University of Ferrara

DIPARTIMENTO DI ECONOMIA E MANAGEMENT

Statistics for Economics and Business

Stefano Bonnini & Valentina Mini

Cluster Analysis: Introduction and hierarchical methods

Lecture 9 – 20th of March 2019

INTRODUCTION

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

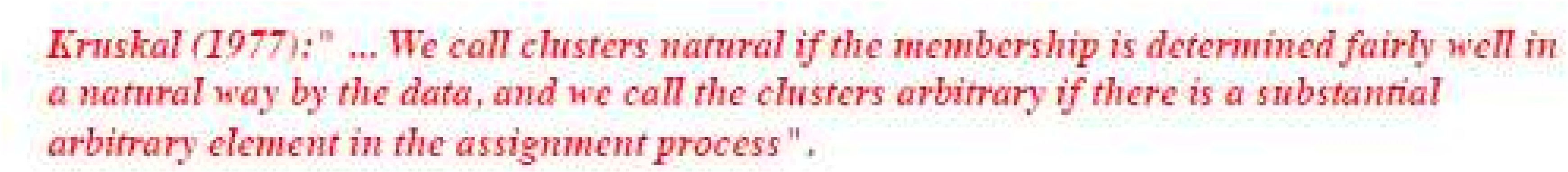
- The groups are called CLUSTERS and are usually not known a priori

- 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

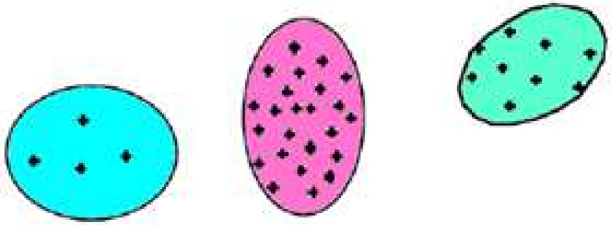
INTRODUCTION

Natural or arbitrary cluster?

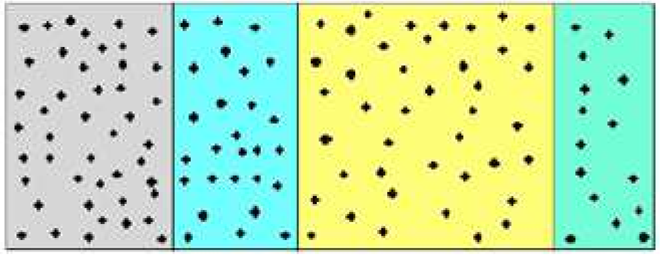


Kruskal (1977):" ... We call clusters natural if the membership is determined fairly well in a natural way by the data, and we call the clusters arbitrary if there is a substantial arbitrary element in the assignment process".

Natural clusters

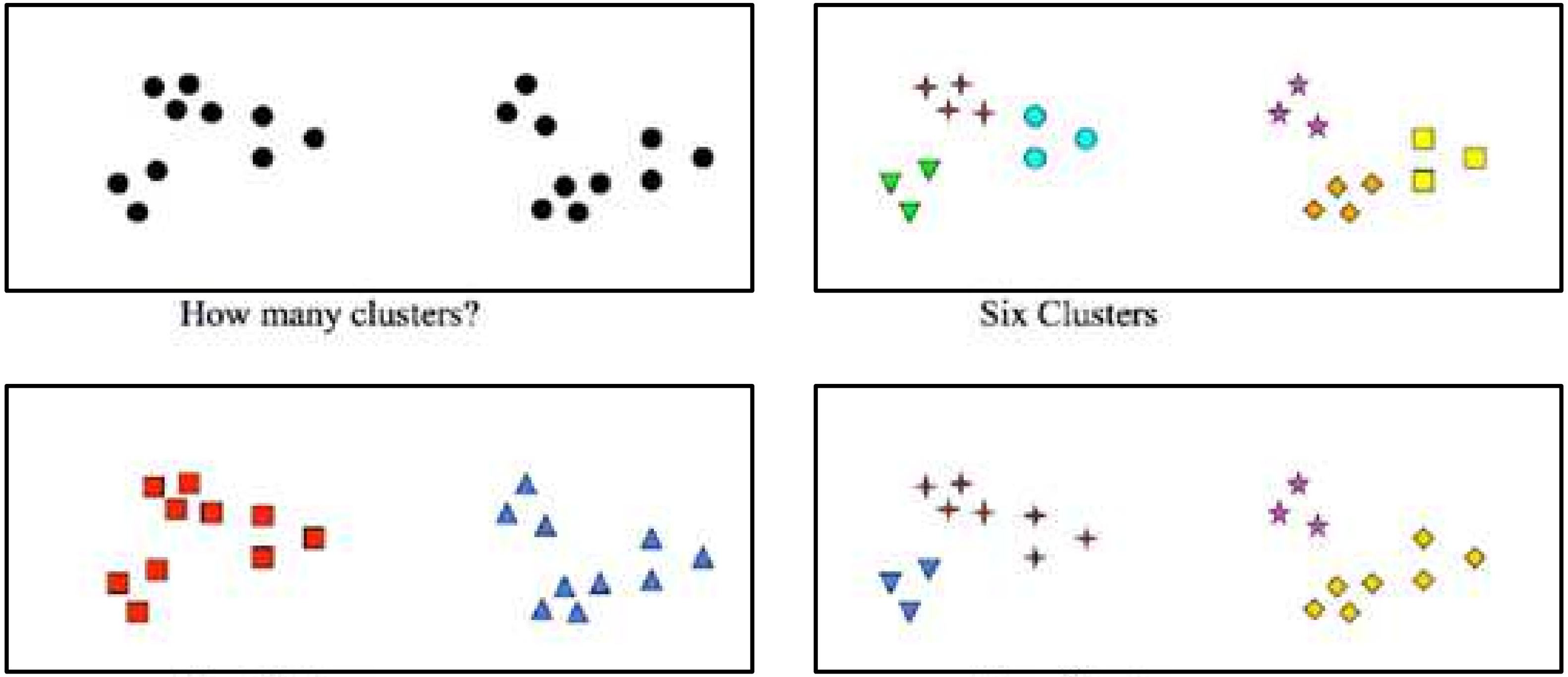


Arbitrary clusters



INTRODUCTION

How many clusters?



How many clusters?

Six Clusters

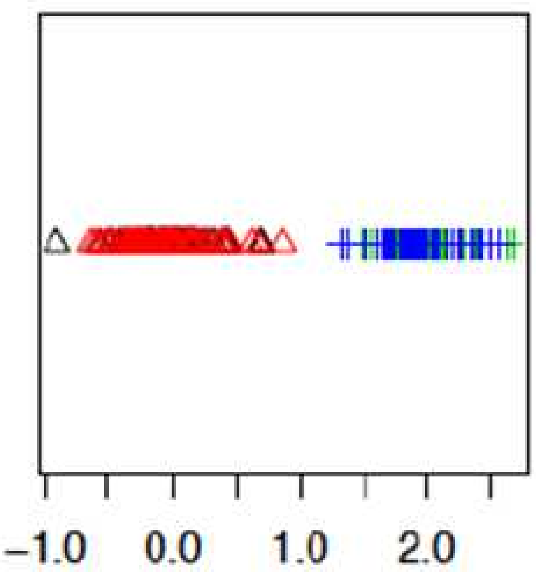
Two Clusters

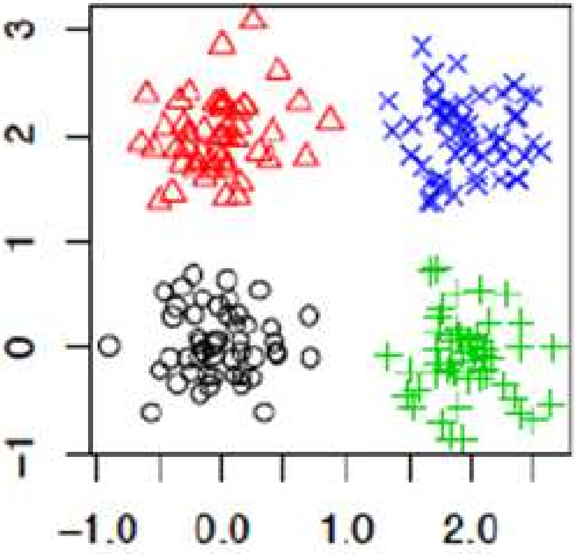
Four Clusters

INTRODUCTION

How many clusters?

Plotting data in a right way





The univariate plot is unable to show all the clusters

INTRODUCTION

- 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

INTRODUCTION

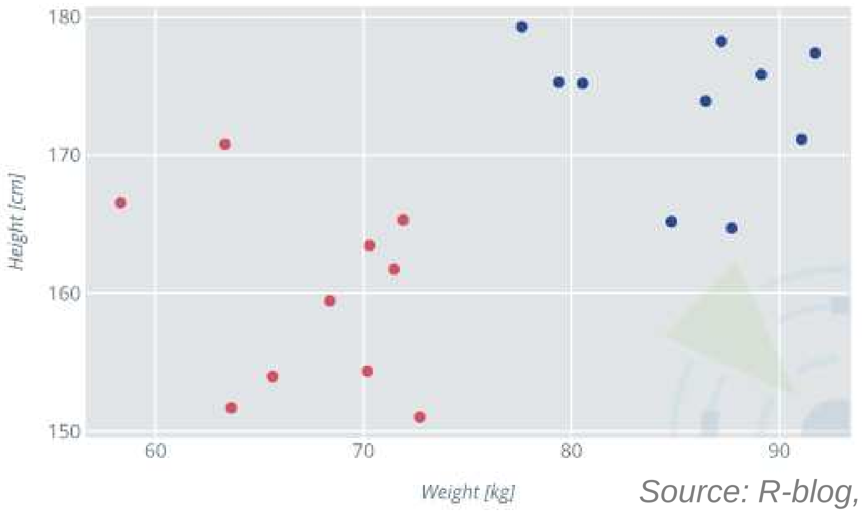
In-class experiment:

Anonymous data collection about your weight (in kilograms) and your height (in centimeters)

ID	Weight (kg)	Height (cm)	sex
1	60	167	F
2	80	187	M
…	…	…	…

INTRODUCTION

- The following scatter plot shows the “in class” simple example (9017).
- It visualizes the **weight** in kilograms on the x-axis and the **height** in centimeters on the y-axis for 20 people.
- The cluster analysis **identifies two groups** that have been marked in red and blue.
- The **interpretation** turns out to be simple in this example: the clusters just correspond to gender (red: women, blue: men).



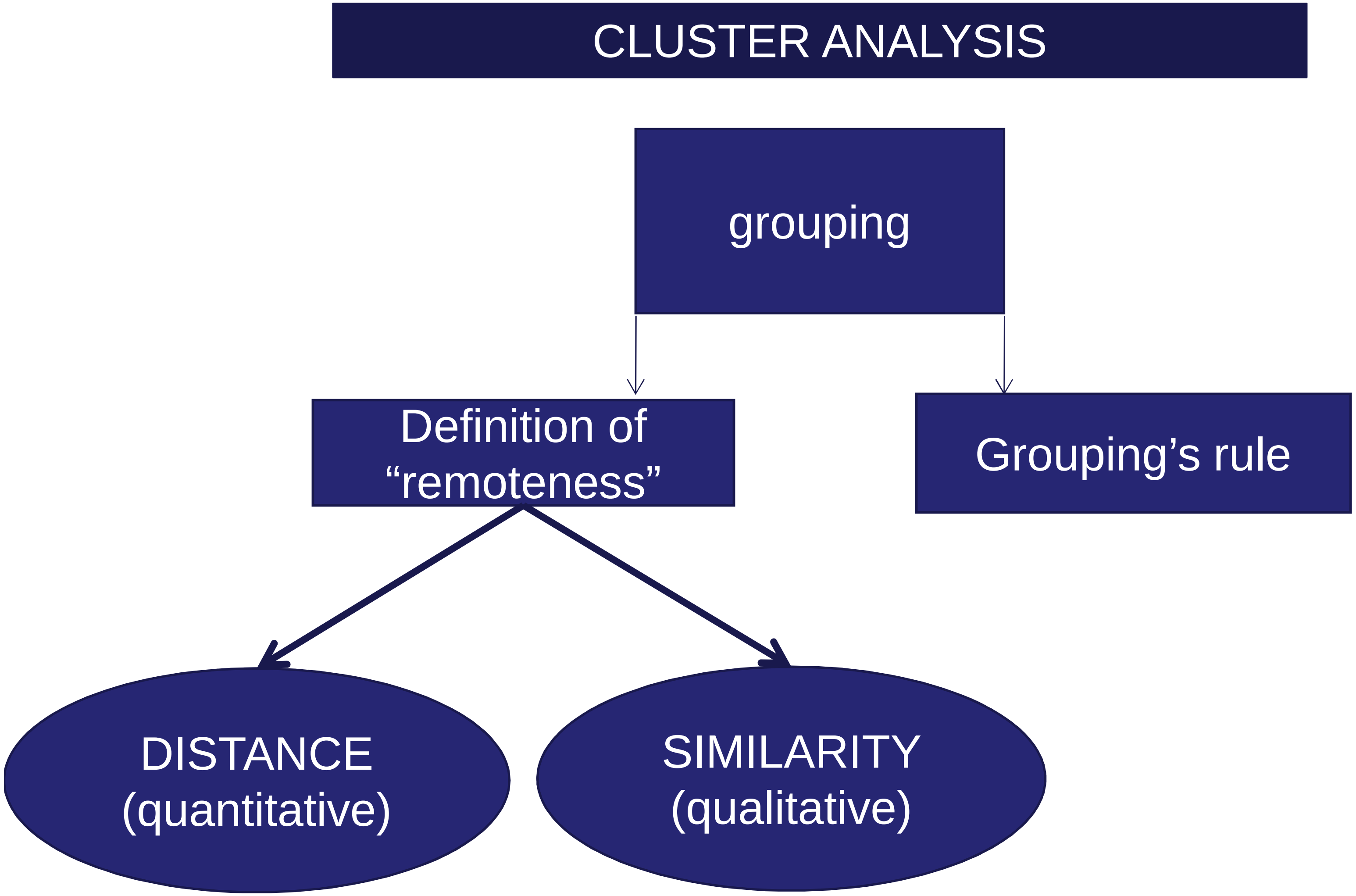
Source: R-blog, 2019

Lab using R

- Create a database in Excel (.xlsx) using the in-class collected data about weight and height
- Save the database in .csv format
- Open R
- Change the directory (selecting the one in which you saved your wh.csv database)
- Check whether R is choosing the right directory (getwd)
- Read the wh database
- Plot your sampled data and comment the graph

INTRODUCTION

- Central concepts in cluster analysis: distance or similarity
- It is central to make a decision about the **measure before the analysis**.



Methods of Cluster Analysis

Clustering methods can be divided into two large groups:

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

1- 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 of an objective function by swapping the observations back and forth between the given clusters.

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)

Methods of Cluster Analysis

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.

Hierarchical cluster methods can be represented graphically by a **dendrogram**.

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

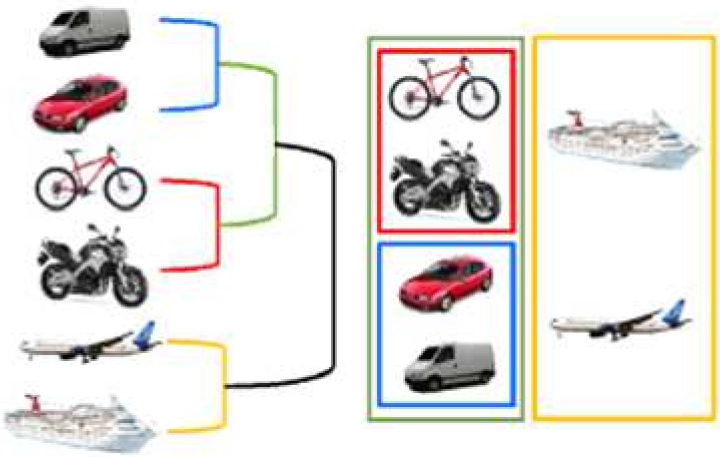
Methods of Cluster Analysis

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

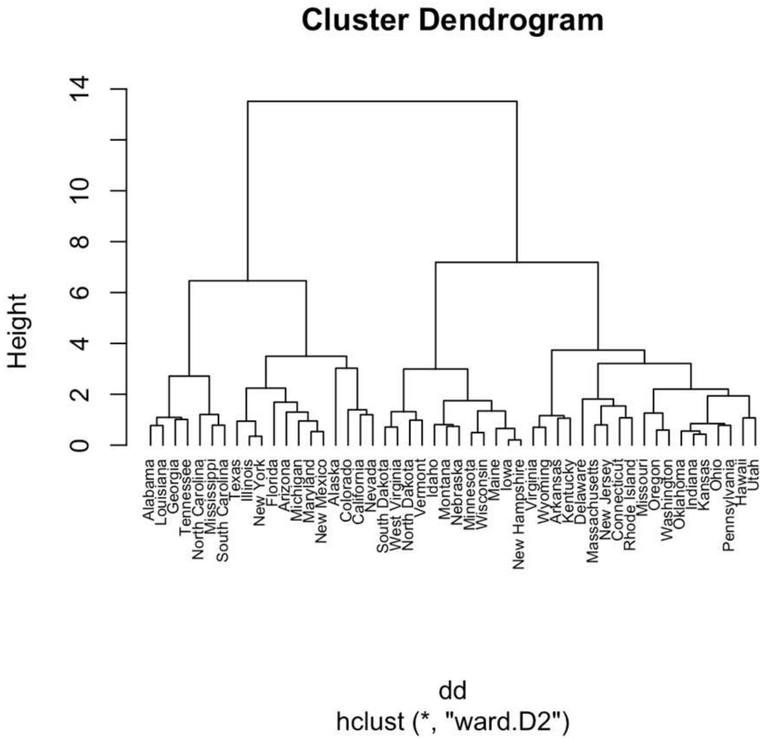
2.A - agglomerative and

2.B - divisive methods.

Ex. Of agglomerative/divisive procedure



Ex. Of a dendogram



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 within 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 of the variance.

Starting point

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

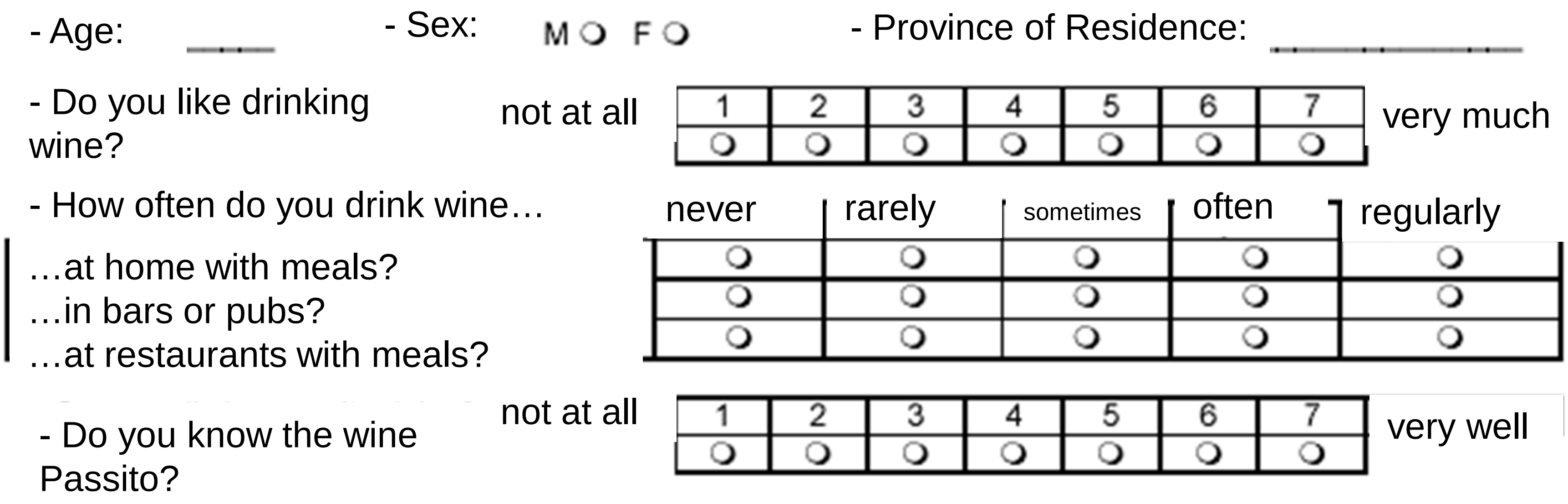
- Table of data: $n \times k$ (n by k) matrix $\boldsymbol{X}=[x_{ij}]$
  - x_{ij} = value of X_j observed on unit *i*
  - $i=1,\dots,n$
  - $j=1,\dots,k$

- 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

Starting point

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
		○	○	○	○	○	○	○	

Starting point

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

Starting point

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	
…		…	…	…	…	…	…	…	…	

Exploring and reducing the dataset

- 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 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*)**

Central questions to perform 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?

1. Which variables must be considered?

We should start form our research questions:

Ex. *: segmentation of the market of wine drinkers by the identification of homogeneous groups of customers*

-***Statistical units****: interviewed customers*
-***Variables****: behavior or preferences about alcohol consumption*

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

- 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

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

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

- *Proximity Indices:*

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

  - For categorical variables
    - ✓ Similarity indices

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

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

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

- 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)

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

- 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*: $_1d_{iu} \geq_2 d_{iu} \geq \cdots \geq_\infty d_{iu}$

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

- 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 $a_jx_{ij}+c_j,\ i=1,\cdots,n, j=1,\cdots,k$. 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 $_{2}d(\boldsymbol{Tx}_i,\boldsymbol{Tx}_u) = {}_{2}d(\boldsymbol{x}_i,\boldsymbol{x}_u)$ with $\boldsymbol{T}$ $k\times k$ matrix such that $\boldsymbol{T}'\boldsymbol{T}=\boldsymbol{I}$

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

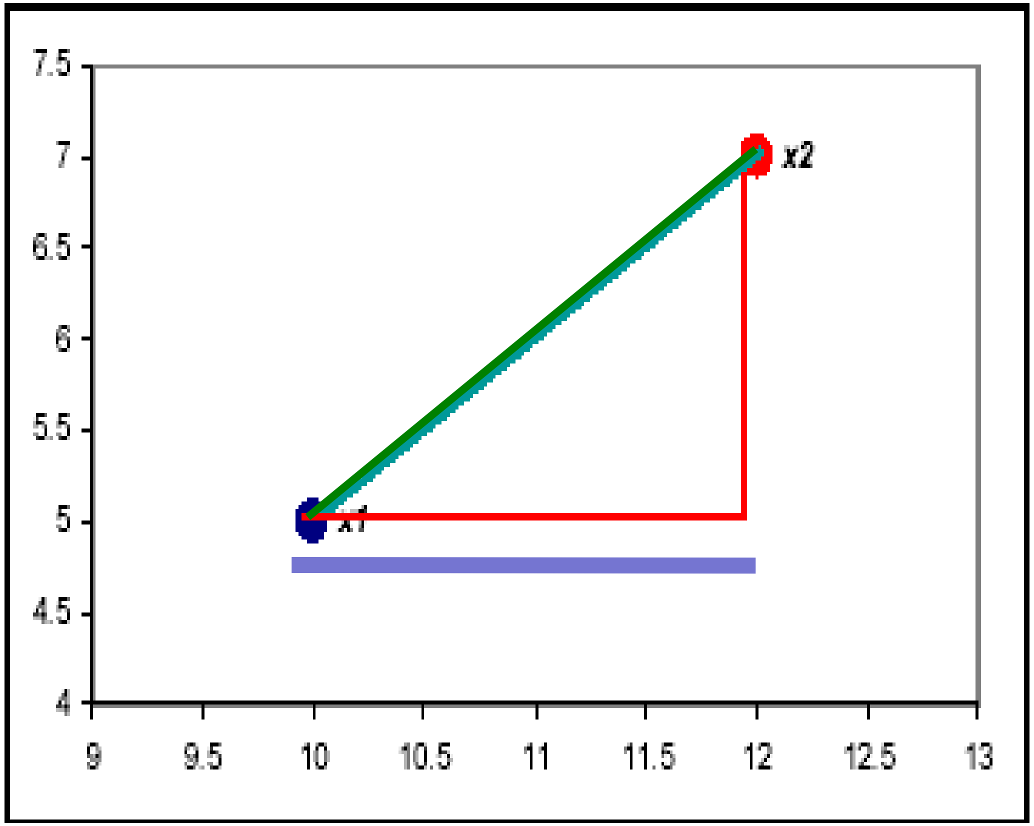
Example

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

${}_1\mathbf{d}_{12} = 4$ (Manhattan)

${}_2\mathbf{d}_{12} = 2.83$ (Euclidean)

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



3. Which method for the groups definition must be applied?

Starting point of hierarchical methods:

$n \times n$ matrix of distances

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

3. Which method for the groups definition must be applied?

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

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

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

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

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

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

3. Which method for the groups definition must be applied?

- 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)

3. Which method for the groups definition must be applied?

- Similarity index (for **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)

3. Which method for the groups definition must be applied?

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* → $\sum_{j=1}^{k} x_{ij}\, x_{uj}$
  - f_{10} = frequency of of characteristics present in *i* but not in *u* → $\sum_{j=1}^{k} x_{ij}\left(1 - x_{uj}\right)$
  - f_{01} = frequency of of characteristics present in *u* but not in *i* → $\sum_{j=1}^{k}\left(1 - x_{ij}\right)x_{uj}$
  - f_{00} = frequency of characteristics jointly absent in *i* and in *u* → $\sum_{j=1}^{k}\left(1 - x_{ij}\right)\left(1 - x_{uj}\right)$

3. Which method for the groups definition must be applied?

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-absences:

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

(index of simple correspondence)

3. Which method for the groups definition must be applied?

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