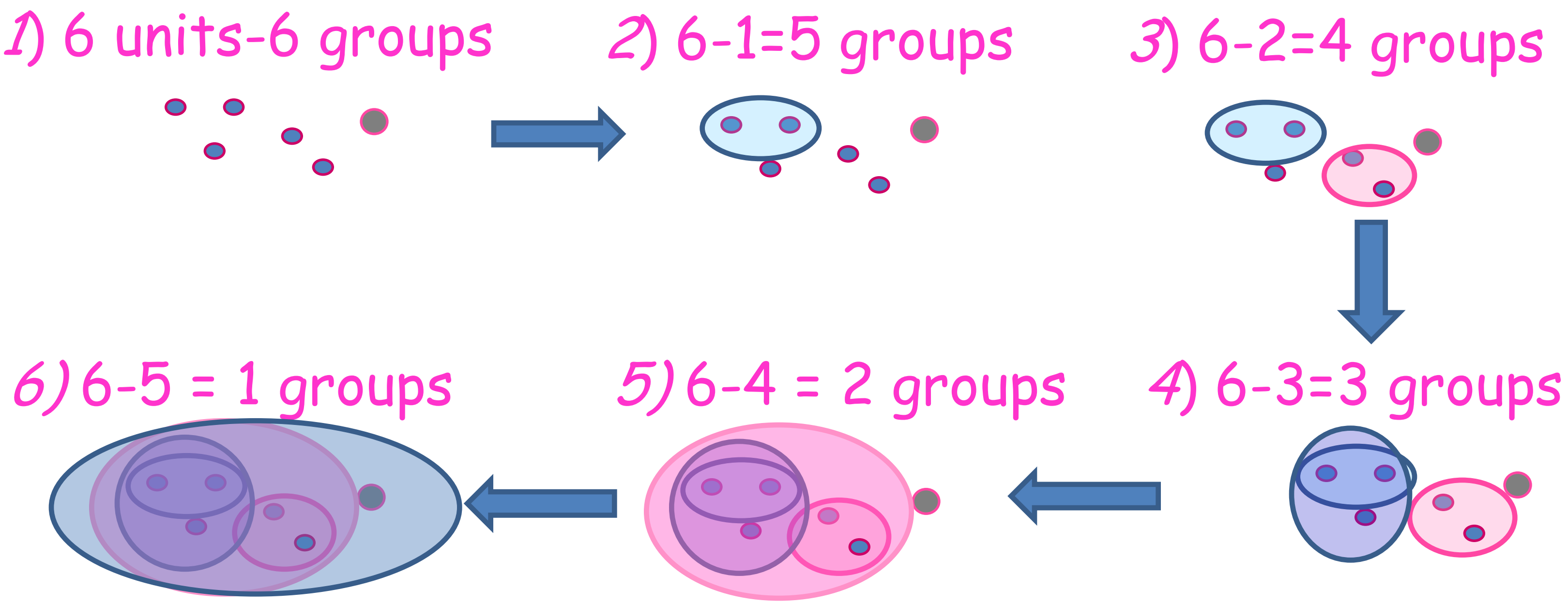
Agglomerative hierarchical method



1) 6 units-6 groups

2) 6-1=5 groups

3) 6-2=4 groups

4) 6-3=3 groups

5) 6-4 = 2 groups

6) 6-5 = 1 groups

Agglomerative algorithm

1. It starts from n-groups for which the distances are known

Considering quantitative data

	PROGRAM			
CHANNEL	**FILM**	**TELEFILM**	**VARIETY**	**NEWS**
RAI1	1158	1280	1577	1703
RAI2	731	1366	1280	1019
RAI3	1454	675	937	1618
RETE4	2053	1289	489	1410
CANALE5	582	1193	2166	3372
ITALIA1	1167	3119	795	1261

Hours of programming for channel and program type

Euclidean distance between RAI 1 and RAI 2 is:

$$d_{12} = \left\{ \sqrt{\sum_{s=1}^{4} (x_{1s} - x_{2s})^2} \right\} = \sqrt{(1158\text{-}731)^2 + (1280 - 1366)^2 + \ldots + (1703 - 1019)^2} = 864$$

D =

CHANNEL	RAI1	RAI2	RAI3	RETE4	CANALE5	ITALIA1
RAI1	0	864	933	1439	1863	2047
RAI2	864	0	1215	1591	2525	1886
RAI3	933	1215	0	990	2370	2491
RETE4	1439	1591	990	0	2972	2061
CANALE5	1863	2525	2370	2972	0	3223
ITALIA1	2047	1886	2491	2061	3223	0

EUCLIDEAN MATRIX DISTANCE!

Agglomerative algorithm

1. It starts from n-groups for which the distances are known
2. Then, it joins the units that have the shortest distance (more similar) deleting their distance from the D matrix.
3. It adds to the matrix D a new row and a new column with the distance of the new group obtained from the other.
4. It executes iteratively the procedure starting from step 2, reducing the matrix D by one unit at each step, until to the final configuration formed by a group constituted by the n-initial units.

PROBLEMS!

How to replace the distance between two units (or two groups) with that of the new group from the other?

Methods:

single linkage, complete linkage, McQuitty, average linkage, centroid and Ward's method.

Denoting with

- C_S=S-th group (in the first step is the first unit)
- N_S=number of units in the S-th group
- C_L=L-th group (in the first step is the first unit)
- N_L=number of units in the L-th group
- C_M= group formed by C_S e C_L groups
- N_M=number of units in the C_M group
- D_{SL}=distance between C_S and C_L in D matrix, which is minimum
- D_{MJ} = distance between C_M (formed) and a generic C_J group

SI HA…

Single linkage method (nearest-neighbor)

D = DISTANCE MATRIX

	RAI1	RAI2	RAI3	RETE4	CANALE5	ITALIA1
RAI1	0	864	933	1439	1863	2047
RAI2		0	1215	1591	2525	1886
RAI3			0	990	2370	2491
RETE4				0	2972	2062
CANALE5					0	3223
ITALIA1						0

$$D_{M,J} = \min(D_{SJ}, D_{LJ})$$

Distance between S (RAI1) and a generic J-th group (e.g RAI3)

Minimum distance: RAI1 and RAI2 = 864

It creates a new array, recalculating the distance between the group RAI1-RAI2 and the other television channels, as MINIMUM of distances presented before the fusion, individually , by RAI1 and RAI2.

	RAI1 e RAI2	RAI3	RETE4	CANALE5	ITALIA1
RAI1 e RAI2	0	933	1439	1863	1886
RAI3		0	990	2370	2491
RETE4			0	2972	2062
CANALE5				0	3223
ITALIA1					0

$$D_{RAI1-2,RAI3} = \min(933,1215)$$

Single linkage method (nearest-neighbor)

	RAI1 e RAI2	RAI3	RETE4	CANALE5	ITALIA1
RAI1 e RAI2	0	933	1439	1863	1886
RAI3		0	990	2370	2491
RETE4			0	2972	2062
CANALE5				0	3223
ITALIA1					0

	RAI1-RAI2-RAI3	RETE4	CANALE5	ITALIA1
RAI1-RAI2-RAI3	0	990	1863	1886
RETE4		0	2972	2062
CANALE5			0	3223
ITALIA1				0

Minimum distance is between the RAI1-RAI2 group and RAI3

Minimum distance is between the RAI1-RAI2-RAI3 group and RETE4

	RAI1-RAI2-RAI3-RETE4	CANALE5	ITALIA1
RAI1-RAI2-RAI3-RETE4	0	1863	1886
CANALE5		0	3223
ITALIA1			0

Minimum distance is between the RAI1-RAI2-RAI3-RETE4 group and CANALE5

	RAI1-RAI2-RAI3-RETE4-CANALE5	ITALIA1
RAI1-RAI2-RAI3-RETE4-CANALE5	0	1886
ITALIA1		0

And, finally, the last group is formed by all units!

1. Stopping rule

The aggregation procedure can be stopped, for this method and for those following, when the maximum distance within the new group exceeds the minimum outer distance between the groups!

The groups that provide the **well structured minimal partition**, which corresponds to the partition with the lowest number of groups in which the maximum distance in the groups is lower than the minimum distance between the groups:

1. (RAI1-RAI2-RAI3-RETE4)
2. (CANALE 5)
3. (ITALIA1)

Aggregating Canale 5 to the first group, the maximum distance in the new group equal to 2,972 (RETE4 and Canale 5) exceeds the minimum distance between the groups of 1,886 (distance between an element of the group, RAI2 , and the outer " ITALIA1 ")

2. Stopping rule: Dendrogram

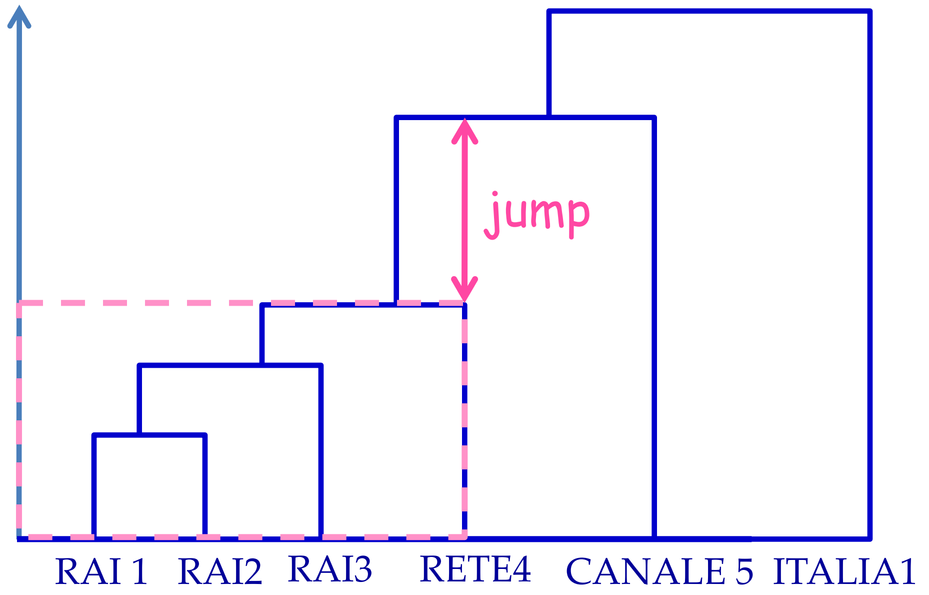
Graphical representation of the sequence of mergers that will help you to understand exactly how many groups to form.

Horizontal axis = units involved in the fusion process

Vertical axis = distance at which occurs the fusion in groups

Examinating the dendrogram it is possible to choose the number of groups that determines a " flat "configuration. In this sense, they will be characterized by sufficient internal homogeneity.

Dendrogram



A selection criterion may consist in stopping the fusion procedure before one of the " jumps" that are generated by combinations of very far groups.

Complete linkage method
(farthest-neighbor)

D = DISTANCE MATRIX

	RAI1	RAI2	RAI3	RETE4	CANALE5	ITALIA1
RAI1	0	864	933	1439	1863	2047
RAI2		0	1215	1591	2525	1886
RAI3			0	990	2370	2491
RETE4				0	2972	2062
CANALE5					0	3223
ITALIA1						0

$$D_{M,J} = \max(D_{SJ}, D_{LJ})$$

Minimum distance: RAI1 and RAI2 = 864

It creates a new array, recalculating the distance between the group RAI1-RAI2 and the other television channels, as MAXIMUM of distances presented before the fusion, individually , by RAI1 and RAI2.

	RAI1 e RAI2	RAI3	RETE4	CANALE5	ITALIA1
RAI1 e RAI2	0	1215	1591	2525	2047
RAI3		0	990	2370	2491
RETE4			0	2972	2062
CANALE5				0	3223
ITALIA1					0

Complete linkage method

	RAI1 e RAI2	RAI3	RETE4	CANALE5	ITALIA1
RAI1 e RAI2	0	1215	1591	2525	2047
RAI3		0	990	2370	2491
RETE4			0	2972	2062
CANALE5				0	3223
ITALIA1					0

Minimum distance is between RAI3-RETE4

	RAI1-RAI2	RAI3-RETE4	CANALE5	ITALIA1
RAI1-RAI2	0	1591	2525	2047
RAI3 - RETE4		0	2972	2491
CANALE5			0	3223
ITALIA1				0

Minimum distance is between the group RAI1-RAI2 and RAI3-RETE4

	RAI1-RAI2-RAI3-RETE4	CANALE5	ITALIA1
RAI1-RAI2-RAI3-RETE4	0	2972	2491
CANALE5		0	3223
ITALIA1			0

Minimum distance is between the group RAI1-RAI2-RAI3-RETE4 and ITALIA1

	RAI1-RAI2-RAI3-RETE4-ITALIA1	CANALE 5
RAI1-RAI2-RAI3-RETE4-ITALIA1	0	3223
CANALE 5		0

Finally, one group is formed by all units!!!

McQuitty

D = DISTANCE MATRIX

	RAI1	RAI2	RAI3	RETE4	CANALE5	ITALIA1
RAI1	0	864	933	1439	1863	2047
RAI2		0	1215	1591	2525	1886
RAI3			0	990	2370	2491
RETE4				0	2972	2062
CANALE5					0	3223
ITALIA1						0

$$D_{M,J} = (D_{SJ} + D_{LJ})/2$$

Minimum distance: RAI1 and RAI2 = 864

It creates a new array, recalculating the distance between the group RAI1-RAI2 and the other television channels, as AVERAGE of distances presented before the fusion, individually , by RAI1 and RAI2.

	RAI1 e RAI2	RAI3	RETE4	CANALE5	ITALIA1
RAI1 e RAI2	0	1074	1515	2194	1967
RAI3		0	990	2370	2491
RETE4			0	2972	2062
CANALE5				0	3223
ITALIA1					0

$$D_{RAI1-2,RAI3} = (933+1215)/2 = 1074$$

McQuitty

	RAI1 e RAI2	RAI3	RETE4	CANALE5	ITALIA1
RAI1 e RAI2	0	1074	1515	2194	1967
RAI3		0	990	2370	2491
RETE4			0	2972	2062
CANALE5				0	3223
ITALIA1					0

	RAI1-RAI2	RAI3-RETE4	CANALE5	ITALIA1
RAI1-RAI2	0	1295	2194	1966.5
RAI3 - RETE4		0	2671	2276.5
CANALE5			0	3223
ITALIA1				0

Minimum distance is between RAI3-RETE4

Minimum distance is between RAI1-RAI2 and RAI3-RETE4

	RAI1-RAI2-RAI3-RETE4	CANALE5	ITALIA1
RAI1-RAI2-RAI3-RETE4	0	2433	2122
CANALE5		0	3223
ITALIA1			0

Minimum distance is between RAI1-RAI2-RAI3-RETE4 and ITALIA1

	RAI1-RAI2-RAI3-RETE4-ITALIA1	CANALE 5
RAI1-RAI2-RAI3-RETE4-ITALIA1	0	2828
CANALE 5		0

Finally, one group is formed by all units!!!

Average linkage

D = DISTANCE MATRIX

	RAI1	RAI2	RAI3	RETE4	CANALE5	ITALIA1
RAI1	0	864	933	1439	1863	2047
RAI2		0	1215	1591	2525	1886
RAI3			0	990	2370	2491
RETE4				0	2972	2062
CANALE5					0	3223
ITALIA1						0

$$D_{M,J} = (D_{SJ} N_S + D_{LJ} N_L) / N_M$$

Minimum distance: RAI1 and RAI2 = 864

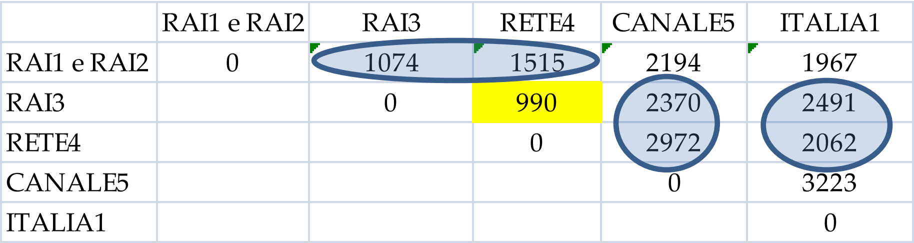
It creates a new array, recalculating the distance between the group RAI1-RAI2 and the other television channels, as WEIGHTED AVERAGE by numerosity of merged groups (in this case $N_S = N_L = 1$), of distances presented before the fusion, individually , by RAI1 and RAI2.

	RAI1 e RAI2	RAI3	RETE4	CANALE5	ITALIA1
RAI1 e RAI2	0	1074	1515	2194	1967
RAI3		0	990	2370	2491
RETE4			0	2972	2062
CANALE5				0	3223
ITALIA1					0

$$D_{RAI1-2,RAI3} = (933 \cdot 1 + 1215 \cdot 1) / 2 =$$

$$= 1074$$

Average linkage



	RAI1 e RAI2	RAI3	RETE4	CANALE5	ITALIA1
RAI1 e RAI2	0	1074	1515	2194	1967
RAI3		0	990	2370	2491
RETE4			0	2972	2062
CANALE5				0	3223
ITALIA1					0

	RAI1-RAI2	RAI3-RETE4	CANALE5	ITALIA1
RAI1-RAI2	0	1295	2194	1966.5
RAI3 - RETE4		0	2671	2276.5
CANALE5			0	3223
ITALIA1				0

Minimum distance is between RAI3-RETE4

Minimum distance is between RAI1-RAI2 and RAI3-RETE4

$$D_{RAI1-2-3-RETE4,CANALE5} = (2194 \cdot 2 + 2671 \cdot 2)/4 = 2433$$

	RAI1-RAI2-RAI3-RETE4	CANALE5	ITALIA1
RAI1-RAI2-RAI3-RETE4	0	2433	2122
CANALE5		0	3223
ITALIA1			0

Minimum distance is between RAI1-RAI2-RAI3-RETE4 and ITALIA1

$$D_{group,CANALE5} = (2433 * 4 + 3223 * 1)/5 = 2591$$

	RAI1-RAI2-RAI3-RETE4-ITALIA1	CANALE 5
RAI1-RAI2-RAI3-RETE4-ITALIA1	0	2591
CANALE 5		0

Finally, one group is formed by all units!!!

Centroid

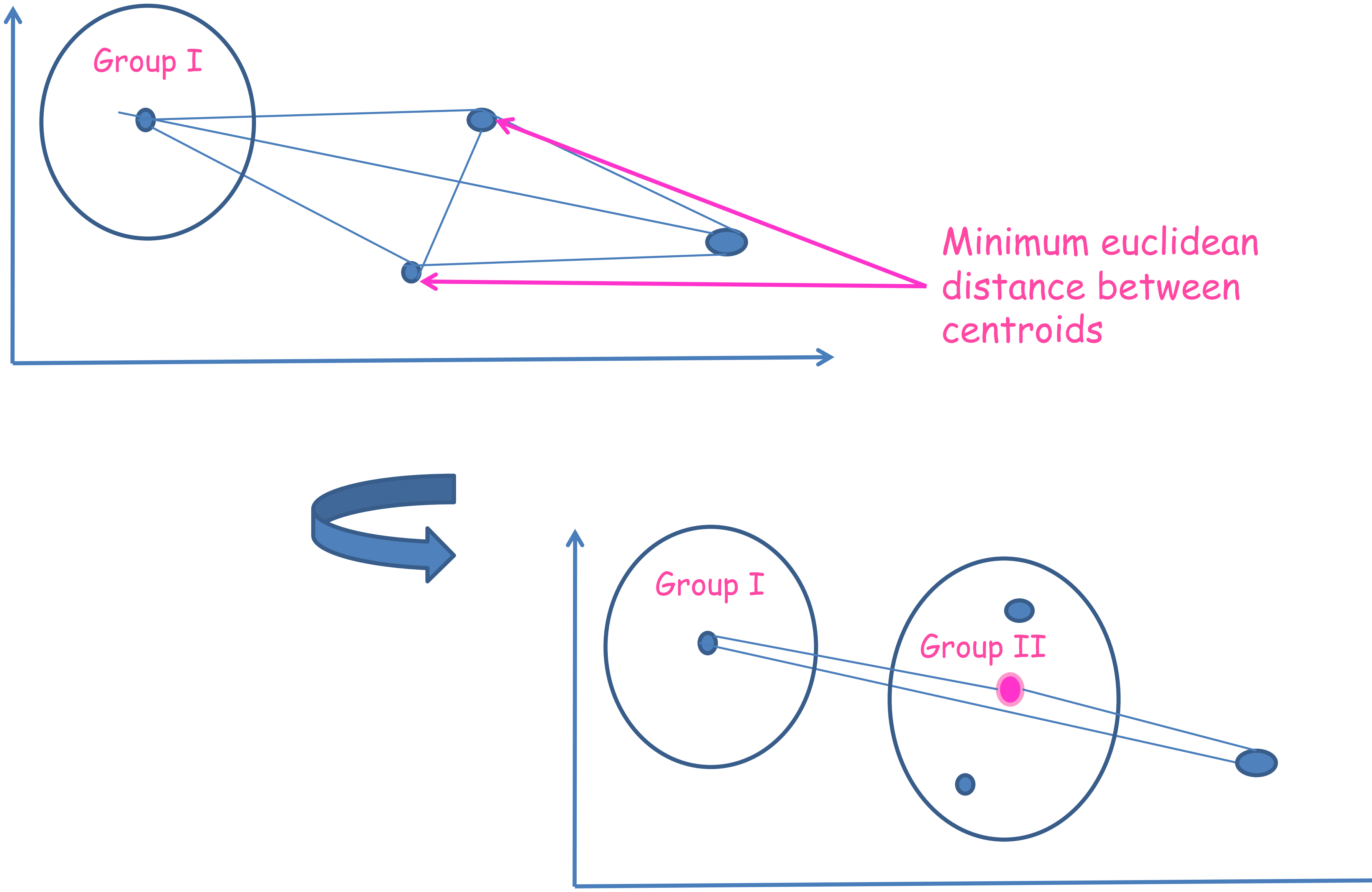
In the centroid method the distance between two clusters is the distance (generally Euclidean) between their centroids.

The centroids are the mean of p variables of the group units

Starting from the *p*-variable matrix, groups are joined with minimum Euclidean distance between centroids

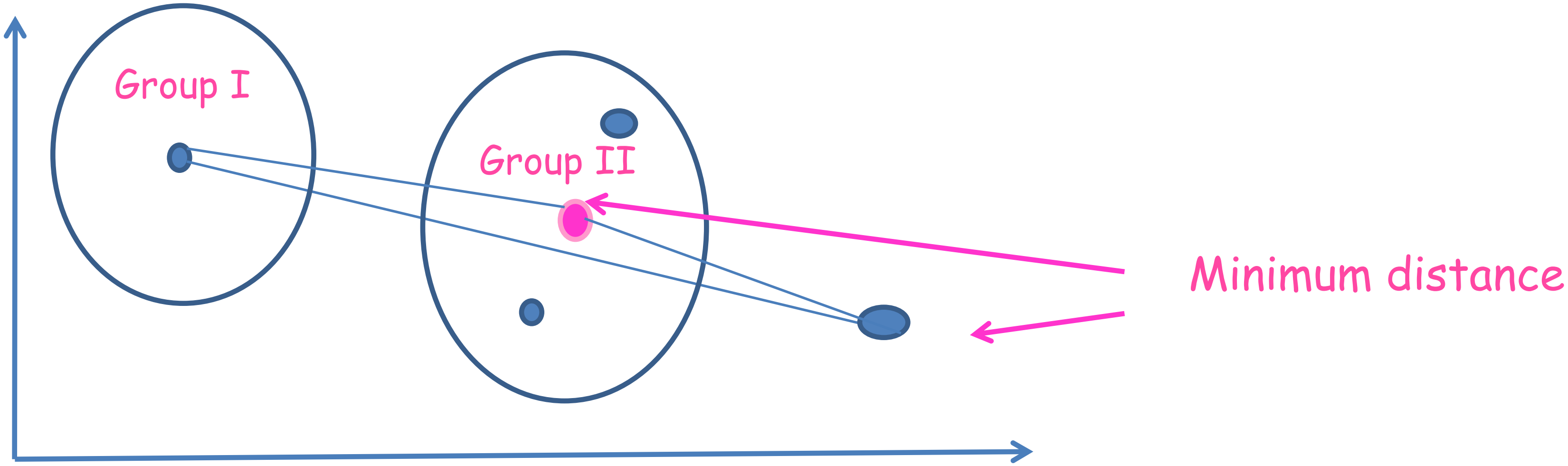
$$\boxed{D_{M,J} = (D_{SJ} N_S + D_{LJ} N_L) / N_M - N_S N_L D_{SL} / N_M^2}$$

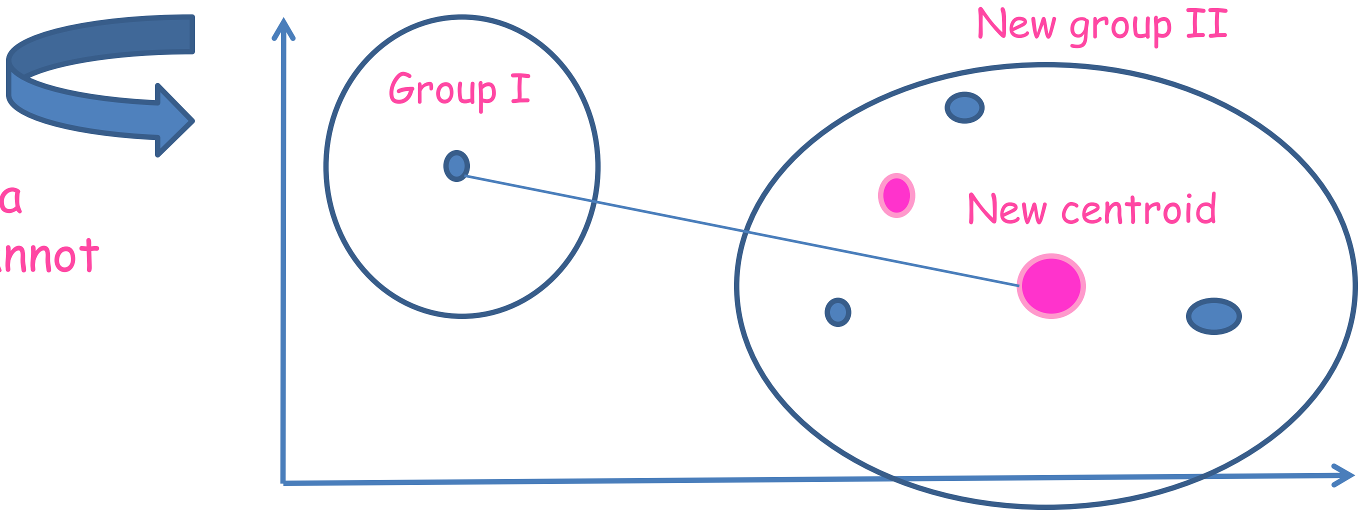
Centroid: 2 variables



Minimum euclidean distance between centroids

Centroid





Units aggregated in a group (group II) cannot be subsequently separated!

Centroid

Considering 2 variables:

Choose the group with the least Euclidean distance (2 and 4) and estimate the centroid (average of the categories presented individually by the two groups before the fusion), obtaining: 3.5=(3+4)/2

UNITS	BRAND	
	A	B
1	5	7
2	3	2
3	1	5
4	4	3

EUCLIDEAN DISTANCE MATRIX

UNITS	1	2	3	4
1	0	5.385165	4.472136	4.123106
2		0	3.605551	1.414214
3			0	3.605551
4				0

CENTROIDS

GROUPS	BRAND	
	A	B
24	3.5	2.5
1	5	7
3	4	3

Distance recalculation, selection of the group on the minimum distance and estimation of new centroid

EUCLIDEAN DISTANCE MATRIX

GROUPS	1	24	3
1	0	4.743416	4.123106
24		0	0.707107
3			0

CENTROIDS

GROUPS	BRAND	
	A	B
243	3.75	2.75
1	5	7