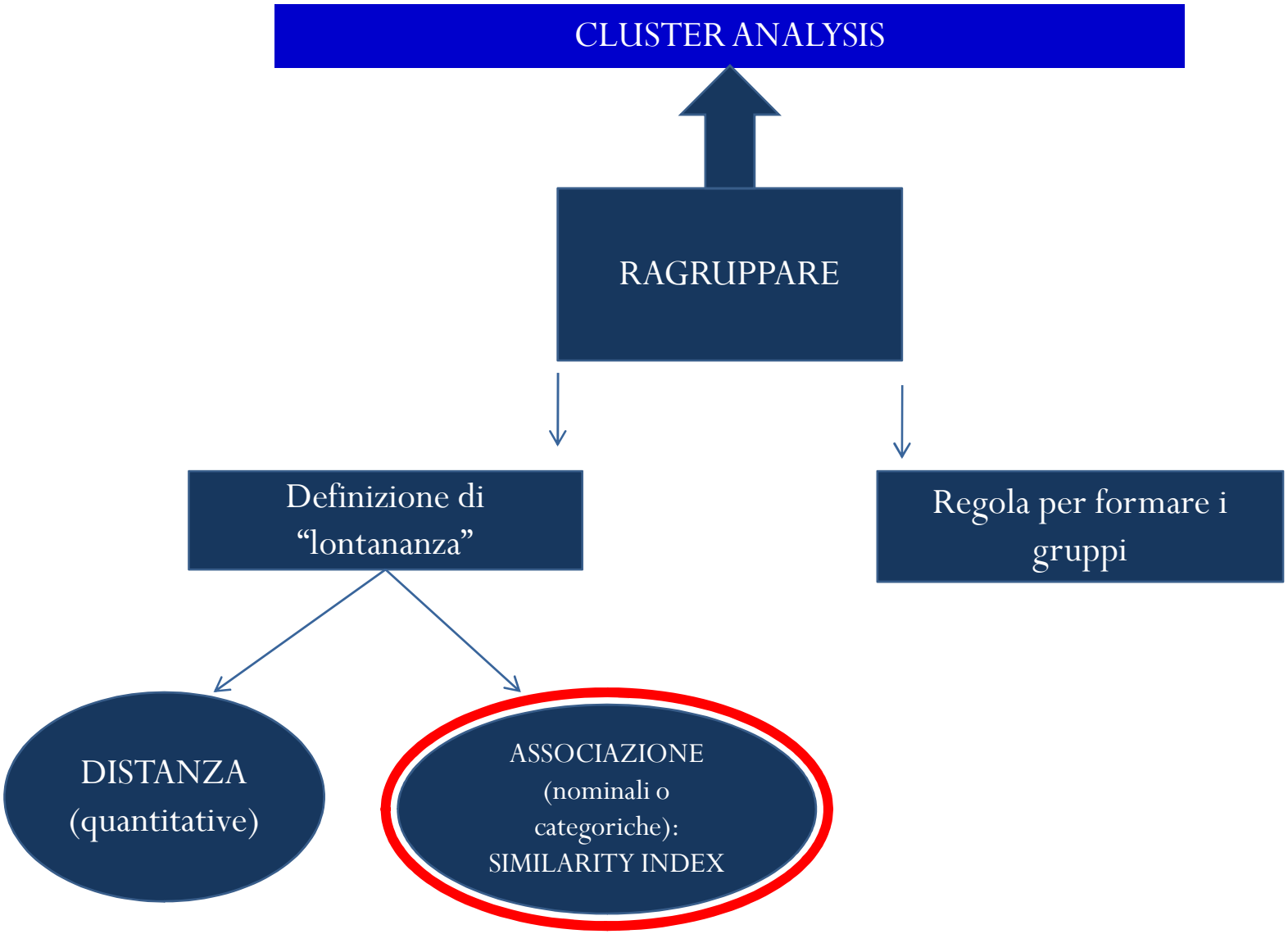
*Distanza di Mahalanobis*

$$d_{ij} = (X_i - X_j)' S^{-2} (X_i - X_j)$$

Nelle prime tre distanze:
le variabili con maggiore variabilità hanno peso maggiore nella misura della dissomiglianza tra unità.

Per evitare questo inconveniente:
preferibile usare osservazioni standardizzate
o utilizzare la distanza di Mahalanobis.

# Procedimento di definizione dei gruppi



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

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

Case of *k* dichotomous variables:

- 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* $\rightarrow \sum_{j=1}^{k} x_{ij}\left(1 - x_{uj}\right)$
  - $f_{01}$ = frequency of of characteristics present in *u* but not in *i* $\rightarrow \sum_{j=1}^{k}\left(1 - x_{ij}\right)x_{uj}$
  - $f_{00}$ = frequency of characteristics jointly absent in *i* and in *u* $\rightarrow \sum_{j=1}^{k}\left(1 - x_{ij}\right)\left(1 - x_{uj}\right)$

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)

Case of *k* categorical (not all dichotomous) variables:

- 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

## Domande centrali per eseguire una CA

Choices in CA:

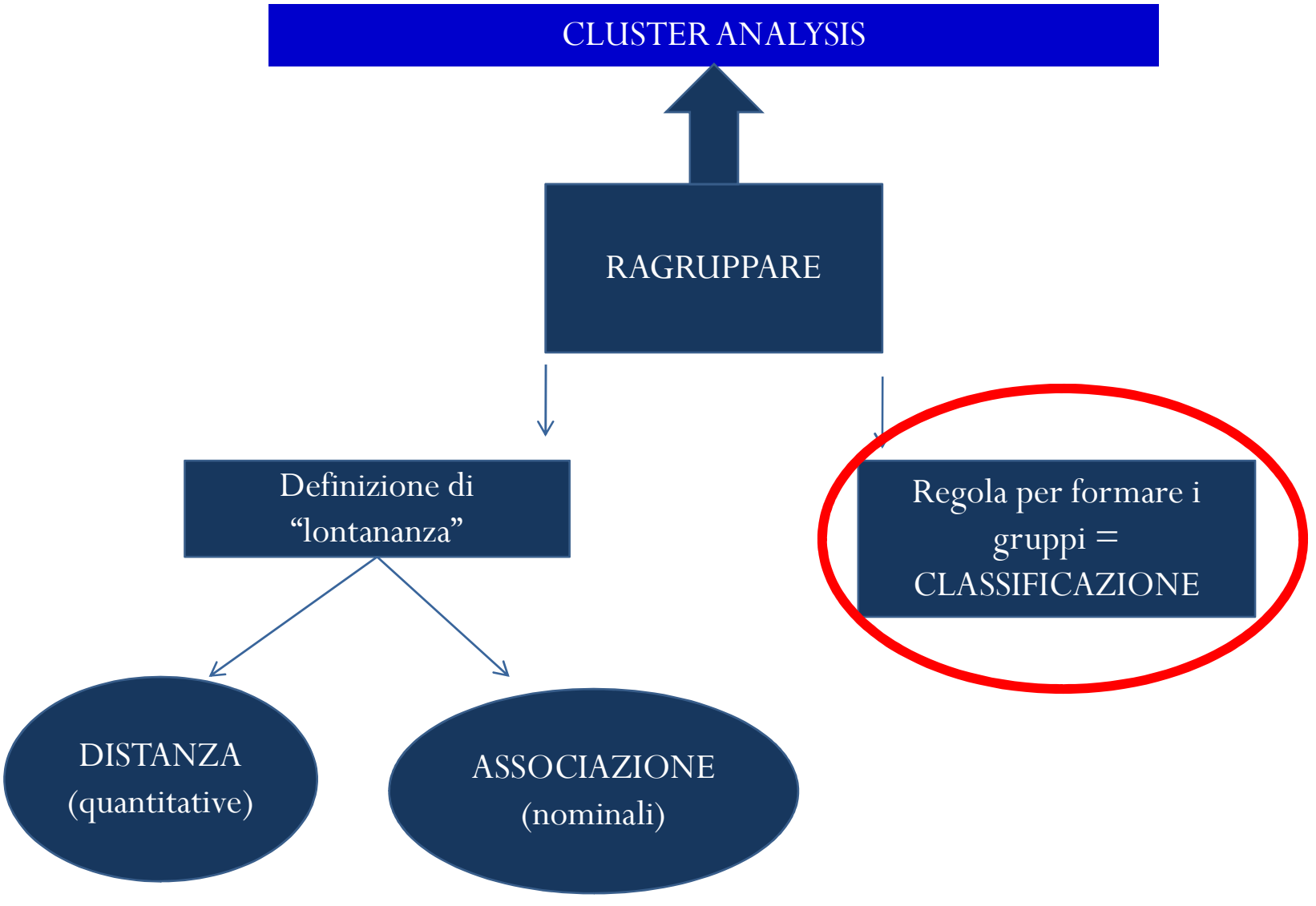
1. Which **informative variables** must be considered?

2. Which **distance or index of similarity** must be used?

3. Which **method** for the groups' definition must be applied?
   a) General criterium: internal cohesion and external separation
   b) Methods:
      - **Hierarchical method**: progressive aggregation of units
      - **Non hierarchical method**: unique partition given the number $g$ of groups

4. How to **evaluate the final partitions** and to **choose** the optimal one?

# Procedimento di definizione dei gruppi



# Procedimento di definizione dei gruppi: classificazione

## Metodi di classificazione

Effettuata la scelta della misura di diversità da utilizzare, si pone la scelta del metodo o algoritmo di classificazione e dell'eventuale criterio di aggregazione/suddivisione.

I metodi di classificazione piu' comuni sono:

1. Metodi gerarchici aggregativi
2. Metodi gerarchici divisivi
3. Metodi non gerarchici.

## Procedimento di definizione dei gruppi: classificazione

I **metodi gerarchici** realizzano fusioni o divisioni successive dei dati. Nel caso dei metodi aggregativi (o "agglomerativi") gli n oggetti iniziali vengono fusi in gruppi via via piu' ampi (alla fine: un unico gruppo); nel caso dei metodi divisivi (o "scissori") vengono definite partizioni sempre piu' fini dell'insieme iniziale (alla fine n clusters contenenti ciascuno un elemento). La caratteristica principale che li distingue dai metodi non gerarchici e' che la assegnazione di un oggetto ad un cluster e' **irrevocabile**. Ovvero, una volta che un oggetto e' entrato a far parte di un cluster, non ne viene piu' rimosso.

## Procedimento di definizione dei gruppi: classificazione

I **metodi non gerarchici** sono solo di tipo aggregativo, e producono un'unica partizione. Procedono a riallocazioni successive delle unità tra i gruppi definiti a priori, fino alla partizione giudicata "ottima" sulla base di un criterio predefinito.

# Procedimento di definizione dei gruppi: classificazione
# METODI GERARCHICI

## Metodi gerarchici aggregativi

Si suppone che l'insieme di oggetti da classificare sia dotato di una misura di dissimilarità. Immaginiamo per semplicità che si tratti di una distanza.

Si costruisce una prima matrice di distanze fra le n unità statistiche. Si aggregano le due unità piu' vicine (ovvero con distanza minima), in un cluster.

Al passo successivo una terza unità entra a far parte del cluster trovato al passo precedente, oppure, due unità vengono fuse per formare un diverso cluster.

Si continua a procedere in questo modo finche' non si sia formato un unico cluster contenente tutte le unità.

Tutto il procedimento poggia sulla definizione del criterio di assegnazione delle unità ai cluster (o di un cluster piccolo ad uno piu' grande).

- 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

# CA hierarchical methods

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

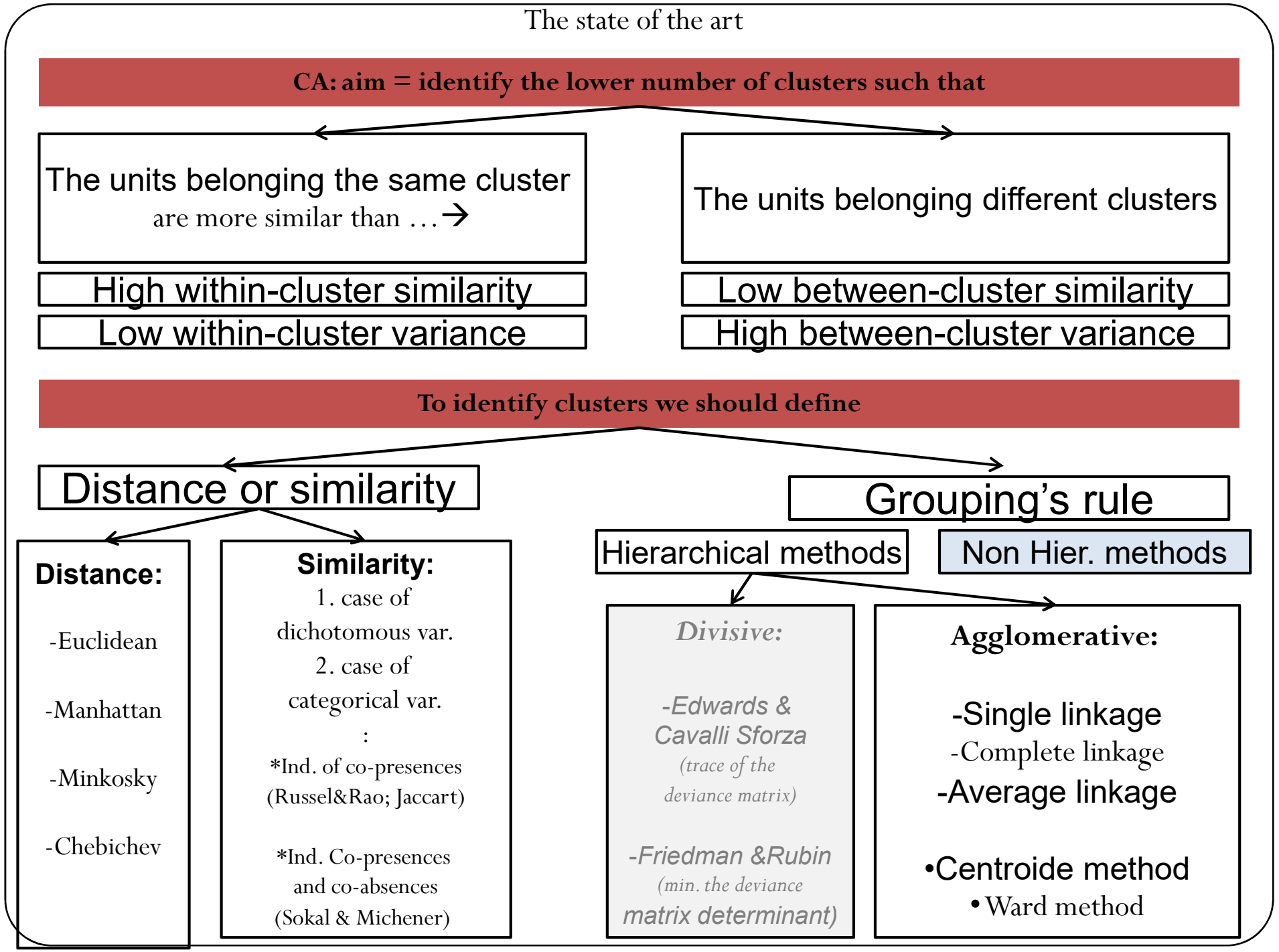
1. The two nearest units (with minimum distance or maximum proximity) are grouped

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


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
Grouping's rule
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)
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

# Procedimento di definizione dei gruppi: classificazione
# METODI GERARCHICI

Esistono diversi possibili criteri, e conseguentemente, diversi algoritmi aggregativi.
Ne vedremo alcuni:

- **Legame singolo**
- **Legame completo**
- **Legame medio**
- **Metodo del centroide.**

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

Let $C_1$ and $C_2$ be two clusters wit $n_1$ 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):
$$d(C_1,C_2)= \Sigma_{i>u}\ d_{iu} / [m(m-1)/2]\ ,\ i,u \in C_1 \cup C_2$$

Ex:

Sptep 1

$D^{(1)} =$

| | 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 |

Lower distance within our statistical units

*After the first step we cluster A and B together and we treat it as a single entity.*

Step 2

*Now we re-compute the distance matrix among pairs of our **4 elements (AB, C, D, E )***

$D^{(2)} =$

| | AB | C | D | E |
|---|---|---|---|---|
| AB | 0 | 3 | 6 | 7 |
| C | 3 | 0 | 4 | 6 |
| D | 6 | 4 | 0 | 2 |
| E | 7 | 6 | 2 | 0 |

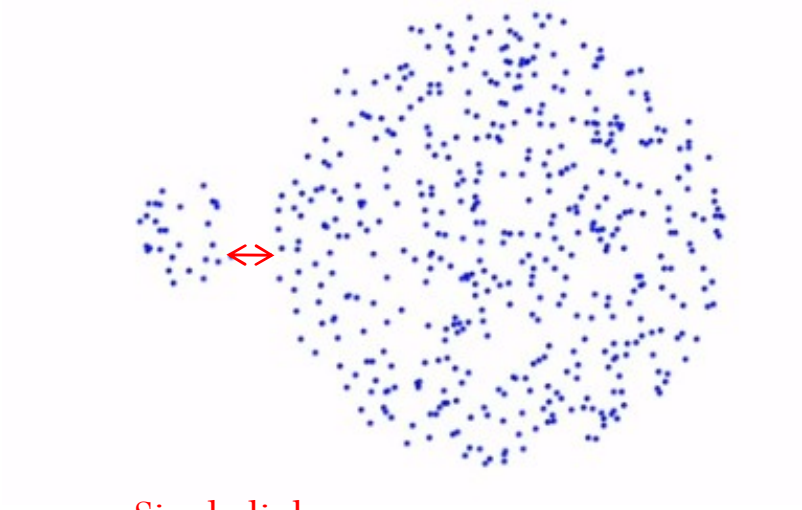
Lower distance within our statistical units

….

*After the second step we cluster D and E together and we treat it as a single entity.*

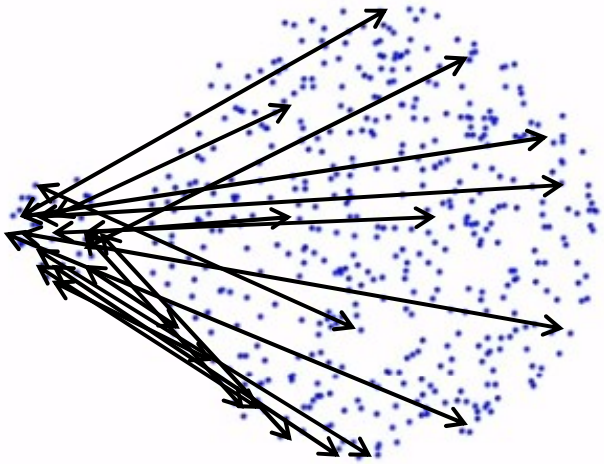
*Now we re-compute the distance matrix among pairs of our **3 elements (AB, C, DE)***

***And so on***

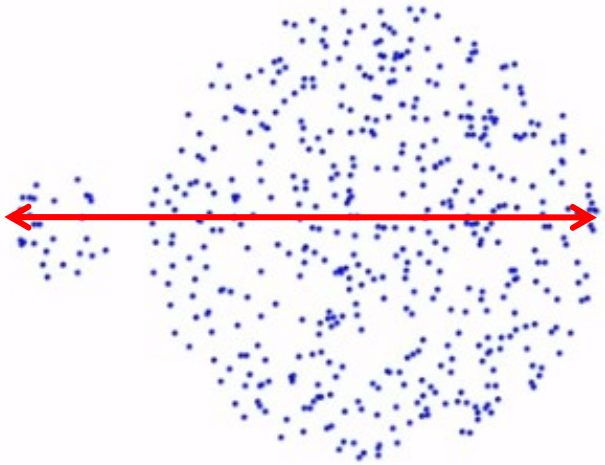


Single linkage

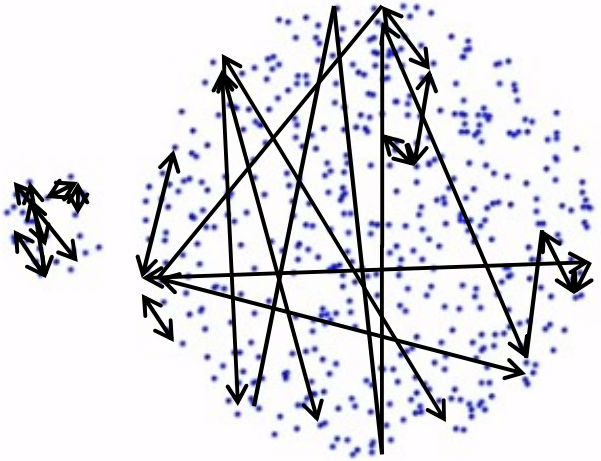
Average linkages:



…. Among all pairs



Complete linkage



# Procedimento di definizione dei gruppi: classificazione
## METODI GERARCHICI

Il metodo del **legame singolo** si basa su un criterio di distanza minima. Supponendo di avere 4 unità : A,B,C,D, e di aver definito un coefficiente di dissimilarità o una misura di distanza tra le unità ($d_{AB}$, $d_{AC}$, ……,$d_{CD}$); supponendo che le unità A e B vengano fuse in un solo cluster, la distanza tra il cluster (AB) e l'unità C e' definita come:

$$d_{(A,B)C} = \min(d_{AC}, d_{BC})$$

Posto che le unità C e D vengano fuse nel cluster (CD), la distanza tra il cluster (AB) ed il cluster (CD) viene definita come:

$$d_{(AB)(CD)} = \min(d_{AC}, d_{AD}\, d_{BC}, d_{BD})$$

Al primo passo si fondono le 2 unità aventi distanza minore, ottenendo cosi' n-1 gruppi.
Si calcola una nuova matrice di distanze tra gli n-1 clusters. Si aggregano i due cluster aventi distanza minima.
E cosi' via, fino ad avere un unico cluster contenente n unità.

## Procedimento di definizione dei gruppi: classificazione
## METODI GERARCHICI

Il metodo del **legame completo** si basa su un criterio di **distanza massima**.Ovvero, supponendo di avere 4 unità : A,B,C,D, e di aver definito un coefficiente di dissimilarità o una misura di distanza tra le unità ($d_{AB}$, $d_{AC}$, ……,$d_{CD}$); supponendo che le unità A e B vengano fuse in un solo cluster, la distanza tra il cluster (AB) e l'unità C e' definita come:

$$d_{(A,B)C} = \max(d_{AC}, d_{BC})$$

mentre la distanza tra il cluster (AB) ed il cluster (CD) viene definita come:

$$d_{(AB)(CD)} = \max(d_{AC}, d_{AD}\, d_{BC}, d_{BD}).$$

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

## Procedimento di definizione dei gruppi: classificazione
## METODI GERARCHICI

Nel metodo del **legame medio** la distanza tra cluster e' definita come media aritmetica delle distanze (o dissimilarità) tra tutte le possibili coppie di elementi appartenenti l'uno ad un cluster, l'altro ad un altro.
Dati 2 cluster A e B, contenenti, rispettivamente, $n_A$ ed $n_B$ unità , indichiamo con l'indice *i* il generico elemento del cluster A, e con l'indice *h* il generico elemento del cluster B, e con $d_{i,h}$ la loro distanza . La distanza tra A e B e' definita come:

$$d_{A,B} = 1/n_A n_B(\Sigma_i \Sigma_h d_{i,h})$$

## Procedimento di definizione dei gruppi: classificazione
## METODI GERARCHICI

Il **metodo del centroide** si applica solo a variabili quantitative e lavora non tanto sulla matrice delle distanze quanto sui singoli vettori di osservazioni. (Nel senso che ad ogni passo ricalcola la matrice delle distanze partendo non dalle distanze precedenti ma dai baricentri di ciascun cluster).
Per ogni gruppo (anche composto da una sola unità) si calcola il baricentro (o *individuo medio*, cioe' un elemento che  come modalità delle diverse variabili, presenta le modalità medie del gruppo). La distanza tra un'unità e un gruppo o tra due gruppi e' calcolata come distanza tra i baricentri.

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

- Centroid method:

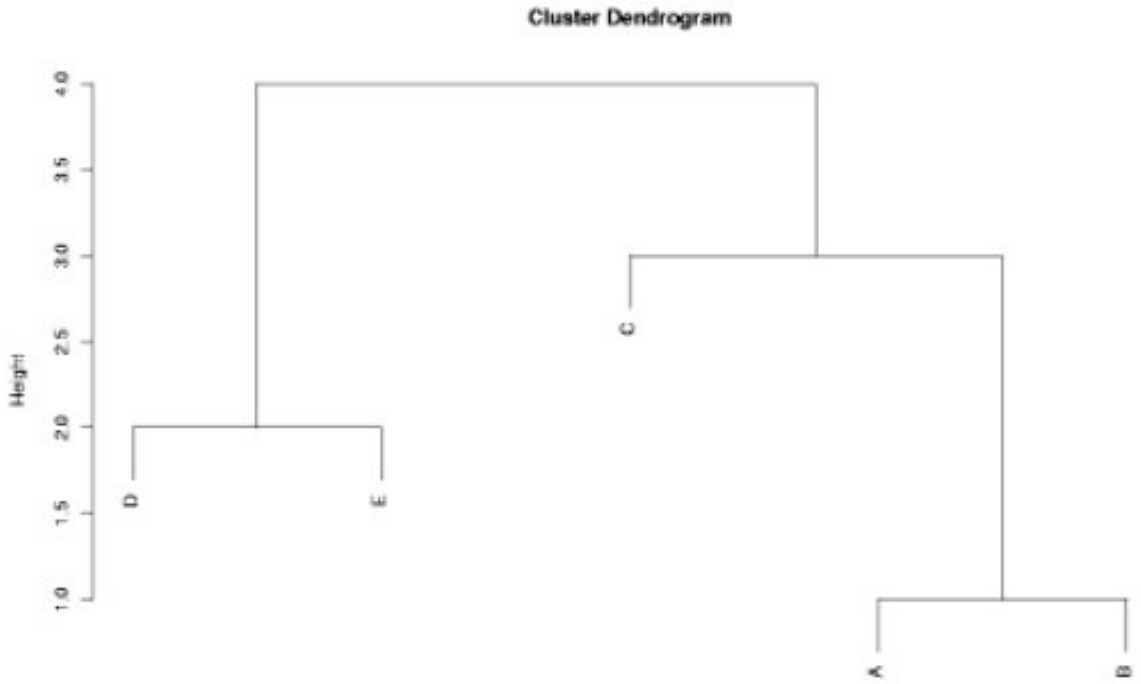
$$d(C_{1,}, C_2) = d(\bar{\boldsymbol{x}}_{1,}, \bar{\boldsymbol{x}}_2)$$

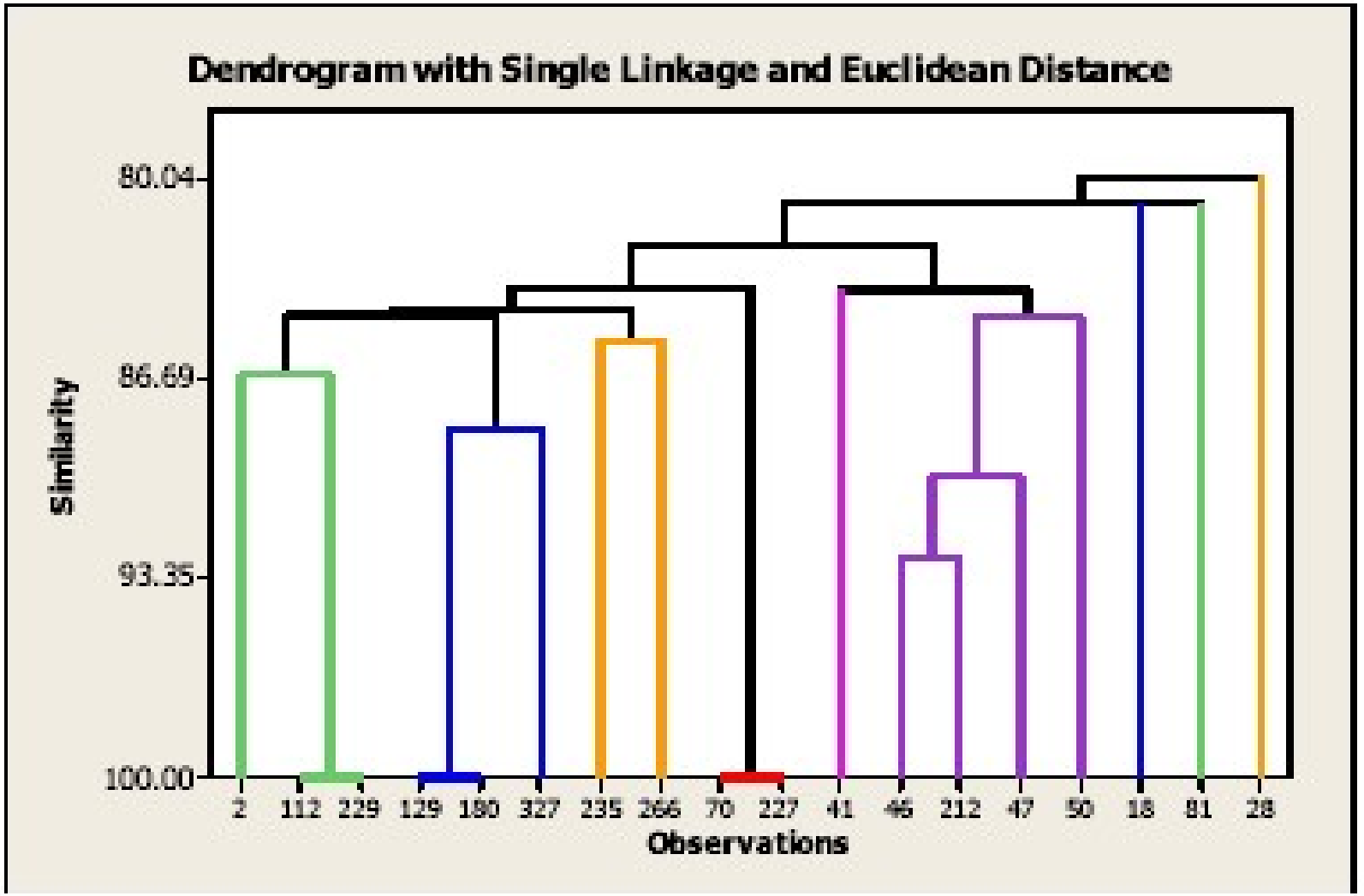
the distance between two clusters is equal to the distance between the two *k*-dimensional vectors of means computed on the $n_1$ units of $C_1$ and the $n_2$ units of $C_2$

# Procedimento di definizione dei gruppi: classificazione
# RAPPRESENTAZIONE GRAFICA

## Il dendrogramma

Il dendrogramma e' una rappresentazione grafica che visualizza secondo ordinate crescenti il livello di aggregazione delle unità o cluster.
Il dendrogramma dell'esempio precedente riguardante il metodo del legame singolo e' il seguente:




Dendrogram with Single Linkage and Euclidean Distance
Similarity
80.04
86.69
93.35
100.00
2 112 229 129 180 327 235 266 70 227 41 46 212 47 50 18 81 28
Observations

# Procedimento di definizione dei gruppi: classificazione

In sostanza visualizza l'intero processo di aggregazione ossia una gerarchia di partizioni. Una singola partizione si ottiene "tagliando" il dendrogramma ad un dato livello dell'**indice di distanza della gerarchia**.

La scelta di quanti gruppi finali ottenere si traduce nel problema: **a quale livello tagliare l'albero?**
Dato che si ha interesse ad avere il **minor** numero di gruppi con **massima** omogeneità, si cerca di tagliare "alle radici" (cioe' in basso) dell'insieme dei "rami" piu' lunghi (cioe' le verticali piu' lunghe). (Nel caso precedente potremmo forse prenderci 3 cluster: (AB), C, (DE), oppure 2: (AB), (CDE)).

# Procedimento di definizione dei gruppi: classificazione

**Definizione: partizione.**
Una partizione P(E) su un insieme E si definisce come l'insieme delle classi di E tale che:

1. Due elementi $A_i$ e $A_h$ di P(E) sono o disgiunti (cioe' $A_i \cap A_h = \varnothing$) oppure coincidenti (cioe' $A_i \cup A_h = A_i = A_h$);
2. L'unione di tutte le parti esaurisce E (cioe' $A_1 \cup A_2 \cup \ldots \cup A_k$).

**Definizione: gerarchia**
Una sequenza (discendente) di partizioni $(P_1,\ldots..,P_s)$ di un insieme E, forma una gerarchia se e solo se per ogni $P_q$ e $P_s$, con $s > q$, ogni elemento $A_i$ di $P_q$ e' contenuto o coincide con un elemento $A_h$ di $P_s$. (Il che in sostanza significa che gli elementi di $P_q$ vengono aggregati tra loro per arrivare alla partizione successiva.)

# Procedimento di definizione dei gruppi: classificazione

**Il metodo di Ward**

Anche questo metodo puo' essere utilizzato come algoritmo gerarchico aggregativo.

Tale metodo e' diretto alla minimizzazione della varianza all'interno dei gruppi.(Pertanto **puo' essere utilizzato solo per variabili quantitative**). Ad ogni passo questo algoritmo tende ad ottimizzare la partizione ottenuta tramite l'aggregazione di due elementi.

Una partizione si considera tanto migliore quanto piu' le classi risultano omogenee al loro interno e differenti l'una dall'altra. In altri termini, quanto più e' elevata la varianza **tra** le classi, e bassa la varianza **interna** (alle classi). E' noto che la varianza totale di un insieme di unità , si puo' scomporre nella somma di due quantità: varianza **interna** (ai cluster) e varianza esterna (cioe' **tra** i cluster). In maniera analoga si scompone la matrice di varianze e covarianze S.

In simboli:

$$S = S_W + S_B$$

Dove S e' la matrice di varianze e covarianze totali;
$S_W$ e' la matrice delle varianze e covarianze "interne";
$S_B$ e' la matrice delle varianze e covarianze "esterne".

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

# Procedimento di definizione dei gruppi: classificazione
# METODI GERARCHICI SCISSORI

## Metodi gerarchici scissori

Tali metodi definiscono partizioni sempre piu' fini dell'insieme iniziale; si suddivide progressivamente l'insieme E in un numero sempre crescente di sottoinsiemi, fino ad ottenere tutti i suoi elementi distinti.
Tali metodi si basano sulla partizione di un insieme in due sottoinsiemi, e sulla suddivisione delle classi precedentemente ottenute, sempre e soltanto in ulteriori bipartizioni.
Anche questi metodi di solito partono dalla scomposizione della devianza, in particolare della matrice di devianze e codevianze (di solito indicata con T).
Sia T = W +B.

## Procedimento di definizione dei gruppi: classificazione
## METODI GERARCHICI SCISSORI

Il metodo di **Edwards e Cavalli-Sforza** sceglie come funzione obiettivo (da minimizzare) la traccia della matrice W. Ad ogni step si effettua la bipartizione che minimizza la traccia della matrice W, cioe' la somma delle devianze interne).

Il metodi di **Friedman e Rubin** ha come obiettivo la minimizzazione del determinante di W. Ma anche – in una variante – la traccia della matrice $BW^{-1}$ .

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 …→

High within-cluster similarity

Low within-cluster variance

The units belonging different clusters

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

# Procedimento di definizione dei gruppi: classificazione
# METODI NON GERARCHICI

## Metodi non gerarchici

Caratteristiche:

- Sono di solito algoritmi aggregativi e producono una sola partizione.
- Ad ogni passo dell'algoritmo rimettono in discussione la partizione ottenuta. Le classi ottenute ad ogni iterazione intermedia vengono infatti cancellate e il processo di aggregazione ricomincia, a partire dai nuovi centri.
- L'inizializzazione del processo di classificazione e' necessariamente data da una qualche scelta di un insieme di g centri iniziali.

## Procedimento di definizione dei gruppi: classificazione
## METODI NON GERARCHICI

Mentre nel caso dei metodi gerarchici l'algoritmo cerca, ad ogni passo, la migliore scissione o aggregazione tra cluster, nel caso dei metodi non gerarchici l'algoritmo partiziona le unità in un numero predefinito di gruppi basandosi sulla ottimizzazione di un qualche criterio (predefinito).

A differenza dei metodi gerarchici, **l'assegnazione di un oggetto ad un cluster non e' irrevocabile**.

## Procedimento di definizione dei gruppi: classificazione
## METODI NON GERARCHICI

esistono diversi algoritmi.
Differiscono tra loro nei seguenti aspetti:

1. come sono inizializzati i centri di partenza;
2. come gli elementi vengono assegnati ai diversi centri;
3. come alcune o tutte le unità vengono eventualmente riassegnate ad un diverso gruppo.

Di solito, scelta una partizione iniziale, si cerca di migliorarla in funzione del criterio di minimizzazione della varianza interna.

## Procedimento di definizione dei gruppi: classificazione
## METODI NON GERARCHICI

# Due metodi principali

1) Aggregazioni dinamiche

Nel metodo delle aggregazioni dinamiche si scelgono g centri provvisori tramite estrazione casuale dalle n unità. Si decide la regola di stop fissando una soglia alla differenza tra $W_{t-1} - W_t$ (differenza tra varianze interne)

2) K-means

Nel metodo k-means si assumono come centri provvisori i primi k individui. Si allocano via via le n-k unità e ad ogni assegnazione si ricalcola subito il centroide del gruppo che si e' modificato. In tal modo si accelera il miglioramento della classificazione. Si calcola la varianza interna e si passa allo step successivo prendendo i baricentri dei gruppi appena ottenuti. La regola di stop si basa sulla differenza tra $W_{t-1} - W_t$ (differenza tra varianze interne)

## Domande centrali per eseguire una CA

Choices in CA:

1. Which **informative variables** must be considered?

2. Which **distance or index of similarity** must be used?

3. Which **method** for the groups' definition must be applied?
   a) General criterium: internal cohesion and external separation
   b) Methods:
      - **Hierarchical method**: progressive aggregation of units
      - **Non hierarchical method**: unique partition given the number $g$ of groups

4. How to **evaluate the final partitions** and to **choose** the optimal one?

# Fasi processo analisi cluster

1. Scelta delle unità di osservazione;
2. Scelta delle variabili; — Operazioni preliminari
3. Omogeneizzazione scale di misura;
4. Scelta della **misura di similarità** o diversità tra unità statistiche;
5. numero di gruppi; — Costruzione dei gruppi
6. Scelta del **criterio di raggruppamento**;
7. Scelta dell'**algoritmo di classificazione** ;
8. Interpretazione dei risultati ottenuti.

# 8. Interpretazione dei risultati ottenuti (fase finale)

## Numero di cluster

La scelta del numero dei cluster k deve essere ben ponderata. In genere è compreso tra un limite minimo di 3-5 (per evitare cluster troppo ampi) ed un massimo (cluster troppo piccoli vanificherebbero il processo di semplificazione che è alla base del raggruppamento).

In generale, più piccoli sono i cluster più informativa risulta la collocazione di una unità rispetto alla tipologia descritta dal clsuter

D'altra parte, classi numerose possono dar luogo a profili irregolari complicando l'uso della cluster nelle applicazioni.

Il numero delle classi è perciò un compromesso tra esigenze contrastanti: da un lato c'è la necessità di trascurare i dettagli non necessari, dall'altro si vuole evitare la perdita di informazioni preziose.

Occorreranno diverse prove prima di pervenire ad una scelta soddisfacente.

Criteria for evaluating the partitioning:

Let $C_1$ and $C_2$ be two clusters wit $n_1$ 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**

## How Many Clusters?

- There is no "right" number of clusters
- In order to get a feeling about **how many clusters make sense**, a **screeplot** is recommended:
  - a screeplot shows the **number of clusters kk on the x-axis** and the **variance <u>within</u> the clusters on the y-axis** (which should be reasonably low).
  - The kk at which we see a kink (so-called **elbow**), is usually used, because **additional clusters hardly contribute to the reduction (explanation) of the variance.**

# 8. Interpretazione dei risultati ottenuti (fase finale)

## Numero di cluster/2

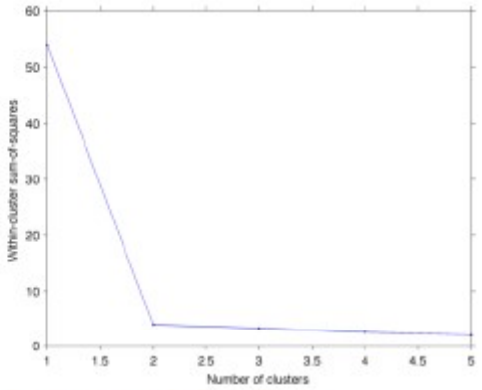
L'impiego delle tecniche gerarchiche è già di per sé una dichiarazione di incertezza sul numero dei gruppi.

Il problema è di capire quando fermare lo sviluppo dell'albero.

Se si intravvede uno spazio molto grande tra un livello ed un altro è lì che bisogna cercare il numero di gruppi ottimale in quanto qui forse si mischiano gruppi molto diversi.

In questo senso può essere d'aiuto un grafico che abbia sulle ascisse il numero di gruppi k e sulle ordinate il livello di aggregazione necessario per formare il k-esimo gruppo.

Un marcato appiattimento della curva, individuato con ampie riserve di soggettività, indicherà il valore giusto di k.




```
function wss = plotScree(X, n)
wss = zeros(1, n);
wss(1) = (size(X, 1)-1) * sum(var(X, [], 1));
for i=2:n
   T = clusterdata(X,'maxclust',i);
   wss(i) = sum((grpstats(T, T, 'numel')-1) .*
sum(grpstats(X, T, 'var'), 2));
end
hold on
plot(wss)
plot(wss, '.')
xlabel('Number of clusters')
ylabel('Within-cluster sum-of-squares')
```

a screeplot shows the **number of clusters kk on the x-axis** and the **variance <u>within</u> the clusters on the y-axis** (which should be reasonably low).

The kk at which we see a kink (so-called **elbow**), is usually used, because **additional clusters hardly contribute to the reduction (explanation) of the variance.**

# 8. Interpretazione dei risultati ottenuti (fase finale)

## VERIFICA DELLA SIGNIFICATIVITA'

Poiché raramente nell'analisi dei gruppi si conosce il numero dei raggruppamenti da individuare, esistono dei criteri per la determinazione dello stesso.

*Verifica della significatività del raggruppamento* (Beale,1969).

*Verifica della significatività degli autovalori.* Per le tecniche non gerarchiche

Nb. Indici costruiti sulla struttura del dendogramma

# Esercitazione in R su cluster analysis

Le fasi del processo di analisi dei gruppi si concretizzano in una serie di decisioni da prendere in merito a diverse scelte.

In particolare:

a) scelta delle entità di analisi;

b) scelta delle variabili caratterizzanti ciascuna entità;

c) omogeneizzazione delle scale di misura utilizzate per esprimere le diverse caratteristiche considerate;

d) scelta della misura di dissimilarità o di distanza tra le entità;

e) definizione del numero di gruppi che si vogliono o di debbono formare;

f) scelta dell'algoritmo di classificazione;

g) interpretazione dei risultati ottenuti;

h) Validazione dei risultati