Segmentation

It is a recursive statistical technique that allows to split n-units into groups, in function of a divisive criterion, which aims to maximize the internal homogeneity of the groups obtained.

Segmentation techniques

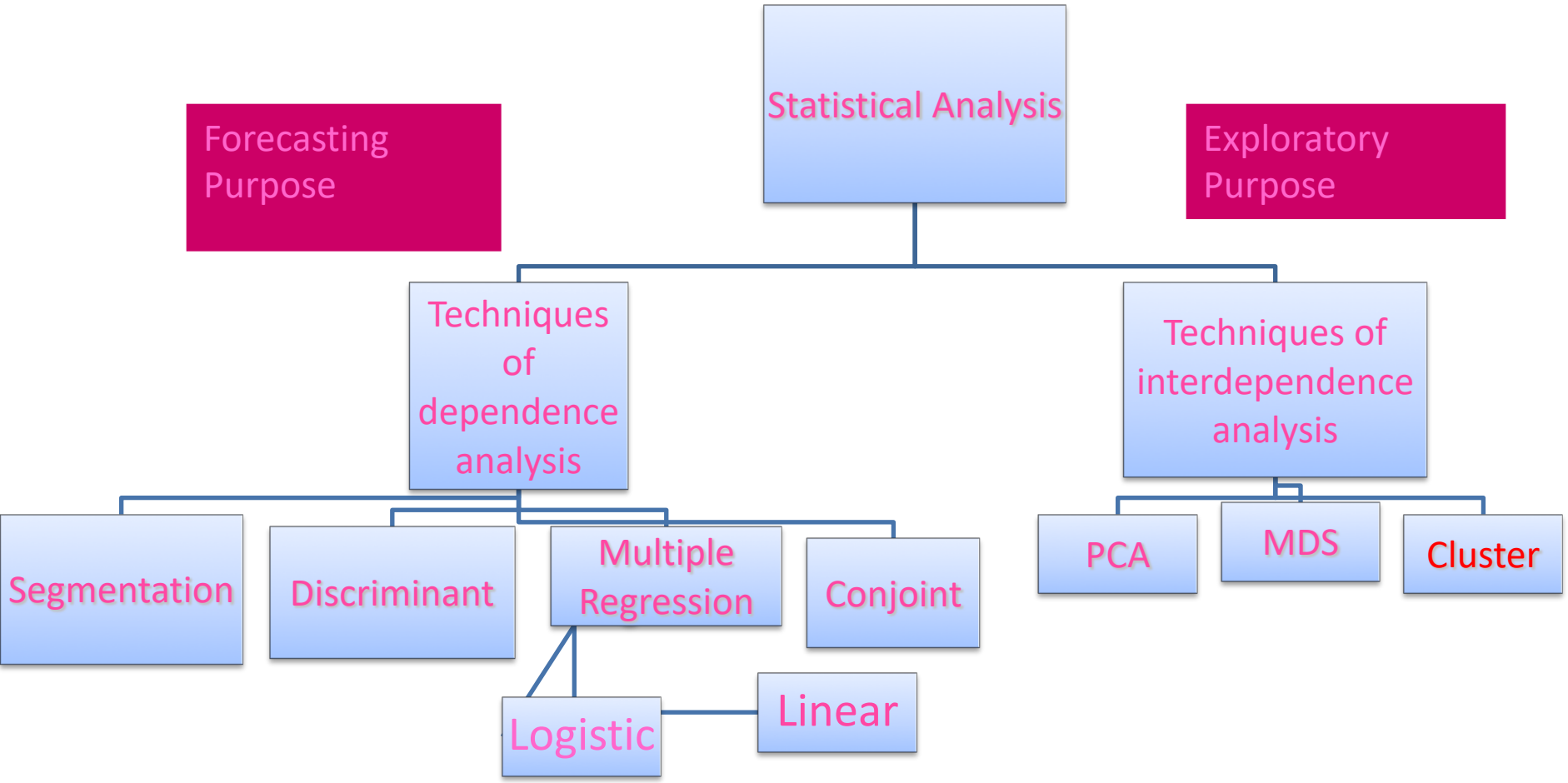
- **Homogenous (a posteriori)**

  →
  - Cluster Analysis (classic)
  - Conjoint Analysis (flexible)

- **Target (a priori)**

  →
  - AID, CHAID, CART, QUEST
  - Discriminant Analysis
  - Multiple e Logistic Regression
  - Neural Networks

Main quantitative analysis techniques



Market segmentation
Cluster Analysis

Cluster Analysis (CLA)

Cluster analysis is, essentially, a set of multivariate techniques aim to explore data allowing to group objects based on the characteristics they possess.

Cluster analysis classifies the objects (e.g., respondents, products, or other entities) so that each object is very similar to the others in the group (cluster) with respect to some predetermined selection criterion.

This grouping criterion is based on the assumption that there are latent groups, so to say " natural ", among the cases. The resulting clusters of objects should then exhibit high internal (whithin – cluster) homogeneity and high external (between – clusters) heterogeneity.

Cluster analysis

Methods:

- Hierarchical
  - agglomerative
  - divisive
- Non-Hierarchical

Hierarchical

The segmentation technique is based on a logic of minimization of the distances between the statistical units within the groups, and of maximization of the distances between groups

It starts from n groups for which the distances are known, represented by the elements of a matrix, denoted as D.

$$D = \begin{bmatrix} 0 & d_{12} & d_{13} \ldots d_{1n} \\ & 0 & d_{23} \ldots d_{2n} \\ & & \ldots \ \ldots \ \ldots . \\ & & 0 \ \ldots \ d_{n-1,n} \\ & & & 0 \end{bmatrix}$$

But, how can the distance measure be obtained??

Distance measure

Satisfaction level of services of "Sheraton"

To measure the distance between two statistical units means to quantify the difference between the levels of satisfaction on the characteristics of the hotel.

Greater is the similarity of the answers, closer are the units!

A distance index has the following properties:

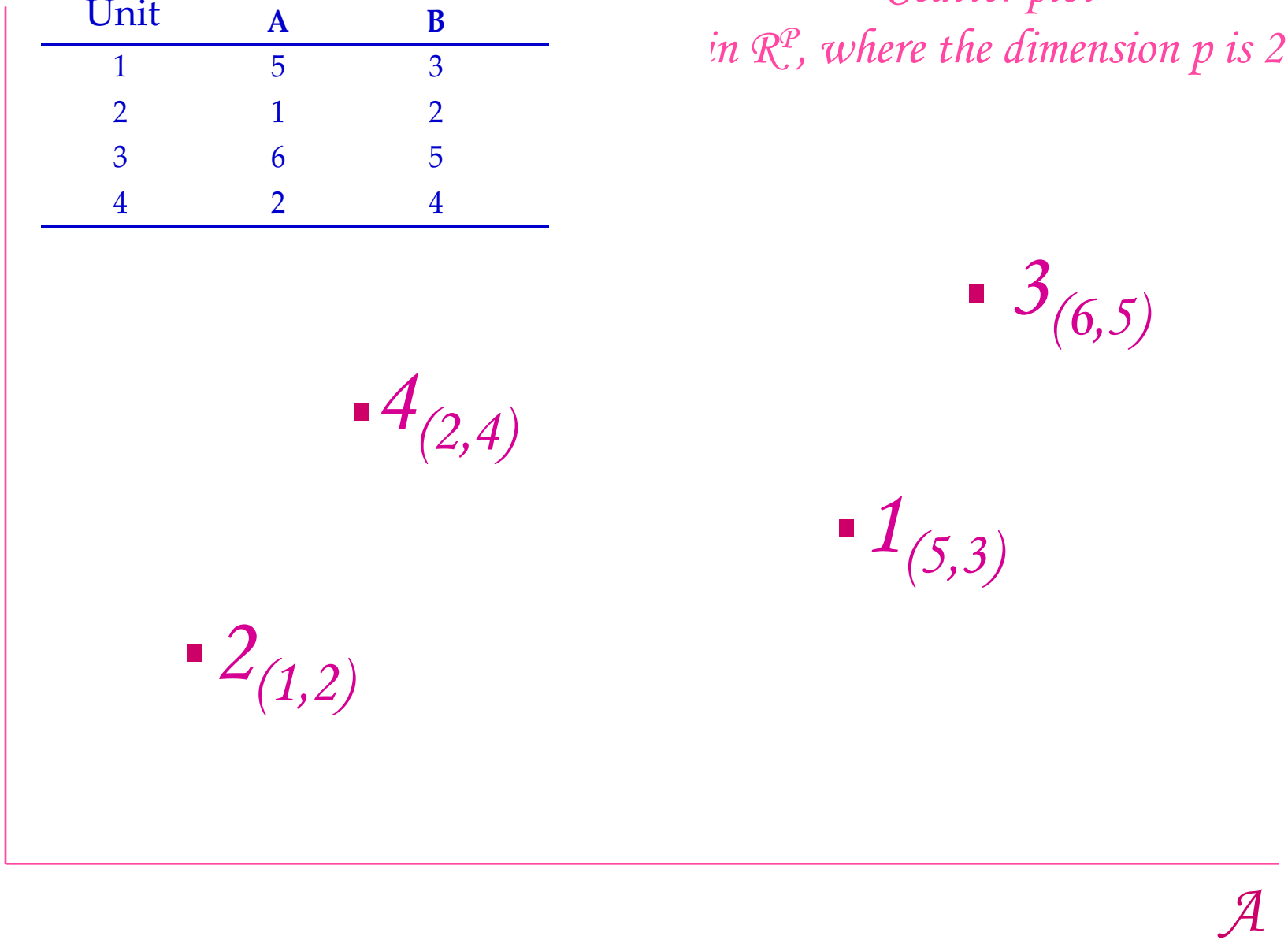
- Non negativity $\longrightarrow$ $d_{ij} \geq 0$
- Identity $\longrightarrow$ $d_{ij} = 0 \ \ for \ \ i=j$
- Symmetry $\longrightarrow$ $d_{ij} = d_{ji}$
- Triangular inequality $\longrightarrow$ $d_{ij} \leq d_{is} + d_{sj}$

A SPACE WITH A DISTANCE CHARACTERIZED BY ALL PROPERTIES IS DEFINED METRIC SPACE

COFFEE

Unit	A	B
1	5	3
2	1	2
3	6	5
4	2	4

Scatter plot
in $\mathcal{R}^p$, where the dimension p is 2



How to obtain the distances???

Distinguishing for character type:

- Nominal (e.g., Sneath Index)
- Dichotomous (e.g., Jaccard Index)
- Ordinal
- Quantitative (e.g., Euclidean distance)
- Mixed

NOMINAL:
SNEATH DISTANCE INDEX

$$d_{ij} = \frac{\sum_{s=1}^{p} d_{ij,s}}{p}$$

$d_{ij,s} = 1$ *if the categories of the s - th variable are different* $(x_{is} \neq x_{sj})$

$d_{ij,s} = 0 \quad se\ x_{is} = x_{sj}$

The corresponding similarity index is

$$c_{ij} = 1 - \frac{\sum_{s=1}^{p} d_{ij,s}}{p}$$

EXAMPLE: SNEATH INDEX

Unit	Coffee	Pasta	Oil	Supermarket
1	Kimbo	Barilla	Carapelli	Conad
2	Lavazza	Garofalo	Monini	Esselunga
3	Segafredo	Barilla	Carapelli	Conad
4	Kimbo	Garofalo	Monini	GS
5	Lavazza	Barilla	Monini	GS

p = 4 variables

$$d_{12} = \frac{\sum_{s=1}^{4} d_{12,s}}{4}$$

$d_{12,1} = 1 \quad d_{12,2} = 1$

$d_{12,3} = 1 \quad d_{12,4} = 1$

The distance between 1 and 2 is:

$$d_{12} = \frac{1+1+1+1}{4} = 1$$

$d_{13,1} = 1 \quad d_{13,2} = 0$

$d_{13,3} = 0 \quad d_{13,4} = 0$

The distance between 1 and 3 is:

$$d_{13} = \frac{1}{4} = 0.25$$

$$D = \begin{pmatrix} 0 & 1 & 0.25. & 0.75 & 0.75 \\ & 0 & 1 & 0.5 & 0.5 \\ & & 0 & 1 & 0.75. \\ & & & 0 & 0.5 \\ & & & & 0 \end{pmatrix}$$

DISTANCE MATRIX!

DICHOTOMOUS

unit i	unit j: 1	unit j: 0
1	a	b
0	c	d

We can get three indeces:

- Simple Matching → $d_{ij} = \frac{b+c}{p}$
- Jaccard → $d_{ij} = \frac{b+c}{a+b+c}$
- Czekanowski → $d_{ij} = \frac{b+c}{2a+b+c}$

EXAMPLE: DICHOTOMOUS

Subject	Meat	Fish	Fruit	Pasta	Milk	Bread
1	yes	yes	no	yes	no	no
2	yes	no	no	yes	yes	yes
3	no	no	no	no	yes	no
4	yes	no	yes	yes	yes	yes

Yes=1 No=0 p=6

The distance between 2 and 4 is:

a=4, b=0, c=1, d=1

Simple matching d_{24} = b+c/p = 1/6

Jaccard d_{24}= b+c/a+b+c = 1/5

Czekanowski d_{24} = b+c/2a+b+c = 1/9

unit i \ unit j	1	0
1	a	b
0	c	d

$$D = \begin{bmatrix} 0 & 3/6 & 4/6 & 4/6 \\ & 0 & 3/6 & 1/6 \\ & & 0 & 4/6 \\ & & & 0 \end{bmatrix}$$

Distance matrix for SIMPLE MATCHING!!!

ORDINAL

GENERALLY ENCODED with cardinal numbers and measured by index for quantitative variables, e.g:

1=none; 2=little; 3= enough; 4=much; 5=most

Arbitrariness:

- the differences between two consecutive categories are always the same (the difference between 1 and 2 is equal to the difference between 4 and 5)

- it is possible to calculate the rario between different categories (the difference between 4=much and 2=little is double of the difference between 2=little and 1=none)

QUANTITATIVE

EUCLIDEAN DISTANCE INDEX

$$d_{ij} = \left\{ \sum_{s=1}^{p} (x_{is} - x_{js})^2 \right\}^{1/2} \quad \forall \, i \, e \, j$$

OR, IN VECTOR TERMS

$$d_{ij} = \left[(x_i - x_j)' (x_i - x_j) \right]^{1/2}$$

EXAMPLE: EUCLIDEAN DISTANCE

COFFEE

Unit	A	B	C
1	5	3	6
2	1	2	4
3	6	5	1
4	2	4	2

THE DISTANCE BETWEEN 1 AND 2 IS

$$d_{12} = \left\{ \sqrt{\sum_{s=1}^{3} (x_{1s} - x_{2s})^2} \right\} = \sqrt{(5\text{-}1)^2 + (3-2)^2 + (6-4)^2} = 4.58$$

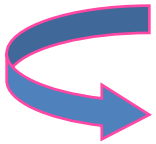
$$D = \begin{pmatrix} 0 & 4{,}58 & 5{,}48 & 5{,}10 \\ & 0 & 6{,}56 & 3 \\ & & 0 & 4{,}24 \end{pmatrix}$$

EUCLIDEAN DISTANCE MATRIX

EUCLIDEAN DISTANCE: PROBLEMS

At the same level variables with different nature and expressed in different measures

To overcome this problem we standardize the data matrix X



$$z_{is} = \frac{x_{is} - \bar{x}_s}{\sigma_s}$$

X ⟶ Z

Unit	COFFEE A	B	C
1	5	3	6
2	1	2	4
3	6	5	1
4	2	4	2
media	3,5	3,5	3,3
σ	2,1	1,1	1,9

Unit	COFFEE A	B	C
1	0,73	-0,45	1,43
2	-1,21	-1,34	0,39
3	1,21	1,34	-1,17
4	-0,73	0,45	-0,65

EUCLIDEAN DISTANCE: PROBLEMS

It does not take into account correlations among variables, for this reason it is better to consider

WEIGHTED EUCLIDEAN DISTANCE

$$d_{ij} = \sqrt{\sum_{s=1}^{p} (x_{is} - x_{js})^2 w_s}$$

In matrix notation,

$$d_{ij} = \left[\left(x_i - x_j \right)' W \left(x_i - x_j \right) \right]^{1/2}$$

MAHALANOBIS DISTANCE

If we consider S, the inverse of the covariance matrix, instead of W matrix, We avoid both the problems, the presence of correlations among variables and the different measure scales

$$d_{ij} = \left[\left(x_i - x_j \right)' S^{-1} \left(x_i - x_j \right) \right]^{1/2}$$

EXAMPLE: MALHANOBIS DISTANCE

Clients	Income	Consumption	Average amount receipt
1	2	1,5	10
2	4	3	9
3	3	1,8	11
4	1	1,3	9,5
5	3,5	2	10

D= Euclidean distance matrix S = covariance matrix R = correlation matrix

$$D = \begin{pmatrix} 0 & 2,69 & 1,45 & 1,14 & 1,58 \\ & 0 & 2,54 & 3,48 & 1,50 \\ & & 0 & 2,55 & 1,14 \\ & & & 0 & 2,64 \\ & & & & 0 \end{pmatrix} \quad S = \begin{pmatrix} 1,16 & 0,56 & -0,03 \\ & 0,35 & -0,18 \\ & & 0,44 \end{pmatrix} \quad R = \begin{pmatrix} 1,00 & 0,87 & -0,04 \\ & 1,00 & -0,45 \\ & & 1,00 \end{pmatrix}$$

$$S^{-1} = \begin{pmatrix} 10,8 & -21,2 & -7,84 \\ & 45,14 & 16,82 \\ & & 8,54 \end{pmatrix}$$

Inverse covariance matrix

EXAMPLE: MALHANOBIS DISTANCE

Clients	Income	Consumption	Average amount receipt
1	2	1,5	10
2	4	3	9
3	3	1,8	11
4	1	1,3	9,5
5	3,5	2	10

$$d_{ij} = \left[\left(x_i - x_j \right)' S^{-1} \left(x_i - x_j \right) \right]^{1/2}$$

Malhanobis distance between 2-4 is:

$$d_{24} = \left((4-1) \quad (3-1.3) \quad (9-9.5) \right) \begin{pmatrix} 11 & -21,2 & -7,84 \\ & 45,14 & 16,82 \\ & & 8,54 \end{pmatrix} \begin{pmatrix} (4-1) \\ (3-1.3) \\ (9-9.5) \end{pmatrix}^{1/2} = 2.94$$

$(x_2 - x_4)'$ $\quad$ S^{-1} $\quad$ $(x_2 - x_4)$

The Malhanobis distance value (2.94) is lower than euclidean distance value (3.48)

MINKOWSKI DISTANCE: quantitative variables

Is a generalization of the previous indeces

$$d_{ij} = \left[\sum_{s=1}^{p} \left| x_{is} - x_{js} \right|^{\lambda} \right]^{1/\lambda}$$

λ=2 ⟶ Euclidean distance index

λ=1 ⟶ Manhattan or city blocks distance index

GOWER INDEX: mixed characters

$$d_{ij} = \frac{\sum_{s=1}^{p} d_{ij,s}}{\sum_{s=1}^{p} \delta_{ij,s}}$$

$\delta_{ij,s}$ = is 1 if i and j can be compared with respect to the s-th attribute and 0 otherwise (if the character is null for both units, eg. , no-no)

The distance between i and j on the s-th attribute is:

$d_{ij,s}$ →

	unit j	
unit i	1	0
1	0	1
0	1	0

$\delta_{ij,s}$ →

	unit j	
unit i	1	0
1	1	1
0	1	0

For quantitative variables → $d_{ij,s} = \frac{|x_{is} - x_{js}|}{range(s)}$

EXAMPLE: GOWER INDEX

UNIT	MEAT CONSUMPTION	COFFEE	INCOME
1	SI	KIMBO	5
2	NO	KIMBO	4
3	SI	ILLY	3
4	SI	LAVAZZA	1

FOR MIXED CHARACTERS…

$$d_{ij,s} = \frac{|x_{is} - x_{js}|}{range(s)} \rightarrow d_{24,reddito} = \frac{|x_{2\,reddito} - x_{4\,reddito}|}{range(INCOME)} = \frac{|4-1|}{4} = \frac{3}{4}$$

$$d_{ij} = \frac{\sum_{s=1}^{p} d_{ij,s}}{\sum_{s=1}^{p} \delta_{ij,s}} \longrightarrow d_{24} = \frac{\sum_{s=1}^{3} d_{24,s}}{\sum_{s=1}^{3} \delta_{24,s}} = \frac{1+1+3/4}{1+1+1} = 0.92$$