STATISTICA E ANALISI DEI DATI

Capitolo 7 - Analisi dei cluster: parte 2

Dott. Stefano Cirillo
Dott. Luigi Di Biasi

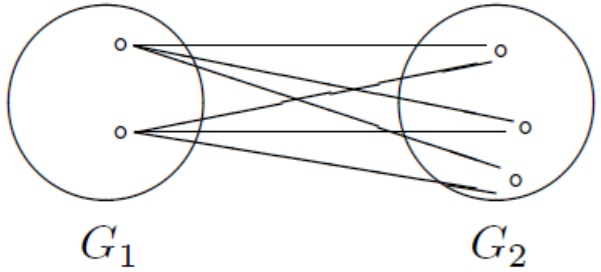
a.a. 2024-2025

CLUSTERING GERARCHICO

Metodo del Legame Medio

Metodo del Legame Medio

- Nel **Metodo del legame medio** o Average neighbour method la distanza tra i gruppi G_1 (contenente n_1 individui) e G_2 (contenente n_2 individui) è definita come la **media aritmetica** tra tutte le distanze tra $n_1 n_2$ che si possono calcolare tra ogni individuo di G_1 e ogni individuo di G_2



- Nella procedura gerarchica si considera inizialmente, ossia al livello 0, un insieme di n cluster $\{I_1\}, \{I_2\}, \ldots, \{I_n\}$
- Al passo successivo si cerca nella matrice D delle distanze il coefficiente di distanza **minima** e si raggruppano nello stesso cluster G_{ij} i due individui I_i e I_j associati secondo tale coefficiente
- Nel caso i coefficienti di distanza minima siano più di uno, si attua una scelta arbitraria tra di essi.

Metodo del Legame Medio

- Al livello 1 quindi si modifica la matrice delle distanze valutando le distanze di G_{ij} da ogni altro individuo I_k non appartenente a G_{ij} mediante la seguente relazione:

$$d_{(ij),k} = \frac{d_{ij} + d_{jk}}{2} \qquad k = 1, 2, \ldots, n;\ k \neq i, j$$

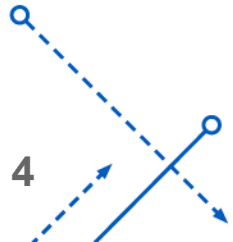
Cioè la distanza dell'individuo I_k dal cluster G_{ij} si ottiene scegliendo la più grande distanza tra d_{ij} e d_{jk}

- Quindi, al livello 1 si costruisce una nuova matrice D_1 di cardinalità (n−1)×(n−1) costituita da G_{ij} (che viene considerato come un unico elemento) e dai restanti (n-2) individui fuori dal cluster G_{ij}
- Ad ogni passo successivo, dopo che i cluster G_u e G_v sono stati uniti scegliendo dalla precedente matrice delle distanze i due cluster più vicini, la distanza tra il nuovo cluster, denotato con G_{uv}, e un altro cluster G_z è così definita:

$$d_{(uv),z} = \frac{N_u}{N_u + N_v} d_{uv} + \frac{N_v}{N_u + N_v} d_{vz}$$

con N_u, N_v il numero di individui in G_u e G_v , $d_{(uv),z}$ la misura di distanza tra gli elementi meno distanti dei cluster G_{uv} e G_z

- La procedura si ripete fino ad ottenere un unico cluster formato da tutti gli individui



Esempio in R

- Applichiamo il metodo gerarchico del legame medio

$$X = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 6 & 3 \\ 8 & 2 \\ 8 & 0 \end{pmatrix} \end{array}$$

```
> hlm<-hclust(d,method="average")
> str(hlm) # visualizza informazioni sull'oggetto cluster
List of 7
 $ merge      : int [1:4, 1:2] -1 -3 1 -5 -2 -4 2 3
 $ height     : num [1:4] 0.877 1.041 2.004 2.307
 $ order      : int [1:5] 5 1 2 3 4
 $ labels     : chr [1:5] "I1" "I2" "I3" "I4" ...
 $ method     : chr "average"
 $ call       : language hclust(d = d, method = "average")
 $ dist.method: chr "euclidean"
 - attr(*, "class")= chr "hclust"
```

		Agglomerazione	Distanza
-1	-2	Al livello 1 si uniscono gli individui I_1 e I_2	0.877

Esempio in R

- Applichiamo il metodo gerarchico del legame medio

$$X = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 6 & 3 \\ 8 & 2 \\ 8 & 0 \end{pmatrix} \end{array}$$

```
> hlm<-hclust(d,method="average")
> str(hlm) # visualizza informazioni sull'oggetto cluster
List of 7
 $ merge      : int [1:4, 1:2] -1 -3 1 -5 -2 -4 2 3
 $ height     : num [1:4] 0.877 1.041 2.004 2.307
 $ order      : int [1:5] 5 1 2 3 4
 $ labels     : chr [1:5] "I1" "I2" "I3" "I4" ...
 $ method     : chr "average"
 $ call       : language hclust(d = d, method = "average")
 $ dist.method: chr "euclidean"
 - attr(*, "class")= chr "hclust"
```

		Agglomerazione	Distanza
-1	-2	Al livello 1 si uniscono gli individui I_1 e I_2	0.877
-3	-4	Al livello 2 si uniscono gli individui I_3 e I_4	1.041

Esempio in R

- Applichiamo il metodo gerarchico del legame medio

$$X = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 6 & 3 \\ 8 & 2 \\ 8 & 0 \end{pmatrix} \end{array}$$

```
> hlm<-hclust(d,method="average")
> str(hlm) # visualizza informazioni sull'oggetto cluster
List of 7
 $ merge      : int [1:4, 1:2] -1 -3 1 -5 -2 -4 2 3
 $ height     : num [1:4] 0.877 1.041 2.004 2.307
 $ order      : int [1:5] 5 1 2 3 4
 $ labels     : chr [1:5] "I1" "I2" "I3" "I4" ...
 $ method     : chr "average"
 $ call       : language hclust(d = d, method = "average")
 $ dist.method: chr "euclidean"
 - attr(*, "class")= chr "hclust"
```

		Agglomerazione	Distanza
-1	-2	Al livello 1 si uniscono gli individui I_1 e I_2	0.877
-3	-4	Al livello 2 si uniscono gli individui I_3 e I_4	1.041
1	2	Al livello 3 si uniscono il primo cluster (formato dagli individui I_1 e I_2) con il secondo cluster (formato dagli individui I_3 e I_4)	2.004

Esempio in R

- Applichiamo il metodo gerarchico del legame medio

$$X = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 6 & 3 \\ 8 & 2 \\ 8 & 0 \end{pmatrix} \end{array}$$

```
> hlm<-hclust(d,method="average")
> str(hlm) # visualizza informazioni sull'oggetto cluster
List of 7
 $ merge      : int [1:4, 1:2] -1 -3 1 -5 -2 -4 2 3
 $ height     : num [1:4] 0.877 1.041 2.004 2.307
 $ order      : int [1:5] 5 1 2 3 4
 $ labels     : chr [1:5] "I1" "I2" "I3" "I4" ...
 $ method     : chr "average"
 $ call       : language hclust(d = d, method = "average")
 $ dist.method: chr "euclidean"
 - attr(*, "class")= chr "hclust"
```

		Agglomerazione	Distanza
-1	-2	Al livello 1 si uniscono gli individui I_1 e I_2	0.877
-3	-4	Al livello 2 si uniscono gli individui I_3 e I_4	1.041
1	2	Al livello 3 si uniscono il primo cluster (formato dagli individui I_1 e I_2) con il secondo cluster (formato dagli individui I_3 e I_4)	2.004
-5	3	Al livello 4 si unisce il terzo cluster (formato dagli individui I_1, I_2, I_3, I_4) con l'individuo I_5	2.307

Esempio in R

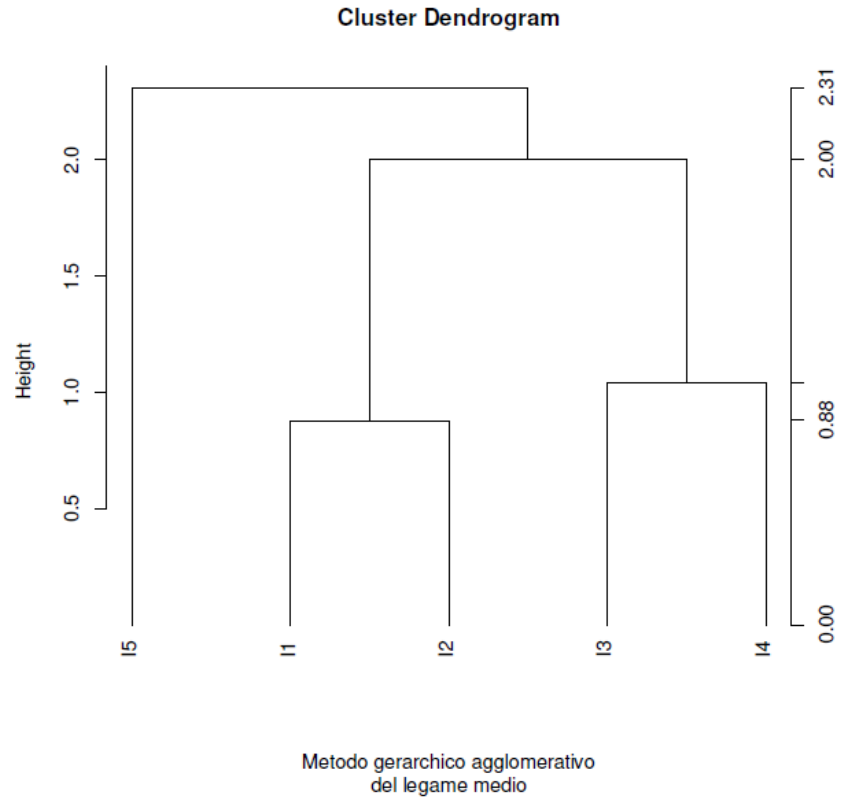
- Applichiamo il metodo gerarchico del legame medio

```
> hlm<-hclust(d,method="average")
> str(hlm) # visualizza informazioni sull'oggetto cluster
List of 7
 $ merge      : int [1:4, 1:2] -1 -3 1 -5 -2 -4 2 3
 $ height     : num [1:4] 0.877 1.041 2.004 2.307
 $ order      : int [1:5] 5 1 2 3 4
 $ labels     : chr [1:5] "I1" "I2" "I3" "I4" ...
 $ method     : chr "average"
 $ call       : language hclust(d = d, method = "average")
 $ dist.method: chr "euclidean"
 - attr(*, "class")= chr "hclust"
```

$$X = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 6 & 3 \\ 8 & 2 \\ 8 & 0 \end{pmatrix} \end{array}$$

- Costruiamo il Dendrogramma:

```
> plot(hlm,hang=-1,xlab="Metodo gerarchico agglomerativo",
+ sub="del legame medio")
> axis(side=4,at=round(c(0,hlm$height),2))
```



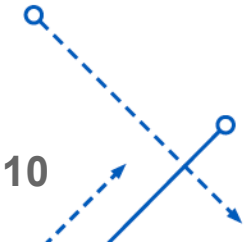
Esempio in R – Sveliamo l'arcano

- Come è avvenuto il processo di agglomerazione con il metodo del legame medio?
- Partiamo dalla matrice delle distanze ottenuta con R

$$D = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \end{array} \begin{array}{c} \begin{array}{ccccc} I_1 & I_2 & I_3 & I_4 & I_5 \end{array} \\ \begin{pmatrix} 0.000000 & \boxed{0.877058} & 2.246203 & 2.151162 & 2.151162 \\ \boxed{0.877058} & 0.000000 & 1.654610 & 1.964247 & 2.633475 \\ 2.246203 & 1.654610 & 0.000000 & 1.041245 & 2.690360 \\ 2.151162 & 1.964247 & 1.041245 & 0.000000 & 1.754116 \\ 2.151162 & 2.633475 & 2.690360 & 1.754116 & 0.000000 \end{pmatrix} \end{array}$$

- **Livello 1**: $d_{12} = 0.877058$ è il più piccolo valore della matrice delle distanze e pertanto I_1 e I_2 sono uniti formando un unico cluster
  - Le distanze tra questo nuovo gruppo e $\{I_3\}, \{I_4\}, \{I_5\}$ sono (dalla marice):

$$d_{(1,2),3} = (d_{1,3} + d_{2,3})/2 = (2.246203 + 1.654610)/2 = 1.950406$$
$$d_{(1,2),4} = (d_{1,4} + d_{2,4})/2 = (2.151162 + 1.964247)/2 = 2.057704$$
$$d_{(1,2),5} = (d_{1,5} + d_{2,5})/2 = (2.151162 + 2.633475)/2 = 2.392319$$



Esempio in R – Sveliamo l'arcano

- Come è avvenuto il processo di agglomerazione con il metodo del legame medio?
- Partiamo dalla matrice delle distanze ottenuta con R

$$D = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \end{array} \begin{array}{c} \begin{array}{ccccc} I_1 & I_2 & I_3 & I_4 & I_5 \end{array} \\ \begin{pmatrix} 0.000000 & \boxed{0.877058} & 2.246203 & 2.151162 & 2.151162 \\ \boxed{0.877058} & 0.000000 & 1.654610 & 1.964247 & 2.633475 \\ 2.246203 & 1.654610 & 0.000000 & 1.041245 & 2.690360 \\ 2.151162 & 1.964247 & 1.041245 & 0.000000 & 1.754116 \\ 2.151162 & 2.633475 & 2.690360 & 1.754116 & 0.000000 \end{pmatrix} \end{array}$$

- **Livello 1**: $d_{12} = 0.877058$ è il più piccolo valore della matrice delle distanze e pertanto I_1 e I_2 sono uniti formando un unico cluster
  - Le distanze tra questo nuovo gruppo e $\{I_3\}, \{I_4\}, \{I_5\}$ sono (dalla marice):

$$d_{(1,2),3} = (d_{1,3} + d_{2,3})/2 = (2.246203 + 1.654610)/2 = 1.950406$$
$$d_{(1,2),4} = (d_{1,4} + d_{2,4})/2 = (2.151162 + 1.964247)/2 = 2.057704$$
$$d_{(1,2),5} = (d_{1,5} + d_{2,5})/2 = (2.151162 + 2.633475)/2 = 2.392319$$

- È quindi possibile costruire una nuova matrice delle distanze D_1 di ordine 4 (considerando un individuo in meno):

$$D_1 = \begin{array}{c} \\ I_{1,2} \\ I_3 \\ I_4 \\ I_5 \end{array} \begin{array}{c} \begin{array}{cccc} I_{1,2} & I_3 & I_4 & I_5 \end{array} \\ \begin{pmatrix} 0.000000 & 1.950406 & 2.057704 & 2.392319 \\ 1.950406 & 0.000000 & \boxed{1.041245} & 2.690360 \\ 2.057704 & \boxed{1.041245} & 0.000000 & 1.754116 \\ 2.392319 & 2.690360 & 1.754116 & 0.000000 \end{pmatrix} \end{array}$$

Esempio in R – Sveliamo l'arcano

- **Livello 2**: $d_{34} = 1{,}041245$ è il più piccolo valore della matrice delle distanze e pertanto I_3 e I_4 sono uniti formando un unico cluster
  - Le distanze tra questo nuovo gruppo e $\{I_1, I_2\}, \{I_5\}$ sono (dalla marice):

$$D_1 = \begin{array}{c} \\ I_{1,2} \\ I_3 \\ I_4 \\ I_5 \end{array} \begin{array}{c} \begin{array}{cccc} I_{1,2} & I_3 & I_4 & I_5 \end{array} \\ \begin{pmatrix} 0.000000 & 1.950406 & 2.057704 & 2.392319 \\ 1.950406 & 0.000000 & \boxed{1.041245} & 2.690360 \\ 2.057704 & \boxed{1.041245} & 0.000000 & 1.754116 \\ 2.392319 & 2.690360 & 1.754116 & 0.000000 \end{pmatrix} \end{array}$$

$$d_{(3,4),(1,2)} = (d_{3,(1,2)} + d_{4,(12)})/2 = (1.950406 + 2.057704)/2 = 2.004055$$
$$d_{(3,4),5} = (d_{3,5} + d_{4,5})/2 = (2.690360 + 1.754116)/2 = 2.222238.$$

Esempio in R – Sveliamo l'arcano

- **Livello 2**: $d_{34} = 1{,}041245$ è il più piccolo valore della matrice delle distanze e pertanto I_3 e I_4 sono uniti formando un unico cluster
  - Le distanze tra questo nuovo gruppo e $\{I_1, I_2\}, \{I_5\}$ sono (dalla marice):

$$D_1 = \begin{array}{c} \\ I_{1,2} \\ I_3 \\ I_4 \\ I_5 \end{array} \begin{array}{c} \begin{array}{cccc} I_{1,2} & I_3 & I_4 & I_5 \end{array} \\ \begin{pmatrix} 0.000000 & 1.950406 & 2.057704 & 2.392319 \\ 1.950406 & 0.000000 & \boxed{1.041245} & 2.690360 \\ 2.057704 & \boxed{1.041245} & 0.000000 & 1.754116 \\ 2.392319 & 2.690360 & 1.754116 & 0.000000 \end{pmatrix} \end{array}$$

$$d_{(3,4),(1,2)} = (d_{3,(1,2)} + d_{4,(12)})/2 = (1.950406 + 2.057704)/2 = 2.004055$$
$$d_{(3,4),5} = (d_{3,5} + d_{4,5})/2 = (2.690360 + 1.754116)/2 = 2.222238.$$

- È quindi possibile costruire una nuova matrice delle distanze D_2 di ordine 3 (considerando due individui in meno):

$$D_2 = \begin{array}{c} \\ I_{1,2} \\ I_{3,4} \\ I_5 \end{array} \begin{array}{c} \begin{array}{ccc} I_{1,2} & I_{3,4} & I_5 \end{array} \\ \begin{pmatrix} 0.000000 & \boxed{2.004055} & 2.392319 \\ \boxed{2.004055} & 0.000000 & 2.222238 \\ 2.392319 & 2.222238 & 0.000000 \end{pmatrix} \end{array}$$

Esempio in R – Sveliamo l'arcano

- **Livello 3**: $d_{(12),(34)} = 2.004055$ è il più piccolo valore della matrice delle distanze e pertanto $I_{(1,2)}$ e $I_{(3,4)}$ sono uniti formando un unico cluster
  - La distanza tra questo nuovo gruppo e $\{I_5\}$ è (dalla marice):

$$D_2 = \begin{array}{c} \\ I_{1,2} \\ I_{3,4} \\ I_5 \end{array} \begin{array}{c} \begin{array}{ccc} I_{1,2} & I_{3,4} & I_5 \end{array} \\ \begin{pmatrix} 0.000000 & \boxed{2.004055} & 2.392319 \\ \boxed{2.004055} & 0.000000 & 2.222238 \\ 2.392319 & 2.222238 & 0.000000 \end{pmatrix} \end{array}$$

$$d_{(1,2,3,4),5} = \frac{2}{4} d_{(1,2),5} + \frac{2}{4} d_{(3,4),5} = (2.392319 + 2.222238)/2 = 2.307279$$

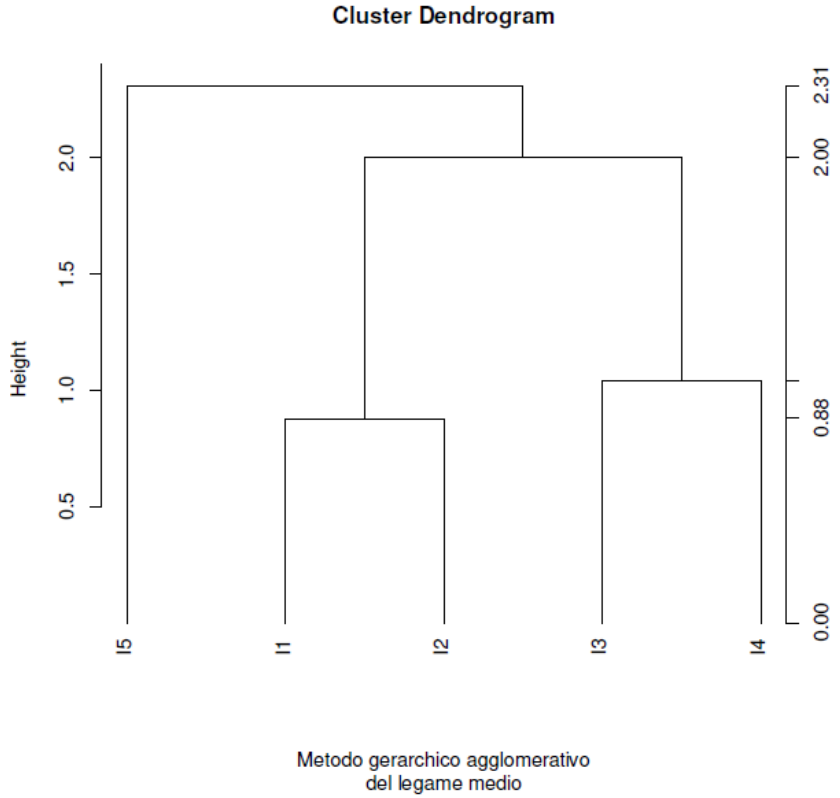
- È quindi possibile costruire una nuova matrice delle distanze D_3 di ordine 2 (considerando due individui in meno):

$$D_3 = \begin{array}{c} \\ I_{1,2,3,4} \\ I_5 \end{array} \begin{array}{c} \begin{array}{cc} I_{1,2,3,4} & I_5 \end{array} \\ \begin{pmatrix} 0.00000 & \boxed{2.307279} \\ \boxed{2.307279} & 0.000000 \end{pmatrix} \end{array}.$$

Esempio in R – Sveliamo l'arcano

- **Livello 4**: Unendo i gruppi $\{I_1, I_2, I_3, I_4\}$ e $\{I_5\}$ si ottiene un unico cluster contenente tutti e 5 gli individui
- La sequenza delle agglomerazioni del metodo del legame medio è stata:

Numero di cluster	Cluster	Livello di distanza
5	$\{I_1\}, \{I_2\}, \{I_3\}, \{I_4\}, \{I_5\}$	
4	$\{I_1, I_2\}, \{I_3\}, \{I_4\}, \{I_5\}$	0.877058
3	$\{I_1, I_2\}, \{I_3, I_4\}, \{I_5\}$	1.041245
2	$\{I_1, I_2, I_3, I_4\}, \{I_5\}$	2.004055
1	$\{I_1, I_2, I_3, I_4, I_5\}$	2.307279

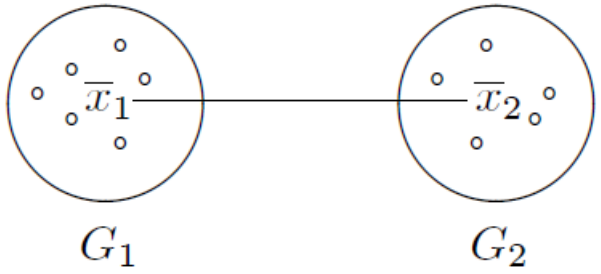


CLUSTERING GERARCHICO

Metodo del Centroide

Metodo del Centroide

- Nel **Metodo del centroide** la distanza tra i gruppi G_1 (contenente n_1 individui) e G_2 (contenente n_2 individui) è definita come la **distanza tra i centroidi**, ossia come le **medie campionarie** sugli individui appartenenti a G_1 e G_2



- Nella procedura gerarchica si considera inizialmente, ossia al livello 0, un insieme di n cluster

$$\{I_1\}, \{I_2\}, \dots, \{I_n\}$$

- Al passo successivo si cerca nella matrice $D^{(2)}$ contenente i **quadrati** delle singole distanze euclidee, il coefficiente di distanza **minima**
  - si raggruppano nello stesso cluster G_{ij} i due individui I_i e I_j associati secondo tale coefficiente
- Nel caso i coefficienti di distanza minima siano più di uno, si attua una scelta arbitraria tra di essi.

Metodo del Centroide

- Al livello 1 quindi si modifica la matrice dei quadrati delle distanze $D^{(2)}$ valutando i quadrati delle distanze di G_{ij} da ogni altro individuo I_k non appartenente a G_{ij} mediante la seguente relazione:

$$d^2_{(ij),k} = \sum_{r=1}^{p} (\bar{x}_{(i,j),r} - \bar{x}_{k,r}) = \frac{1}{2}\left(d^2_{ik} + d^2_{jk}\right) - \frac{1}{4} d^2_{ij} \qquad k \neq i,j$$

Dove la media campionaria è:

$$\bar{x}_{(i,j),r} = \frac{x_{i,r} + x_{j,r}}{2} \qquad e \qquad \bar{x}_{k,r} = x_{k,r} \qquad k \neq i,j$$

Cioè la distanza dell'individuo I_k dal cluster G_{ij} si ottiene scegliendo la più grande distanza tra d_{ij} e d_{jk}

- Quindi, al livello 1 si costruisce una nuova matrice X_1:
  - ottenendo una matrice di cardinalità $(n-1) * p$.con

$$X_1 = \begin{array}{c} \\ I_1 \\ I_2 \\ \vdots \\ I_{i,j} \\ \vdots \\ I_n \end{array} \begin{array}{c} \begin{array}{cccc} C_1 & C_2 & \dots & C_p \end{array} \\ \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \overline{x}_{(i,j),1} & \overline{x}_{(i,j),2} & \dots & \overline{x}_{(i,j),p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \end{array}$$

Metodo del Centroide

- Ad ogni passo successivo,
  - dopo che i cluster G_u e G_v sono stati uniti scegliendo dalla precedente matrice delle distanze i due cluster più vicini
  - la distanza tra il nuovo cluster, denotata con G_{uv}, e un altro cluster G_z è così definita:

$$d^2_{(uv),z} = \frac{N_u}{N_u + N_v} d^2_{u,z} + \frac{N_v}{N_u + N_v} d^2_{v,z} + \frac{N_u N_v}{(N_u + N_v)^2} d^2_{u,v}$$

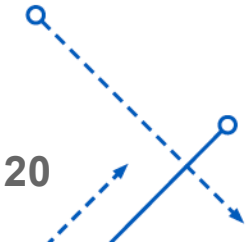
  con N_u, N_v e N_z denotano il numero di individui in G_u,G_v e G_z
- La procedura si ripete fino ad ottenere un unico cluster formato da tutti gli individui

Esempio in R

- Applichiamo il metodo gerarchico del centroide alle misurazioni ottenute sui 5 individui I_1, I_2, I_3, I_4, I_5
  - Definiamo la matrice dei dati:

$$X = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 36 & 20 \\ 35 & 25 \\ 40 & 21 \\ 37 & 28 \\ 33 & 24 \end{pmatrix} \end{array}$$

```
> X<-data.frame(c1=c(36,35,40,37,33),c2=c(20,25,21,28,24))
> row.names(X)<-c("I1","I2","I3","I4","I5")
> X # visualizza il data frame X
   c1 c2
I1 36 20
I2 35 25
I3 40 21
I4 37 28
I5 33 24
```



Esempio in R

- Applichiamo il metodo gerarchico del centroide alle misurazioni ottenute sui 5 individui I_1, I_2, I_3, I_4, I_5

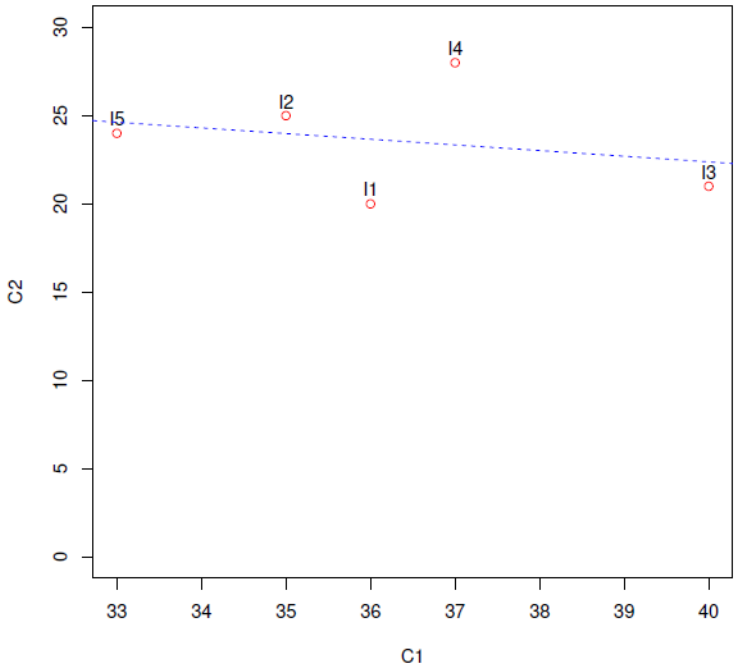
  - Definiamo la matrice dei dati:

$$X = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 36 & 20 \\ 35 & 25 \\ 40 & 21 \\ 37 & 28 \\ 33 & 24 \end{pmatrix} \end{array}$$

```
> X<-data.frame(c1=c(36,35,40,37,33),c2=c(20,25,21,28,24))
> row.names(X)<-c("I1","I2","I3","I4","I5")
> X # visualizza il data frame X
   c1 c2
I1 36 20
I2 35 25
I3 40 21
I4 37 28
I5 33 24
```

  - Rappresentare i cinque punti relativi agli individui I_1, I_2, I_3, I_4, I_5

```
> plot(X$c1,X$c2,col="red",xlab="C1",
+ ylab="C2",ylim=c(0,30))
> text(X$c1,X$c2+0.8,c("I1","I2","I3","I4","I5"))
> abline(lm(X$c2~X$c1),lty=2,col="blue")
```



Esempio in R

- Calcoliamo ora la matrice contenente i quadrati delle distanze euclidee
  - Definiamo la matrice dei dati:

$$X = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 36 & 20 \\ 35 & 25 \\ 40 & 21 \\ 37 & 28 \\ 33 & 24 \end{pmatrix} \end{array}$$

```
> d<-dist(X,method="euclidean",diag=TRUE,upper=TRUE)
>
> d2<-d^2
> d2 # visualizza la matrice con i quadrati delle distanze euclidee
    I1 I2 I3 I4 I5
I1   0 26 17 65 25
I2  26  0 41 13  5
I3  17 41  0 58 58
I4  65 13 58  0 32
I5  25  5 58 32  0
```

Esempio in R

- Calcoliamo ora la matrice contenente i quadrati delle distanze euclidee
  - Definiamo la matrice dei dati:

$$X = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 36 & 20 \\ 35 & 25 \\ 40 & 21 \\ 37 & 28 \\ 33 & 24 \end{pmatrix} \end{array}$$

```
> d<-dist(X,method="euclidean",diag=TRUE,upper=TRUE)
>
> d2<-d^2
> d2 # visualizza la matrice con i quadrati delle distanze euclidee
    I1 I2 I3 I4 I5
I1   0 26 17 65 25
I2  26  0 41 13  5
I3  17 41  0 58 58
I4  65 13 58  0 32
I5  25  5 58 32  0
```

  - Applichiamo ora il metodo gerarchico del centroide utilizzando i quadrati delle distanze euclidee:

```
> hc<-hclust(d2,method="centroid")
> str(hc) # visualizza informazioni sull'oggetto cluster
List of 7
 $ merge      : int [1:4, 1:2] -2 -1 -4 2 -5 -3 1 3
 $ height     : num [1:4] 5 17 21.3 35.7
 $ order      : int [1:5] 1 3 4 2 5
 $ labels     : chr [1:5] "I1" "I2" "I3" "I4" ...
 $ method     : chr "centroid"
 $ call       : language hclust(d = d2, method = "centroid")
 $ dist.method: chr "euclidean"
 - attr(*, "class")= chr "hclust"
```

Esempio in R

- Applichiamo ora il metodo gerarchico del centroide utilizzando i quadrati delle distanze euclidee:

$$X = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 36 & 20 \\ 35 & 25 \\ 40 & 21 \\ 37 & 28 \\ 33 & 24 \end{pmatrix} \end{array}$$

```
> hc<-hclust(d2,method="centroid")
> str(hc) # visualizza informazioni sull'oggetto cluster
List of 7
 $ merge      : int [1:4, 1:2] -2 -1 -4 2 -5 -3 1 3
 $ height     : num [1:4] 5 17 21.3 35.7
 $ order      : int [1:5] 1 3 4 2 5
 $ labels     : chr [1:5] "I1" "I2" "I3" "I4" ...
 $ method     : chr "centroid"
 $ call       : language hclust(d = d2, method = "centroid")
 $ dist.method: chr "euclidean"
 - attr(*, "class")= chr "hclust"
```

		Agglomerazione	Distanza
-2	-5	Al livello 1 si uniscono gli individui I_2 e I_5	5.0

Esempio in R

- Applichiamo ora il metodo gerarchico del centroide utilizzando i quadrati delle distanze euclidee:

$$X = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 36 & 20 \\ 35 & 25 \\ 40 & 21 \\ 37 & 28 \\ 33 & 24 \end{pmatrix} \end{array}$$

```
> hc<-hclust(d2,method="centroid")
> str(hc) # visualizza informazioni sull'oggetto cluster
List of 7
 $ merge      : int [1:4, 1:2] -2 -1 -4 2 -5 -3 1 3
 $ height     : num [1:4] 5 17 21.3 35.7
 $ order      : int [1:5] 1 3 4 2 5
 $ labels     : chr [1:5] "I1" "I2" "I3" "I4" ...
 $ method     : chr "centroid"
 $ call       : language hclust(d = d2, method = "centroid")
 $ dist.method: chr "euclidean"
 - attr(*, "class")= chr "hclust"
```

		Agglomerazione	Distanza
-2	-5	Al livello 1 si uniscono gli individui I_2 e I_5	5.0
-1	-3	Al livello 2 si uniscono gli individui I_1 e I_3	17.0

Esempio in R

- Applichiamo ora il metodo gerarchico del centroide utilizzando i quadrati delle distanze euclidee:

$$X = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 36 & 20 \\ 35 & 25 \\ 40 & 21 \\ 37 & 28 \\ 33 & 24 \end{pmatrix} \end{array}$$

```
> hc<-hclust(d2,method="centroid")
> str(hc) # visualizza informazioni sull'oggetto cluster
List of 7
 $ merge      : int [1:4, 1:2] -2 -1 -4 2 -5 -3 1 3
 $ height     : num [1:4] 5 17 21.3 35.7
 $ order      : int [1:5] 1 3 4 2 5
 $ labels     : chr [1:5] "I1" "I2" "I3" "I4" ...
 $ method     : chr "centroid"
 $ call       : language hclust(d = d2, method = "centroid")
 $ dist.method: chr "euclidean"
 - attr(*, "class")= chr "hclust"
```

		Agglomerazione	Distanza
-2	-5	Al livello 1 si uniscono gli individui I_2 e I_5	5.0
-1	-3	Al livello 2 si uniscono gli individui I_1 e I_3	17.0
-4	1	Al livello 3 si uniscono l'individuo I_4 con il primo cluster (formato dagli individui I_2 e I_5)	21.3

Esempio in R

- Applichiamo ora il metodo gerarchico del centroide utilizzando i quadrati delle distanze euclidee:

$$X = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 36 & 20 \\ 35 & 25 \\ 40 & 21 \\ 37 & 28 \\ 33 & 24 \end{pmatrix} \end{array}$$

```
> hc<-hclust(d2,method="centroid")
> str(hc) # visualizza informazioni sull'oggetto cluster
List of 7
 $ merge      : int [1:4, 1:2] -2 -1 -4 2 -5 -3 1 3
 $ height     : num [1:4] 5 17 21.3 35.7
 $ order      : int [1:5] 1 3 4 2 5
 $ labels     : chr [1:5] "I1" "I2" "I3" "I4" ...
 $ method     : chr "centroid"
 $ call       : language hclust(d = d2, method = "centroid")
 $ dist.method: chr "euclidean"
 - attr(*, "class")= chr "hclust"
```

		Agglomerazione	Distanza
-2	-5	Al livello 1 si uniscono gli individui I_2 e I_5	5.0
-1	-3	Al livello 2 si uniscono gli individui I_1 e I_3	17.0
-4	1	Al livello 3 si uniscono l'individuo I_4 con il primo cluster (formato dagli individui I_2 e I_5)	21.3
2	3	Al livello 4 si unisce il secondo cluster (formato dagli individui I_1 e I_3) con il terzo cluster (formato dagli individui I_2,I_4 e I_5)	35.7

Esempio in R

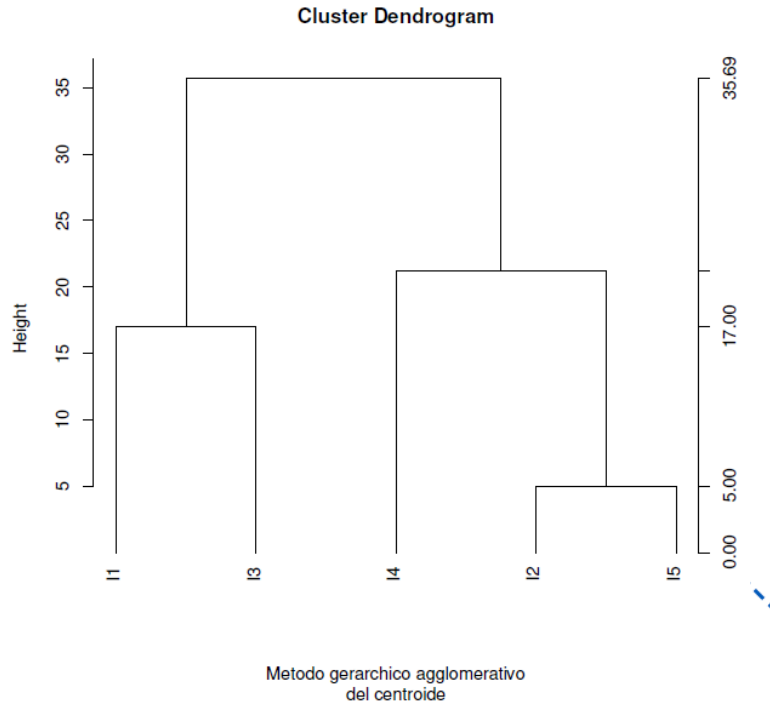
- Applichiamo ora il metodo gerarchico del centroide utilizzando i quadrati delle distanze euclidee:

$$X = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 36 & 20 \\ 35 & 25 \\ 40 & 21 \\ 37 & 28 \\ 33 & 24 \end{pmatrix} \end{array}$$

```
> hc<-hclust(d2,method="centroid")
> str(hc) # visualizza informazioni sull'oggetto cluster
List of 7
 $ merge      : int [1:4, 1:2] -2 -1 -4 2 -5 -3 1 3
 $ height     : num [1:4] 5 17 21.3 35.7
 $ order      : int [1:5] 1 3 4 2 5
 $ labels     : chr [1:5] "I1" "I2" "I3" "I4" ...
 $ method     : chr "centroid"
 $ call       : language hclust(d = d2, method = "centroid")
 $ dist.method: chr "euclidean"
 - attr(*, "class")= chr "hclust"
```

- Costruiamo il Dendrogramma:

```
> plot(hc,hang=-1,xlab="Metodo gerarchico agglomerativo",
+ sub="del centroide")
> axis(side=4,at=round(c(0,hc$height),2))
```



Esempio in R – Sveliamo l'arcano

- Come è avvenuto il processo di agglomerazione con il metodo del centroide?
- Partiamo dalla matrice dei quadrati delle distanze euclidee ottenuta con R

	I1	I2	I3	I4	I5
I1	0	26	17	65	25
I2	26	0	41	13	5
I3	17	41	0	58	58
I4	65	13	58	0	32
I5	25	5	58	32	0

- **Livello 1**: $d^2_{25} = 5$ è il più piccolo valore della matrice delle distanze e pertanto I_2 e I_5 sono uniti formando un unico cluster
- Otteniamo quindi una nuova matrice:

$$X = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 36 & 20 \\ 35 & 25 \\ 40 & 21 \\ 37 & 28 \\ 33 & 24 \end{pmatrix} \end{array}$$

$$d_{I_2 I_5} = \sqrt{(x_2 + x_1)^2 + (y_2 + y_1)^2} =$$

$$= \sqrt{(35 + 33)^2 + (25 + 24)^2} = 5$$

Esempio in R – Sveliamo l'arcano

- Come è avvenuto il processo di agglomerazione con il metodo del centroide?
- Partiamo dalla matrice dei quadrati delle distanze euclidee ottenuta con R

	I1	I2	I3	I4	I5
I1	0	26	17	65	25
I2	26	0	41	13	5
I3	17	41	0	58	58
I4	65	13	58	0	32
I5	25	5	58	32	0

- **Livello 1**: $d^2_{25} = 5$ è il più piccolo valore della matrice delle distanze e pertanto I_2 e I_5 sono uniti formando un unico cluster
- Otteniamo quindi una nuova matrice:

$$X = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 36 & 20 \\ 35 & 25 \\ 40 & 21 \\ 37 & 28 \\ 33 & 24 \end{pmatrix} \end{array} \longrightarrow X_1 = \begin{array}{c} \\ I_1 \\ I_{2,5} \\ I_3 \\ I_4 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 36 & 20 \\ 34 & 24.5 \\ 40 & 21 \\ 37 & 28 \end{pmatrix} \end{array}$$

- Poiché $x_{(2,5),1} = \frac{35+33}{2} = 34$ e $x_{(2,5),2} = \frac{25+24}{2} = 24.5$ si ottiene la nuova matrice X_1

Esempio in R – Sveliamo l'arcano

- Partiamo dalla nuova matrice e calcoliamo la matrice dei quadrati delle distanze euclidee ottenuta con R

	I1	I25	I3	I4
I1	0.00	24.25	17.00	65.00
I25	24.25	0.00	48.25	21.25
I3	17.00	48.25	0.00	58.00
I4	65.00	21.25	58.00	0.00

- **Livello 2**: $d_{13}^2 = 17$ è il più piccolo valore della matrice delle distanze e pertanto I_1 e I_3 sono uniti formando un unico cluster
- Otteniamo quindi una nuova matrice:

$$X_1 = \begin{array}{c} \\ I_1 \\ I_{2,5} \\ I_3 \\ I_4 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 36 & 20 \\ 34 & 24.5 \\ 40 & 21 \\ 37 & 28 \end{pmatrix} \end{array} \longrightarrow X_2 = \begin{array}{c} \\ I_{1,3} \\ I_{2,5} \\ I_4 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 38 & 20.5 \\ 34 & 24.5 \\ 37 & 28 \end{pmatrix} \end{array}$$

- Poiché $x_{(1,3),1} = \frac{36+40}{2} = 38$ e $x_{(2,5),2} = \frac{20+21}{2} = 20.5$ si ottiene la nuova matrice X_2

Esempio in R – Sveliamo l'arcano

- Partiamo dalla nuova matrice e calcoliamo la matrice dei quadrati delle distanze euclidee ottenuta con R

	I13	I25	I4
I13	0.00	32.00	57.25
I25	32.00	0.00	21.25
I4	57.25	21.25	0.00

- **Livello 3**: $d^2_{(2,5),4} = 21.25$ è il più piccolo valore della matrice delle distanze e pertanto I_{25} e I_4 sono uniti formando un unico cluster
- Otteniamo quindi una nuova matrice:

$$X_2 = \begin{array}{c} \\ I_{1,3} \\ I_{2,5} \\ I_4 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 38 & 20.5 \\ 34 & 24.5 \\ 37 & 28 \end{pmatrix} \end{array} \longrightarrow X_3 = \begin{array}{c} \\ I_{1,3} \\ I_{2,4,5} \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 38 & 20.5 \\ 105/3 & 77/3 \end{pmatrix} \end{array}$$

- Poiché $\overline{x_{((2,5),4),1}} = \frac{2*34}{3} + \frac{37}{3} = \frac{105}{3}$ e $\overline{x_{((2,5),4),2}} = \frac{2*24.5}{3} + \frac{28}{3} = \frac{77}{3}$ si ottiene la nuova matrice X_3
- Calcoliamo la matrice dei quadrati delle distanze euclidee ottenendo:
  - Quindi tutti individui si uniscono in un unico cluster ad un livello di distanza

	I13	245
I13	0.00000	35.69444
245	35.69444	0.00000

$$d^2 = 35.69444$$

CLUSTERING GERARCHICO

Metodo della Mediana

Metodo della Mediana

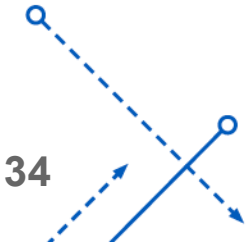
- Il **Metodo della Mediana** è simile al metodo del centroide con la differenza che la procedura è indipendente dalla numerosità dei cluster
- Quando due gruppi si aggregano, il nuovo centroide è calcolato come la semisomma dei due centroidi precedenti
- Il Livello 1 è lo stesso del metodo del centroide
- Ad ogni livello successivo dopo che i cluster G_ue G_v sono stati uniti scegliendo dalla matrice dei quadrati delle distanze euclidee i due cluster più vicini
  - la distanza tra il nuovo cluster G_{uv} e G_z si calcola:

$$d^{(2)}_{(uv),z} = \frac{d^2_{u,z}}{2} + \frac{d^2_{v,z}}{2} - \frac{d^2_{u,v}}{4}$$

Dove:

$$\bar{x}_{(uv),r} = \frac{(\bar{x}_{(u),r} + \bar{x}_{(v),r})}{2} \qquad r = 1,2,\dots,p$$

- La procedura si ripete fino ad ottenere un unico cluster formato da tutti gli individui



Esempio in R

- Calcoliamo ora la matrice contenente i quadrati delle distanze euclidee
  - Definiamo la matrice dei dati:

$$X = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 36 & 20 \\ 35 & 25 \\ 40 & 21 \\ 37 & 28 \\ 33 & 24 \end{pmatrix} \end{array}$$

```
> d<-dist(X,method="euclidean",diag=TRUE,upper=TRUE)
>
> d2<-d^2
> d2 # visualizza la matrice con i quadrati delle distanze euclidee
   I1 I2 I3 I4 I5
I1  0 26 17 65 25
I2 26  0 41 13  5
I3 17 41  0 58 58
I4 65 13 58  0 32
I5 25  5 58 32  0
```

  - Applichiamo ora il metodo gerarchico della mediana utilizzando i quadrati delle distanze euclidee:

```
> hmed<-hclust(d2,method="median")
>
> str(hmed) # visualizza informazioni sull'oggetto cluster
List of 7
 $ merge      : int [1:4, 1:2] -2 -1 -4 2 -5 -3 1 3
 $ height     : num [1:4] 5 17 21.3 39.3
 $ order      : int [1:5] 1 3 4 2 5
 $ labels     : chr [1:5] "I1" "I2" "I3" "I4" ...
 $ method     : chr "median"
 $ call       : language hclust(d = d2, method = "median")
 $ dist.method: chr "euclidean"
 - attr(*, "class")= chr "hclust"
```

Esempio in R

- Calcoliamo ora la matrice contenente i quadrati delle distanze euclidee

```
> hmed<-hclust(d2,method="median")
>
> str(hmed) # visualizza informazioni sull'oggetto cluster
List of 7
 $ merge      : int [1:4, 1:2] -2 -1 -4 2 -5 -3 1 3
 $ height     : num [1:4] 5 17 21.3 39.3
 $ order      : int [1:5] 1 3 4 2 5
 $ labels     : chr [1:5] "I1" "I2" "I3" "I4" ...
 $ method     : chr "median"
 $ call       : language hclust(d = d2, method = "median")
 $ dist.method: chr "euclidean"
 - attr(*, "class")= chr "hclust"
```

$$X = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 36 & 20 \\ 35 & 25 \\ 40 & 21 \\ 37 & 28 \\ 33 & 24 \end{pmatrix} \end{array}$$

		Agglomerazione	Distanza
-2	-5	Al livello 1 si uniscono gli individui I_2 e I_3	5.0

Esempio in R

- Calcoliamo ora la matrice contenente i quadrati delle distanze euclidee

```
> hmed<-hclust(d2,method="median")
>
> str(hmed) # visualizza informazioni sull'oggetto cluster
List of 7
 $ merge      : int [1:4, 1:2] -2 -1 -4 2 -5 -3 1 3
 $ height     : num [1:4] 5 17 21.3 39.3
 $ order      : int [1:5] 1 3 4 2 5
 $ labels     : chr [1:5] "I1" "I2" "I3" "I4" ...
 $ method     : chr "median"
 $ call       : language hclust(d = d2, method = "median")
 $ dist.method: chr "euclidean"
 - attr(*, "class")= chr "hclust"
```

$$X = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 36 & 20 \\ 35 & 25 \\ 40 & 21 \\ 37 & 28 \\ 33 & 24 \end{pmatrix} \end{array}$$

		Agglomerazione	Distanza
-2	-5	Al livello 1 si uniscono gli individui I_2 e I_3	5.0
-1	-3	Al livello 2 si uniscono gli individui I_1 e I_5	17.0

Esempio in R

- Calcoliamo ora la matrice contenente i quadrati delle distanze euclidee

```
> hmed<-hclust(d2,method="median")
>
> str(hmed) # visualizza informazioni sull'oggetto cluster
List of 7
 $ merge      : int [1:4, 1:2] -2 -1 -4 2 -5 -3 1 3
 $ height     : num [1:4] 5 17 21.3 39.3
 $ order      : int [1:5] 1 3 4 2 5
 $ labels     : chr [1:5] "I1" "I2" "I3" "I4" ...
 $ method     : chr "median"
 $ call       : language hclust(d = d2, method = "median")
 $ dist.method: chr "euclidean"
 - attr(*, "class")= chr "hclust"
```

$$X = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 36 & 20 \\ 35 & 25 \\ 40 & 21 \\ 37 & 28 \\ 33 & 24 \end{pmatrix} \end{array}$$

		Agglomerazione	Distanza
-2	-5	Al livello 1 si uniscono gli individui I_2 e I_3	5.0
-1	-3	Al livello 2 si uniscono gli individui I_1 e I_5	17.0
-4	1	Al livello 3 si uniscono l'individuo I_4 con il primo cluster (formato dagli individui I_2 e I_5)	21.3

Esempio in R

- Calcoliamo ora la matrice contenente i quadrati delle distanze euclidee

```
> hmed<-hclust(d2,method="median")
>
> str(hmed) # visualizza informazioni sull'oggetto cluster
List of 7
 $ merge      : int [1:4, 1:2] -2 -1 -4 2 -5 -3 1 3
 $ height     : num [1:4] 5 17 21.3 39.3
 $ order      : int [1:5] 1 3 4 2 5
 $ labels     : chr [1:5] "I1" "I2" "I3" "I4" ...
 $ method     : chr "median"
 $ call       : language hclust(d = d2, method = "median")
 $ dist.method: chr "euclidean"
 - attr(*, "class")= chr "hclust"
```

$$X = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 36 & 20 \\ 35 & 25 \\ 40 & 21 \\ 37 & 28 \\ 33 & 24 \end{pmatrix} \end{array}$$

		Agglomerazione	Distanza
-2	-5	Al livello 1 si uniscono gli individui I_2 e I_3	5.0
-1	-3	Al livello 2 si uniscono gli individui I_1 e I_5	17.0
-4	1	Al livello 3 si uniscono l'individuo I_4 con il primo cluster (formato dagli individui I_2 e I_5)	21.3
2	3	Al livello 4 si unisce il secondo cluster (formato dagli individui I_1 e I_3) con il terzo cluster (formato dagli individui I_2,I_4 e I_5)	39.3

Esempio in R

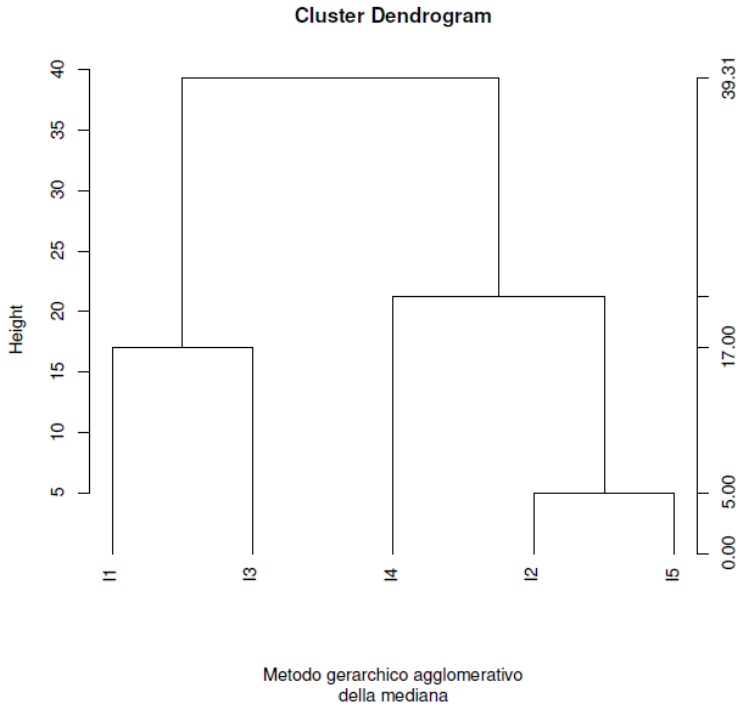
- Calcoliamo ora la matrice contenente i quadrati delle distanze euclidee

```
> hmed<-hclust(d2,method="median")
>
> str(hmed) # visualizza informazioni sull'oggetto cluster
List of 7
 $ merge      : int [1:4, 1:2] -2 -1 -4 2 -5 -3 1 3
 $ height     : num [1:4] 5 17 21.3 39.3
 $ order      : int [1:5] 1 3 4 2 5
 $ labels     : chr [1:5] "I1" "I2" "I3" "I4" ...
 $ method     : chr "median"
 $ call       : language hclust(d = d2, method = "median")
 $ dist.method: chr "euclidean"
 - attr(*, "class")= chr "hclust"
```

$$X = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 36 & 20 \\ 35 & 25 \\ 40 & 21 \\ 37 & 28 \\ 33 & 24 \end{pmatrix} \end{array}$$

- Costruiamo ora il dendrogramma utilizzando le seguenti linee di codice:

```
> plot(hmed,hang=-1,xlab="Metodo gerarchico agglomerativo",
+ sub="della mediana")
> axis(side=4,at=round(c(0,hmed$height),2))
```



Esempio in R – Sveliamo l'arcano

- Come è avvenuto il processo di agglomerazione con il metodo della mediana?
- Partiamo dalla matrice dei quadrati delle distanze euclidee ottenuta con R

	I1	I2	I3	I4	I5
I1	0	26	17	65	25
I2	26	0	41	13	5
I3	17	41	0	58	58
I4	65	13	58	0	32
I5	25	5	58	32	0

- **Livello 1**: $d_{25}^2 = 5$ è il più piccolo valore della matrice delle distanze e pertanto I_2 e I_5 sono uniti formando un unico cluster
- Otteniamo quindi una nuova matrice:

$$X = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 36 & 20 \\ 35 & 25 \\ 40 & 21 \\ 37 & 28 \\ 33 & 24 \end{pmatrix} \end{array} \longrightarrow X_1 = \begin{array}{c} \\ I_1 \\ I_{2,5} \\ I_3 \\ I_4 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 36 & 20 \\ 34 & 24.5 \\ 40 & 21 \\ 37 & 28 \end{pmatrix} \end{array}$$

$$\text{mediana}(C_1) = \frac{33+35}{2} = 34.0 \qquad \text{mediana}(C_2) = \frac{24+25}{2} = 24.5$$

$$M_A = (34.0, 24.5)$$

Esempio in R

- Calcoliamo ora la matrice contenente i quadrati delle distanze euclidee
  - Definiamo la matrice dei dati:

$$X = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 36 & 20 \\ 35 & 25 \\ 40 & 21 \\ 37 & 28 \\ 33 & 24 \end{pmatrix} \end{array}$$

$$d^2(I_1, I_2) = 1 + 25 = 26$$
$$d^2(I_1, I_3) = 16 + 1 = 17$$
$$d^2(I_1, I_4) = 1 + 64 = 65$$
$$d^2(I_1, I_5) = 9 + 16 = 25$$
$$d^2(I_2, I_3) = 25 + 16 = 41$$
$$d^2(I_2, I_4) = 4 + 9 = 13$$
$$d^2(I_2, I_5) = 4 + 1 = \mathbf{5}$$
$$d^2(I_3, I_4) = 9 + 49 = 58$$
$$d^2(I_3, I_5) = 49 + 9 = 58$$
$$d^2(I_4, I_5) = 16 + 16 = 32$$

→

	I1	I2	I3	I4	I5
I1	0	26	17	65	25
I2	26	0	41	13	5
I3	17	41	0	58	58
I4	65	13	58	0	32
I5	25	5	58	32	0

  - La distanza minima è tra I_2 e I_5, dunque vengono uniti in:

$$A = \{I_2, I_5\}, \qquad M_A = \left(\frac{33+35}{2}, \frac{24+25}{2}\right) = (34.0, 24.5)$$

Esempio in R – Sveliamo l'arcano

- Partiamo dalla nuova matrice e calcoliamo la matrice dei quadrati delle distanze euclidee ottenuta con R

$$d^2(I_1, A) = (36-34)^2 + (20-24.5)^2 = 4 + 20.25 = 24.25$$
$$d^2(I_3, A) = (40-34)^2 + (21-24.5)^2 = 36 + 12.25 = 48.25$$
$$d^2(I_4, A) = (37-34)^2 + (28-24.5)^2 = 9 + 12.25 = \mathbf{21.25}$$

→

	I1	I25	I3	I4
I1	0.00	24.25	17.00	65.00
I25	24.25	0.00	48.25	21.25
I3	17.00	48.25	0.00	58.00
I4	65.00	21.25	58.00	0.00

- **Livello 2**: $d^2_{13} = 17$ è il più piccolo valore della matrice delle distanze e pertanto I_1 e I_3 sono uniti formando un unico cluster

$$\text{mediana}(C_1) = \frac{36+40}{2} = 38.0 \qquad \text{mediana}(C_2) = \frac{20+21}{2} = 20.5$$

$$M_B = (38.0, 20.5)$$

- Otteniamo quindi una nuova matrice:

$$X_1 = \begin{array}{c} \\ I_1 \\ I_{2,5} \\ I_3 \\ I_4 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 36 & 20 \\ 34 & 24.5 \\ 40 & 21 \\ 37 & 28 \end{pmatrix} \end{array} \rightarrow X_2 = \begin{array}{c} \\ I_{1,3} \\ I_{2,5} \\ I_4 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 38 & 20.5 \\ 34 & 24.5 \\ 37 & 28 \end{pmatrix} \end{array}$$

Esempio in R – Sveliamo l'arcano

- Partiamo dalla nuova matrice e calcoliamo la matrice dei quadrati delle distanze euclidee ottenuta con R

$$d^2(B,A) = (38-34)^2 + (20.5-24.5)^2 = 16+16 = 32$$
$$d^2(B,I_4) = (38-37)^2 + (20.5-28)^2 = 1+56.25 = 57.25$$
$$d^2(A,I_4) = \mathbf{21.25}$$

→

	I13	I25	I4
I13	0.00	32.00	57.25
I25	32.00	0.00	21.25
I4	57.25	21.25	0.00

- **Livello 3**: $d^2_{(2,5),4} = 21.25$ è il più piccolo valore della matrice delle distanze e pertanto I_{25} e I_4 sono uniti formando un unico cluster

$$X = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 36 & 20 \\ 35 & 25 \\ 40 & 21 \\ 37 & 28 \\ 33 & 24 \end{pmatrix} \end{array}$$

$$I_2 = (35,25), \quad I_5 = (33,24), \quad I_4 = (37,28)$$

Mediana per componente:

$$M_{254} = (\text{med}(35,33,37), \text{med}(25,24,28)) = (35,25)$$

Calcolo della distanza tra i due cluster rimasti:

$$d^2(M_{13}, M_{254}) = (38-35)^2 + (20.5-25)^2 = 9 + 20.25 = 29.25$$

$$D^{(3)} = \begin{array}{c|cc} & I_{13} & I_{254} \\ \hline I_{13} & 0 & 29.25 \\ I_{254} & 29.25 & 0 \end{array}$$

La distanza minima è:

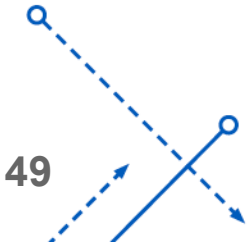
$$\min D^{(3)} = 29.25 = d^2(I_{13}, I_{254})$$

CLUSTERING GERARCHICO

Metodo di Lance e Williams

Metodo di Lance e Williams

- **Lance e Williams** hanno considerato uno schema ricorsivo in cui il calcolo della matrice dei quadrati delle distanze d^2_{k+1} al livello (k+1)-esimo dipende solo dai valori di d^2_k della matrice delle distanze al livello k-esimo
- Questo schema ricorsivo include:
  - i metodi del legame singolo
  - del legame completo
  - del legame medio
  - del centroide
  - della mediana
- Al Livello 0 si considera un insieme di n clusters $\{I_1\}, \{I_2\}, \dots, \{I_n\}$
- Al passo successivo si cerca nella matrice $D^{(2)}$ contenente i quadrati delle singole distanze euclidee, il coefficiente di distanza **minima** e si raggruppano nello stesso cluster G_{ij} i due individui I_i e I_j associati secondo tale coefficiente
- Nel caso i coefficienti di distanza minima siano più di uno, si attua una scelta arbitraria tra di essi.



Metodo di Lance e Williams

- Al Livello $k+1$ con $k = 0, 1, 2, \ldots, n-2$ dopo che i clusters G_u e G_v sono stati uniti scegliendo dalla precedente matrice delle distanze i clusters più vicini, la distanza tra il nuovo cluster G_{uv} e uno C_z è:

$$d^2_{(uv),z} = \alpha_u d^2_{uz} + \alpha_v d^2_{vz} + \beta d^2_{uv} + \gamma |d^2_{uz} - d^2_{vz}|$$

dove $\alpha_u, \alpha_v, \beta, \gamma$ sono dei parametri che dipendono dalla procedura di clustering gerarchico scelta

- Metodo del legame singolo:

$$d^2_{(uv),z} = \frac{1}{2} d^2_{u,z} + \frac{1}{2} d^2_{v,z} - \frac{1}{2} |d^2_{u,z} - d^2_{v,z}| \qquad \alpha_u = \frac{1}{2}, \quad \alpha_v = \frac{1}{2}, \quad \beta = 0, \quad \gamma = -\frac{1}{2}.$$

- Metodo del legame completo:

$$d^2_{(uv),z} = \frac{1}{2} d^2_{u,z} + \frac{1}{2} d^2_{v,z} + \frac{1}{2} |d^2_{u,z} - d^2_{v,z}| \qquad \alpha_u = \frac{1}{2}, \quad \alpha_v = \frac{1}{2}, \quad \beta = 0, \quad \gamma = \frac{1}{2}.$$

- Metodo del legame medio:

$$d^2_{(uv),z} = \frac{N_u}{N_u + N_v} d^2_{u,z} + \frac{N_v}{N_u + N_v} d^2_{v,z} \qquad \alpha_u = \frac{N_u}{N_u + N_v}, \quad \alpha_v = \frac{N_u}{N_u + N_v}, \quad \beta = 0, \quad \gamma = 0.$$

con N_u, N_v il numero di individui in G_u e G_v

Metodo di Lance e Williams

- Al Livello $k+1$ con $k = 0, 1, 2, \ldots, n-2$ dopo che i clusters G_u e G_v sono stati uniti scegliendo dalla precedente matrice delle distanze i clusters più vicini, la distanza tra il nuovo cluster G_{uv} e uno C_z è:

$$d^2_{(uv),z} = \alpha_u d^2_{uz} + \alpha_v d^2_{vz} + \beta d^2_{uv} + \gamma |d^2_{uz} - d^2_{vz}|$$

dove $\alpha_u, \alpha_v, \beta, \gamma$ sono dei parametri che dipendono dalla procedura di clustering gerarchico scelta

- Metodo del centroide:

$$d^2_{(uv),z} = \frac{N_u}{N_u + N_v} d^2_{u,z} + \frac{N_v}{N_u + N_v} d^2_{v,z} - \frac{N_u N_v}{(N_u + N_v)^2} d^2_{u,v}$$

$$\alpha_u = \frac{N_u}{N_u + N_v}, \quad \alpha_v = \frac{N_v}{N_u + N_v}, \quad \beta = -\frac{N_u N_v}{(N_u + N_v)^2}, \quad \gamma = 0.$$

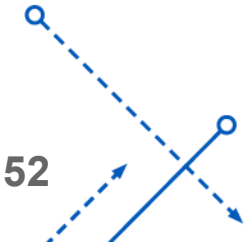
- Metodo della Mediana:

$$d^2_{(uv),z} = \frac{1}{2} d^2_{u,z} + \frac{1}{2} d^2_{v,z} - \frac{1}{4} d^2_{u,v}$$

$$\alpha_u = \frac{1}{2}, \quad \alpha_v = \frac{1}{2}, \quad \beta = -\frac{1}{4}, \quad \gamma = 0.$$

Quale Metodo utilizzare?

- **Metodo del Legame Singolo (Single Linkage)**
  - È utile quando si vuole identificare **cluster di forma allungata** o in presenza di dati che formano **cluster concatenati** (come nelle reti sociali o nei grafi)
  - È adatto per individuare **cluster connessi** da punti vicini, anche se questo può portare all'effetto chaining
  - **Caratteristiche**:
    - Crea cluster unendo i punti più vicini tra i gruppi.
    - Ha il problema dell'**effetto chaining**: tende a formare cluster allungati con punti intermedi anche se i dati sono naturalmente separati
- **Metodo del Legame Completo (Complete Linkage)**
  - È indicato quando si vuole ottenere **cluster compatti e ben separati**.
  - È adatto per dati in cui si desidera **minimizzare** la distanza massima all'interno di ciascun cluster, garantendo che i punti all'interno del cluster siano vicini tra loro
  - **Caratteristiche**:
    - Considera la distanza massima tra punti dei due cluster; tende a formare cluster compatti.
    - È meno sensibile all'effetto chaining e produce cluster con separazioni più nette

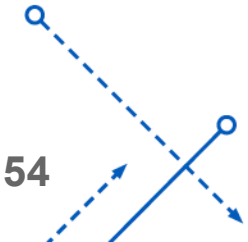


Quale Metodo utilizzare?

- **Metodo del Legame Medio (Average Linkage)**
  - È adatto per dati che non hanno cluster di forma regolare, poiché si adatta bene a forme diverse e minimizza gli svantaggi del legame singolo e del legame completo.
  - Funziona bene per **dati con cluster di forma e densità variabile**
  - **Caratteristiche**:
    - Calcola la media delle distanze tra tutti i punti di un cluster e quelli dell'altro
    - Offre una soluzione bilanciata tra il legame singolo e il legame completo, senza l'effetto chaining ma con una buona coesione dei cluster
- **Metodo del Centroide (Centroid Linkage)**
  - È utile quando si vuole avere un punto centrale rappresentativo per ciascun cluster e i dati hanno una struttura **simmetrica e non troppo densa**.
  - Utile per **dati con cluster sferici** o per analisi esplorativa che richiede di trovare un centro di massa per ogni gruppo
  - **Caratteristiche**:
    - Unisce i cluster in base alla distanza tra i centroidi (medie dei punti nei cluster).
    - I cluster risultanti tendono ad avere una forma più regolare e compatta, ma potrebbero non adattarsi bene a cluster irregolari

Quale Metodo utilizzare?

- **Metodo della Mediana (Median Linkage)**
  - È indicato in presenza di **outlier** o dati rumorosi, poiché la mediana è meno influenzata dai valori estremi rispetto alla media
  - Adatto quando si vuole una maggiore robustezza ai valori anomali.
  - **Caratteristiche**:
    - Utilizza il punto mediano (o mediana) di ciascun cluster per calcolare la distanza tra cluster.
    - Simile al metodo del centroide, ma più resistente agli outlier.

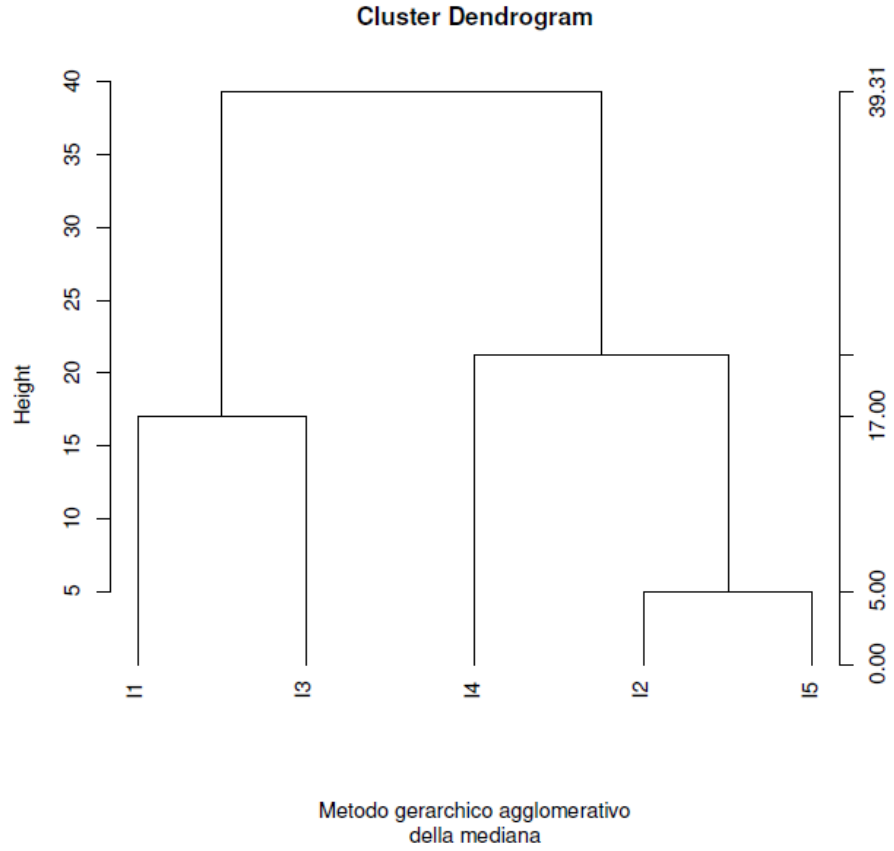


Osservazioni

- La scelta del metodo gerarchico agglomerativo dipende dagli scopi che il ricercatore si propone poichè ogni metodo definisce un diverso concetto di omogeneità all'interno dei cluster
- Non esiste un metodo migliore, ma ogni metodo ha i suoi vantaggi e i suoi svantaggi
  - Se non si ha nessuna informazione sulla struttura dell'insieme da investigare e soprattutto se non si conosce la forma dei cluster da individuare, è interessante applicare il **metodo del legame singolo** e il **metodo del legame completo**
- Con il metodo del **legame completo** i cluster sono sicuramente ben separati ma l'algoritmo privilegia l'omogeneità tra gli elementi interni ai vari gruppi
- Le tecniche di tipo gerarchico sono sicuramente appropriate per dati numerici di **tipo biologico** o **zoologico** per i quali si può ragionevolmente assumere che esista una struttura gerarchica

Osservazioni

- Occorre infine sottolineare che i metodi gerarchici hanno due vantaggi principali:
  - Fornire una visione completa dell'insieme in termini di distanze, seppure condizionata dalla scelta del metodo scelto;
  - Non comportare la scelta a priori del numero di cluster oppure la scelta a priori dei parametri per la determinazione automatica del loro numero.



STATISTICA E ANALISI DEI DATI

Analisi del Dendrogramma

Analisi Dendrogramma

- Consideriamo un particolare dendrogramma ottenuto a partire dalla funzione **hclust**
- La funzione **rect.hclust()** permette di disegnare dei rettangoli intorno ai cluster, individuati in base all'altezza *h* alla quale si opera il taglio del dendrogramma oppure in base al numero *k* di cluster che si vogliono ottenere:

hclust(z, h = NULL, k = NULL, border = "color")

- Dove
  - **z** è l'oggetto creato (output) dalla funzione **hclust**;
  - **h** è l'altezza alla quale si inserisce il taglio;
  - **k** è il numero di cluster che si vogliono ottenere;
  - **border** è il colore dei contorni dei rettangoli.

Analisi Dendrogramma

- Consideriamo la seguente matrice contenente due caratteristiche C_1 e C_2 osservate per 8 individui $I_1, I_2, I_3, I_4, I_5, I_6, I_7, I_8$

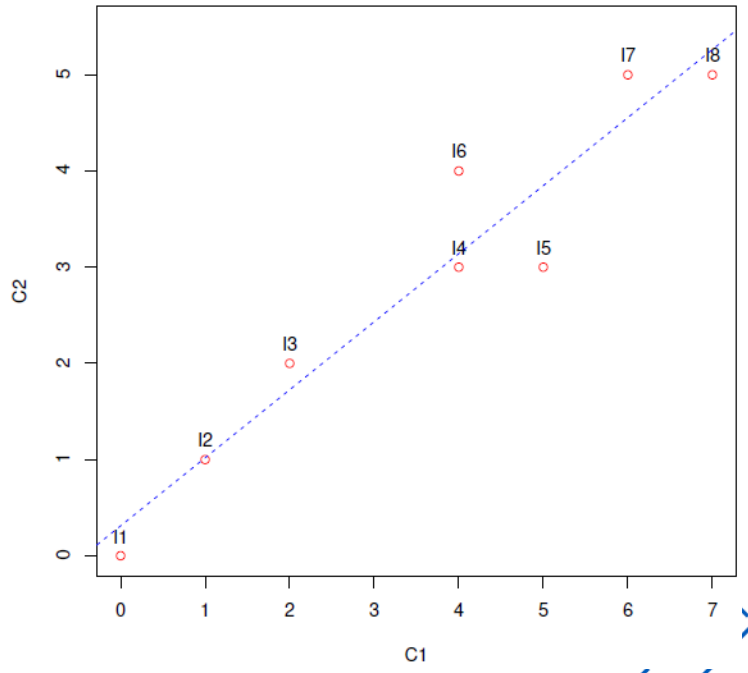
$$X = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \\ I_6 \\ I_7 \\ I_8 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 0 & 0 \\ 1 & 1 \\ 2 & 2 \\ 4 & 3 \\ 5 & 3 \\ 4 & 4 \\ 6 & 5 \\ 7 & 5 \end{pmatrix} \end{array}$$

```
> X<-data.frame(c1=c(0,1,2,4,5,4,6,7),c2=c(0,1,2,3,3,4,5,5))
> row.names(X)<-c("I1","I2","I3","I4","I5","I6","I7","I8")
> X # visualizza il data frame X
   c1 c2
I1  0  0
I2  1  1
I3  2  2
I4  4  3
I5  5  3
I6  4  4
I7  6  5
I8  7  5
```

- Rappresentiamo i punti in uno Scatterplot

```
> plot(X$c1,X$c2,col="red",xlab="C1",
+ ylab="C2",ylim=c(0,5.5))
> text(X$c1,X$c2+0.2,c("I1","I2","I3","I4","I5","I6","I7","I8"))
>
> abline(lm(X$c2~X$c1),lty=2,col="blue")
```

- in cui è anche indicata la retta di regressione



Analisi Dendrogramma

- Calcoliamo ora la matrice contenente i quadrati delle distanze euclidee e applichiamo il metodo gerarchico del centroide

$$X = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \\ I_6 \\ I_7 \\ I_8 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 0 & 0 \\ 1 & 1 \\ 2 & 2 \\ 4 & 3 \\ 5 & 3 \\ 4 & 4 \\ 6 & 5 \\ 7 & 5 \end{pmatrix} \end{array}$$

```
> d<-dist(X,method="euclidean",diag=TRUE,upper=TRUE)
> d2<-d^2
> d2 #visualizza la matrice con i quadrati delle distanze euclidee
    I1  I2  I3  I4  I5  I6  I7  I8
I1   0   2   8  25  34  32  61  74
I2   2   0   2  13  20  18  41  52
I3   8   2   0   5  10   8  25  34
I4  25  13   5   0   1   1   8  13
I5  34  20  10   1   0   2   5   8
I6  32  18   8   1   2   0   5  10
I7  61  41  25   8   5   5   0   1
I8  74  52  34  13   8  10   1   0
>
> tree <- hclust(d2, method = "centroid")
>
> str(tree) # visualizza la struttura dell'oggetto tree
List of 7
 $ merge      : int [1:7, 1:2] -4 -7 -6 -1 -3 2 5 -5 -8 1 ...
 $ height     : num [1:7] 1 1 1.25 2 4.5 ...
 $ order      : int [1:8] 3 1 2 7 8 6 4 5
 $ labels     : chr [1:8] "I1" "I2" "I3" "I4" ...
 $ method     : chr "centroid"
 $ call       : language hclust(d = d2, method = "centroid")
 $ dist.method: chr "euclidean"
 - attr(*, "class")= chr "hclust"
```

Clustering gerarchico

Analisi Dendrogramma

- Calcoliamo ora la matrice contenente i quadrati delle distanze euclidee e applichiamo il metodo gerarchico del centroide.

$$X = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \\ I_6 \\ I_7 \\ I_8 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 0 & 0 \\ 1 & 1 \\ 2 & 2 \\ 4 & 3 \\ 5 & 3 \\ 4 & 4 \\ 6 & 5 \\ 7 & 5 \end{pmatrix} \end{array}$$

```
> d<-dist(X,method="euclidean",diag=TRUE,upper=TRUE)
> d2<-d^2
> d2 #visualizza la matrice con i quadrati delle distanze euclidee
    I1  I2  I3  I4  I5  I6  I7  I8
I1   0   2   8  25  34  32  61  74
I2   2   0   2  13  20  18  41  52
I3   8   2   0   5  10   8  25  34
I4  25  13   5   0   1   1   8  13
I5  34  20  10   1   0   2   5   8
I6  32  18   8   1   2   0   5  10
I7  61  41  25   8   5   5   0   1
I8  74  52  34  13   8  10   1   0
>
> tree <- hclust(d2, method = "centroid")
>
> str(tree) # visualizza la struttura dell'oggetto tree
List of 7
 $ merge      : int [1:7, 1:2] -4 -7 -6 -1 -3 2 5 -5 -8 1 ...
 $ height     : num [1:7] 1 1 1.25 2 4.5 ...
 $ order      : int [1:8] 3 1 2 7 8 6 4 5
 $ labels     : chr [1:8] "I1" "I2" "I3" "I4" ...
 $ method     : chr "centroid"
 $ call       : language hclust(d = d2, method = "centroid")
 $ dist.method: chr "euclidean"
 - attr(*, "class")= chr "hclust"
```

		Agglomerazione	Distanza
-4	-5	Al livello 1 si uniscono gli individui I_4 e I_5 (primo cluster)	1.000000
-7	-8	Al livello 2 si uniscono gli individui I_7 e I_8 (secondo cluster)	1.000000
-6	1	Al livello 3 si uniscono l'individuo I_6 con il primo cluster (formato dagli individui I_4 e I_5) costituendo il terzo cluster	1.250000
-1	-2	Al livello 4 si uniscono gli individui I_1 e I_2 (quarto cluster)	2.000000
-3	4	Al livello 5 si unisce l'individuo I_3 con il quarto cluster (formato dagli individui I_1 e I_2) costituendo il quinto cluster	4.500000
2	3	Al livello 6 si unisce il secondo cluster (formato dagli individui I_7 e I_8) con il terzo cluster (formato dagli individui I_4,I_5 e I_6) costituendo il sesto cluster	7.472222
5	6	Al livello 7 si unisce il quinto cluster (formato dagli individui I_1, I_2 e I_3) con il sesto cluster (formato dagli individui I_4,I_5, I_6, I_7 e I_8)	26.640000

Analisi Dendrogramma

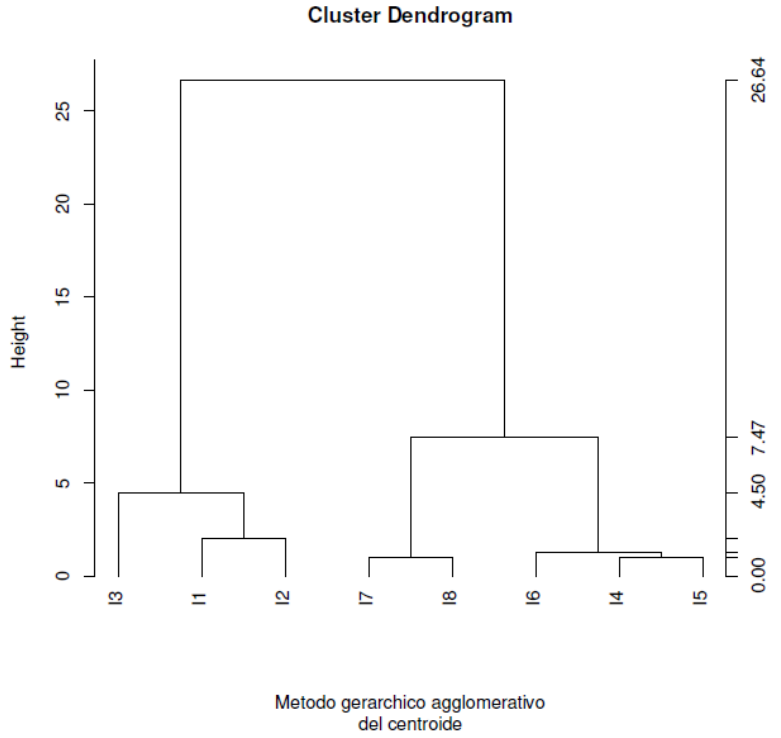
- Costruiamo ora il dendrogramma:

```
> plot(tree,hang=-1,xlab="Metodo gerarchico agglomerativo",
+ sub="del centroide")
> axis(side=4,at=round(c(0,tree$height),2))
```

- La sequenza delle agglomerazioni del metodo del centroide è:

$$X = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \\ I_6 \\ I_7 \\ I_8 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 0 & 0 \\ 1 & 1 \\ 2 & 2 \\ 4 & 3 \\ 5 & 3 \\ 4 & 4 \\ 6 & 5 \\ 7 & 5 \end{pmatrix} \end{array}$$

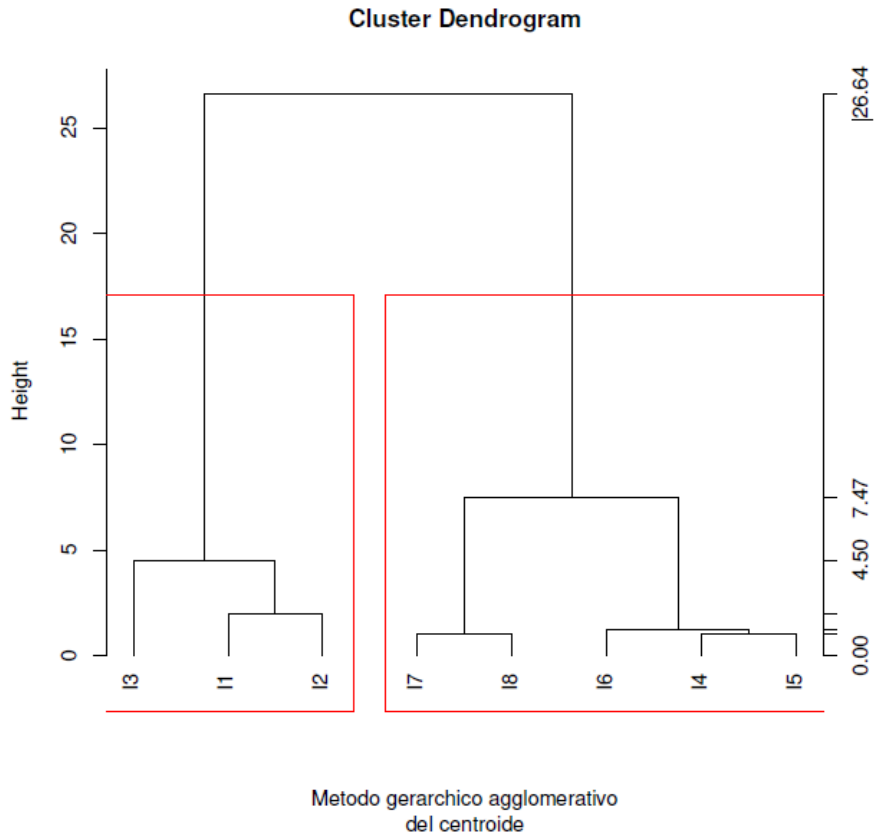
Numero di cluster	Cluster	Livello di distanza
8	$\{I_1\},\{I_2\},\{I_3\},\{I_4\},\{I_5\},\{I_6\},\{I_7\},\{I_8\}$	
7	$\{I_1\},\{I_2\},\{I_3\},\{I_4,I_5\},\{I_6\},\{I_7\},\{I_8\}$	1.000000
6	$\{I_1\},\{I_2\},\{I_3\},\{I_4,I_5\},\{I_6\},\{I_7,I_8\}$	1.000000
5	$\{I_1\},\{I_2\},\{I_3\},\{I_4,I_5,I_6\},\{I_7,I_8\}$	1.250000
4	$\{I_1,I_2\},\{I_3\},\{I_4,I_5,I_6\},\{I_7,I_8\}$	2.000000
3	$\{I_1,I_{2,3}\},\{I_4,I_5,I_6\},\{I_7,I_8\}$	4.500000
2	$\{I_1,I_{2,3}\},\{I_4,I_5,I_6,I_7,I_8\}$	7.472222
1	$\{I_1,I_{2,3},I_4,I_5,I_6,I_7,I_8\}$	26.640000



Analisi Dendrogramma

- Supponiamo di voler evidenziare due cluster mediante rettangoli **rossi**:

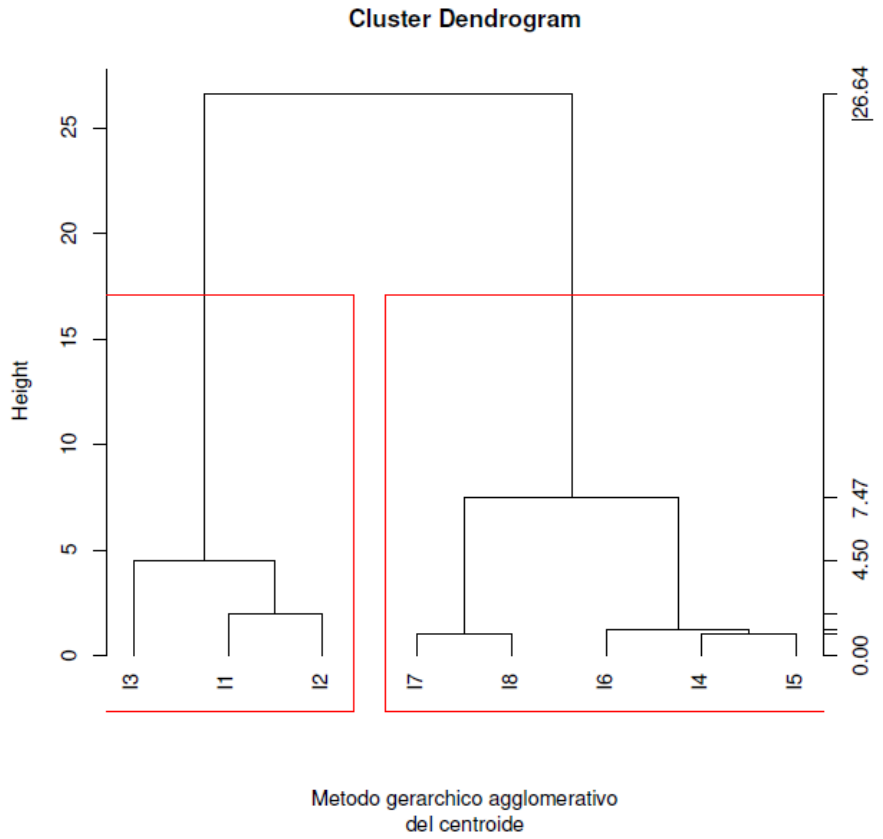
```
> rect.hclust(tree, k = 2, border = "red")
```



Analisi Dendrogramma

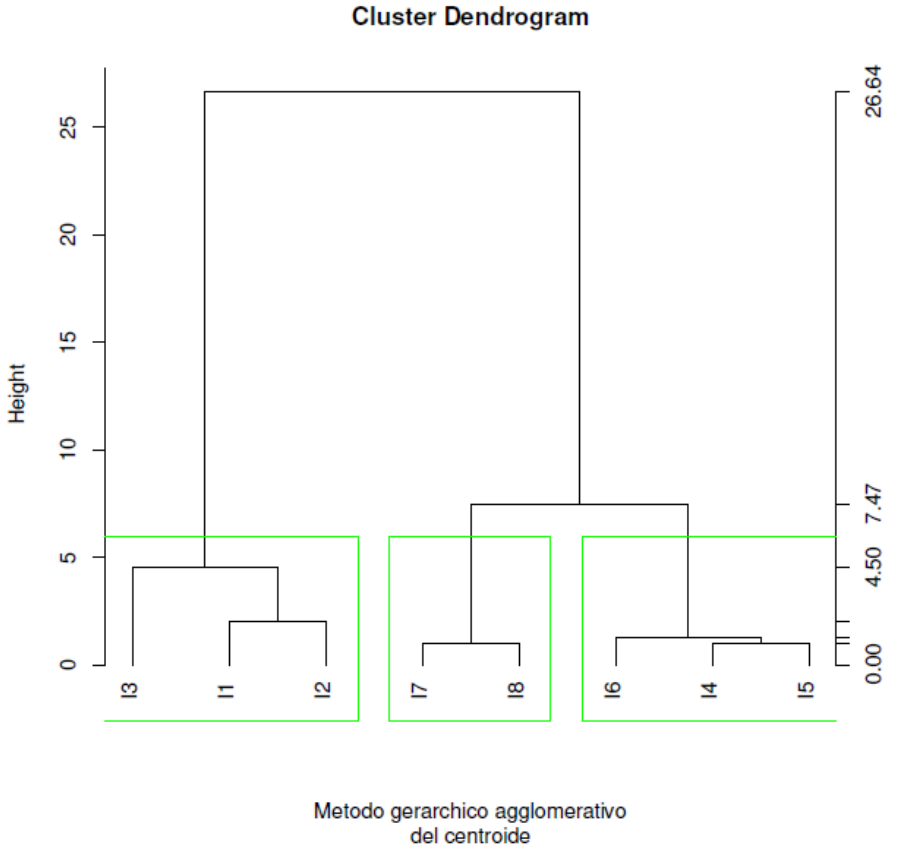
- Supponiamo di voler evidenziare due cluster mediante rettangoli **rossi**:

```
> rect.hclust(tree, k = 2, border = "red")
```



- Supponiamo di voler evidenziare tre cluster mediante rettangoli **verdi**:

```
> rect.hclust(tree, k = 3, border = "green")
```



Inserire gli individui nei cluster

- La funzione **cutree** in R è utilizzata per tagliare un dendrogramma in modo da ottenere un numero specifico di cluster
  - Questa funzione è particolarmente utile nell'analisi di clustering gerarchico, in cui si costruisce un dendrogramma per rappresentare la struttura gerarchica dei dati
- Per ottenere una suddivisione degli individui in cluster in corrispondenza di un determinato livello di distanza oppure in corrispondenza di un prefissato numero di cluster si può utilizzare:

cutree(tree, k = NULL, h = NULL)

Dove

- **tree** è l'oggetto creato (che individua il dendrogramma) dalla funzione **hclust**;
- **h** è l'altezza alla quale il dendrogramma viene tagliato;
- **k** è il numero prefissato di cluster e permette di tagliare il dendrogramma in modo da produrre esattamente k cluster

- Output: della funzione **cutree**() è un vettore contenente numeri interi positivi associati ai cluster in cui sono stati inseriti i vari individui

Inserire gli individui nei cluster

- Esempio:
  - Considerando l'esempio precedente con la matrice X contenente due caratteristiche C_1 e C_2 osservate per 8 individui $I_1, I_2, I_3, I_4, I_5, I_6, I_7, I_8$

$$X = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \\ I_6 \\ I_7 \\ I_8 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 0 & 0 \\ 1 & 1 \\ 2 & 2 \\ 4 & 3 \\ 5 & 3 \\ 4 & 4 \\ 6 & 5 \\ 7 & 5 \end{pmatrix} \end{array}$$

```
> cutree(tree, k = 2, h = NULL)
I1 I2 I3 I4 I5 I6 I7 I8
 1  1  1  2  2  2  2  2
```

  - La funzione cutree individua la seguente partizione in due cluster:

$$G_1 = \{I_1, I_2, I_3\} \quad G_2 = \{I_4, I_5, I_6, I_7, I_8\}$$

- Se consideriamo un livello di distanza fisssato, ma non definiamo il numero di cluster, abbiamo:

```
> cutree(tree, k = NULL, h =15)
I1 I2 I3 I4 I5 I6 I7 I8
 1  1  1  2  2  2  2  2
```

  - La funzione cutree individua la seguente partizione in due cluster:

$$G_1 = \{I_1, I_2, I_3\} \quad G_2 = \{I_4, I_5, I_6, I_7, I_8\}$$

Numero di unità nei cluster

- Per ottenere il numero di unità (individui) in ciascun cluster si può applicare la funzione **table**() al risultato della funzione **cutree**() ottenendo:

```
> table(cutree(tree, k =2, h = NULL))
1 2
3 5
```

ID Cluster

che mostra che nel primo cluster sono presenti tre individui e nel secondo cluster cinque individui

$$X = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \\ I_6 \\ I_7 \\ I_8 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 0 & 0 \\ 1 & 1 \\ 2 & 2 \\ 4 & 3 \\ 5 & 3 \\ 4 & 4 \\ 6 & 5 \\ 7 & 5 \end{pmatrix} \end{array}$$

- La funzione *cutree*() con un vettore di valori come parametro *k:*

```
> cutree(tree, k = 1:8)
   1 2 3 4 5 6 7 8
I1 1 1 1 1 1 1 1 1
I2 1 1 1 1 2 2 2 2
I3 1 1 1 2 3 3 3 3
I4 1 2 2 3 4 4 4 4
I5 1 2 2 3 4 4 4 5
I6 1 2 2 3 4 5 5 6
I7 1 2 3 4 5 6 6 7
I8 1 2 3 4 5 6 7 8
```

Permette di individuare i cluster degli individui all'aumentare del numero di cluster

Misure di sintesi associate ai cluster

- E' possibile ricavare misure di sintesi (ad esempio, la **media campionaria**, la **varianza campionaria**, la **deviazione standard**, etc.) sulle colonne dei singoli cluster ottenuti tagliando il dendrogramma tramite la funzione cutree()
- Si utilizza la funzione aggregate():

aggregate(X, by, FUNCTION)

dove:

- **X** rappresenta una matrice numerica o un data frame;
- **by** è una lista di indici sulla base dei quali le colonne di X vanno aggregate;
- **FUNCTION** è la funzione da applicare alle colonne di X per i vari gruppi individuati in base a **by**.

- Output:
  - è una struttura contenente i valori ottenuti applicando la funzione FUN (ad esempio, la media campionaria, la varianza campionaria, la deviazione standard, etc.) ad ognuna delle caratteristiche associate ai diversi cluster che sono stati aggregati.

Misure di sintesi associate ai cluster

- Esempio:

$$X = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \\ I_6 \\ I_7 \\ I_8 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 0 & 0 \\ 1 & 1 \\ 2 & 2 \\ 4 & 3 \\ 5 & 3 \\ 4 & 4 \\ 6 & 5 \\ 7 & 5 \end{pmatrix} \end{array}$$

```
> X<-data.frame(c1=c(0,1,2,4,5,4,6,7),c2=c(0,1,2,3,3,4,5,5))
> row.names(X)<-c("I1","I2","I3","I4","I5","I6","I7","I8")
> X # visualizza il data frame X
> d<-dist(X,method="euclidean",diag=TRUE,upper=TRUE)
> d2<-d^2
> d2 #visualizza la matrice con i quadrati delle distanze euclidee
    I1 I2 I3 I4 I5 I6 I7 I8
I1   0  2  8 25 34 32 61 74
I2   2  0  2 13 20 18 41 52
I3   8  2  0  5 10  8 25 34
I4  25 13  5  0  1  1  8 13
I5  34 20 10  1  0  2  5  8
I6  32 18  8  1  2  0  5 10
I7  61 41 25  8  5  5  0  1
I8  74 52 34 13  8 10  1  0
>
> tree <- hclust(d2, method = "centroid")
>
> str(tree) # visualizza la struttura dell'oggetto tree
List of 7
 $ merge       : int [1:7, 1:2] -4 -7 -6 -1 -3 2 5 -5 -8 1 ...
 $ height      : num [1:7] 1 1 1.25 2 4.5 ...
 $ order       : int [1:8] 3 1 2 7 8 6 4 5
 $ labels      : chr [1:8] "I1" "I2" "I3" "I4" ...
 $ method      : chr "centroid"
 $ call        : language hclust(d = d2, method = "centroid")
 $ dist.method: chr "euclidean"
 - attr(*, "class")= chr "hclust"
```

→ Clustering gerarchico

Misure di sintesi associate ai cluster

- Calcoliamo le medie campionarie, le varianze campionarie e le deviazioni standard delle caratteristiche dei due cluster precedentemente individuati:

$$G_1 = \{I_1, I_2, I_3\} \quad G_2 = \{I_4, I_5, I_6, I_7, I_8\}$$

$$X = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \\ I_6 \\ I_7 \\ I_8 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 0 & 0 \\ 1 & 1 \\ 2 & 2 \\ 4 & 3 \\ 5 & 3 \\ 4 & 4 \\ 6 & 5 \\ 7 & 5 \end{pmatrix} \end{array}$$

```
> taglio<-cutree(tree, k =2, h = NULL)
> tagliolist<-list(taglio) # lista di indici  per  i gruppi
>
> aggregate(X, tagliolist, mean)
  Group.1  c1 c2
1       1 1.0  1
2       2 5.2  4
>
> aggregate(X, tagliolist, var)
  Group.1  c1 c2
1       1 1.0  1
2       2 1.7  1
>
> aggregate(X, tagliolist, sd)
  Group.1      c1 c2
1       1 1.00000  1
2       2 1.30384  1
```

Misure di sintesi associate ai cluster

- Calcoliamo le medie campionarie, le varianze campionarie e le deviazioni standard delle caratteristiche dei due cluster precedentemente individuati:

$$G_1 = \{I_1, I_2, I_3\} \quad G_2 = \{I_4, I_5, I_6, I_7, I_8\}$$

$$X = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \\ I_6 \\ I_7 \\ I_8 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 0 & 0 \\ 1 & 1 \\ 2 & 2 \\ 4 & 3 \\ 5 & 3 \\ 4 & 4 \\ 6 & 5 \\ 7 & 5 \end{pmatrix} \end{array}$$

```
> taglio<-cutree(tree, k =2, h = NULL)
> tagliolist<-list(taglio) # lista di indici  per  i gruppi
>
> aggregate(X, tagliolist, mean)
  Group.1  c1  c2
1       1 1.0   1
2       2 5.2   4
>
> aggregate(X, tagliolist, var)
  Group.1  c1  c2
1       1 1.0   1
2       2 1.7   1
>
> aggregate(X, tagliolist, sd)
  Group.1      c1  c2
1       1 1.00000   1
2       2 1.30384   1
```

Coordinate centroide G_1

Coordinate centroide G_2

Rappresentare graficamente i cluster

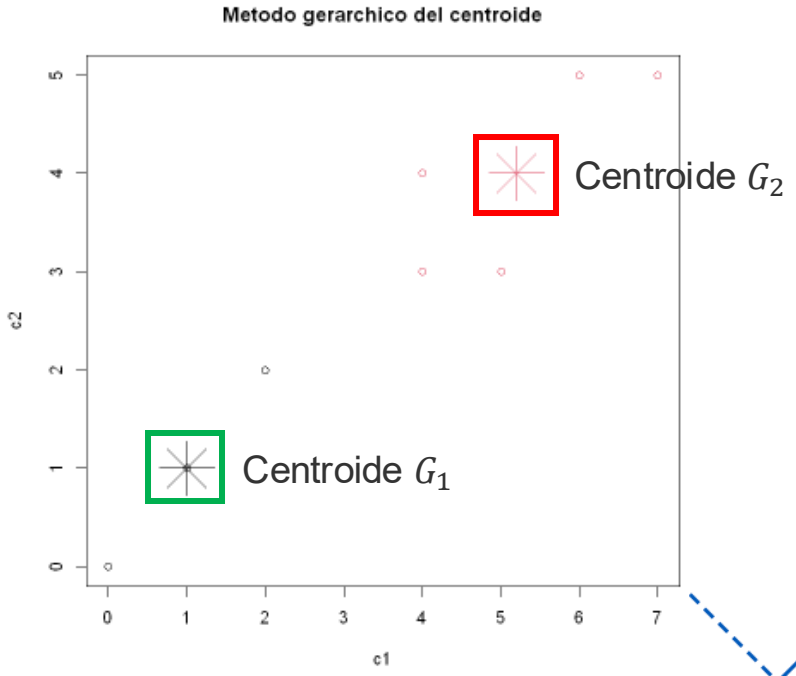
- Se esistono due sole caratteristiche nella matrice dei dati, per rappresentare graficamente i due cluster ottenuti con la funzione **cutree**() ed anche i centroidi dei due cluster, si utilizzano le seguenti linee di codice:

$$X = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \\ I_6 \\ I_7 \\ I_8 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 0 & 0 \\ 1 & 1 \\ 2 & 2 \\ 4 & 3 \\ 5 & 3 \\ 4 & 4 \\ 6 & 5 \\ 7 & 5 \end{pmatrix} \end{array}$$

```
> agmean<-aggregate(X, tagliolist, mean)[, -1]
>
> plot(X, col = taglio, main = "Metodo gerarchico del centroide")
> points(agmean,col = 1:2,pch=8,cex=1)
```

dove

- **X** è la matrice delle misure
- **taglio** è l'oggetto creato con la funzione cutree()
- **col**() individua i colori da associare ai differenti cluster
- **pch** controlla il tipo di carattere (marker) da utilizzare
- **cex** la grandezza del testo e dei simboli generati

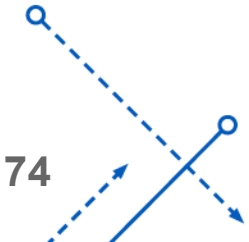


STATISTICA E ANALISI DEI DATI

Screeplot

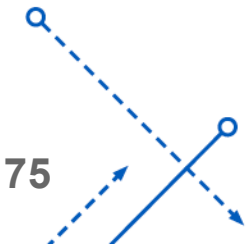
Screeplot

- Un metodo euristico per scegliere una buona partizione del dendrogramma considera una procedura empirica consistente nel costruire un grafico, detto ***screeplot***
  - sull'asse delle ordinate i numeri di gruppi ottenibili con il metodo gerarchico
  - sull'asse delle ascisse le distanze a cui avvengono le successive aggregazioni tra i gruppi
- Se nel passaggio da *k* gruppi a *k−1* gruppi si registra un forte incremento della distanza di aggregazione è consigliabile tagliare il dendrogramma in *k* gruppi
- È possibile realizzare lo screeplot quando si usa il metodo del legame singolo, del legame completo o del legame medio in cui è utilizzata la funzione distanza



Screeplot

- Un metodo euristico per scegliere una buona partizione del dendrogramma considera una procedura empirica consistente nel costruire un grafico, detto ***screeplot***
  - sull'asse delle ordinate i numeri di gruppi ottenibili con il metodo gerarchico
  - sull'asse delle ascisse le distanze a cui avvengono le successive aggregazioni tra i gruppi
- Se nel passaggio da *k* gruppi a *k−1* gruppi si registra un forte incremento della distanza di aggregazione è consigliabile tagliare il dendrogramma in *k* gruppi
- È possibile realizzare lo screeplot quando si usa il metodo del legame singolo, del legame completo o del legame medio in cui è utilizzata la funzione distanza
- Nel metodo del centroide e della mediana (che utilizzano i quadrati delle distanze) le successive agglomerazioni potrebbero verificarsi ad un livello di distanza minore o uguale rispetto alle precedenti agglomerazioni (**e quindi la misura potrebbe essere meno precisa**)
- **Nota**: La procedura empirica basata sullo screeplot non sempre fornisce la suddivisione in cluster più adeguata; è sempre preferibile utilizzare le misure di non omogeneità statistiche precedentemente descritte



Screeplot

- Consideriamo:

$$\delta_k = d_{k-1} - d_k \qquad (k = 2, \dots, n)$$

Dove d_k rappresenta il livello di distanza a cui è stata effettuata l'agglomerazione in *k* gruppi e *n* è il numero iniziale di individui

- Quando δ_k è elevato significa che i due gruppi d_{k-1} e d_k sono sufficientemente dissimili
  - Pertanto è possibile tagliare il dendrogramma all'altezza (al livello di distanza) corrispondente alla partizione in *k* gruppi
- Lo Screeplot fornisce una visione di insieme delle altezze a cui sono avvenute le agglomerazioni e si potrebbe scegliere il valore di *j* per il quale:

$$\delta_j = \max(\delta_2, \delta_3, \dots, \delta_n)$$

Screeplot

- Consideriamo la seguente matrice contenente due caratteristiche C_1 e C_2 osservate per 8 individui $I_1, I_2, I_3, I_4, I_5, I_6, I_7, I_8$

$$X = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \\ I_6 \\ I_7 \\ I_8 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 0 & 0 \\ 1 & 1 \\ 2 & 2 \\ 4 & 3 \\ 5 & 3 \\ 4 & 4 \\ 6 & 5 \\ 7 & 5 \end{pmatrix} \end{array}$$

```
> X<-data.frame(c1=c(0,1,2,4,5,4,6,7),c2=c(0,1,2,3,3,4,5,5))
> row.names(X)<-c("I1","I2","I3","I4","I5","I6","I7","I8")
> X # visualizza il data frame X
   c1 c2
I1  0  0
I2  1  1
I3  2  2
I4  4  3
I5  5  3
I6  4  4
I7  6  5
I8  7  5
```

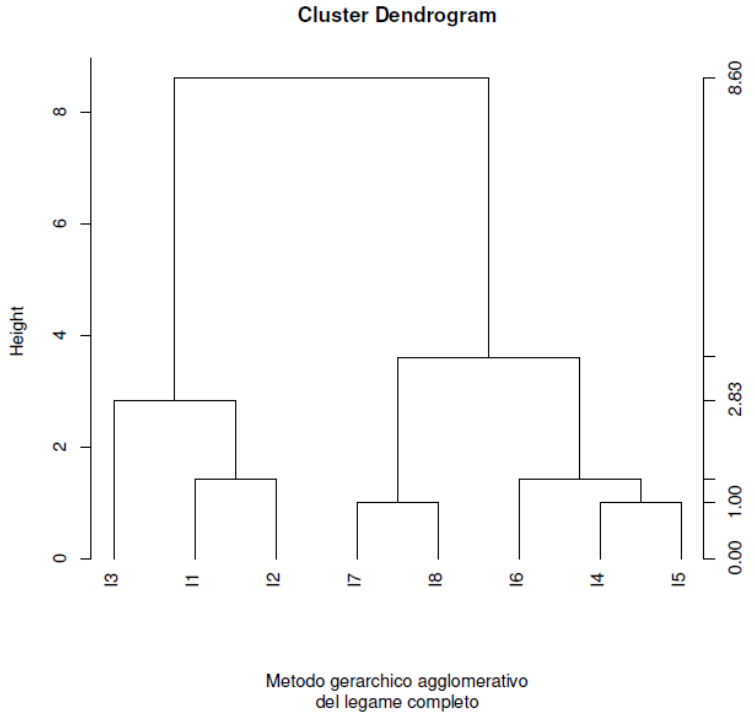
- Calcoliamo ora la matrice delle distanze euclidee e applichiamo il metodo gerarchico del legame completo

```
> d<-dist(X,method="euclidean",diag=TRUE,upper=TRUE)
>
> hlc<-hclust(d,method="complete")
> hlc$height
[1] 1.000000 1.000000 1.414214 1.414214 2.828427 3.605551 8.602325
```

Screeplot

- Costruiamo ora il dendrogramma:

```
> plot(hlc,hang=-1,xlab="Metodo gerarchico agglomerativo",
+ sub="del legame completo")
> axis(side=4,at=round(c(0,hlc$height),2))
```



Screeplot

- Costruiamo ora il dendrogramma:

```
> plot(hlc,hang=-1,xlab="Metodo gerarchico agglomerativo",
+ sub="del legame completo")
> axis(side=4,at=round(c(0,hlc$height),2))
```

- Costruiamo lo screeplot:

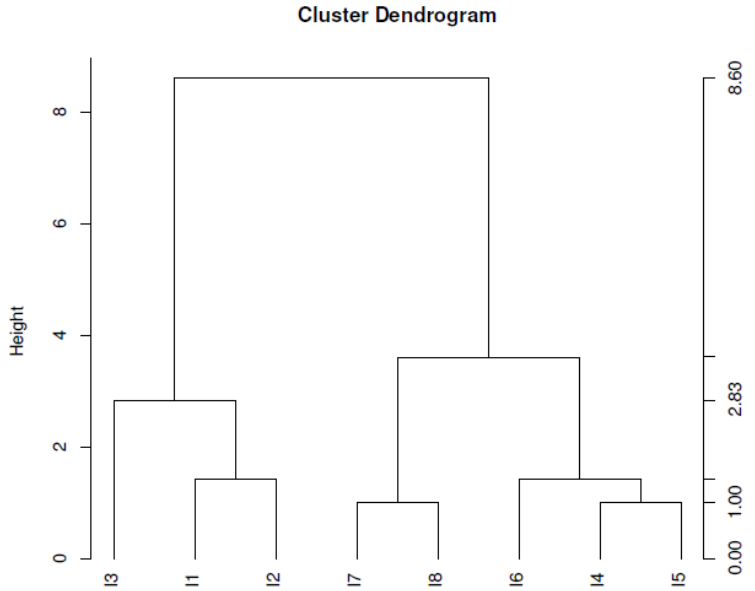
```
> plot(c(0,hlc$height),seq(8,1),type="b",
+ main="Screeplot",xlab="Distanza di aggregazione",
+ ylab="Numero di cluster", col="red")
```

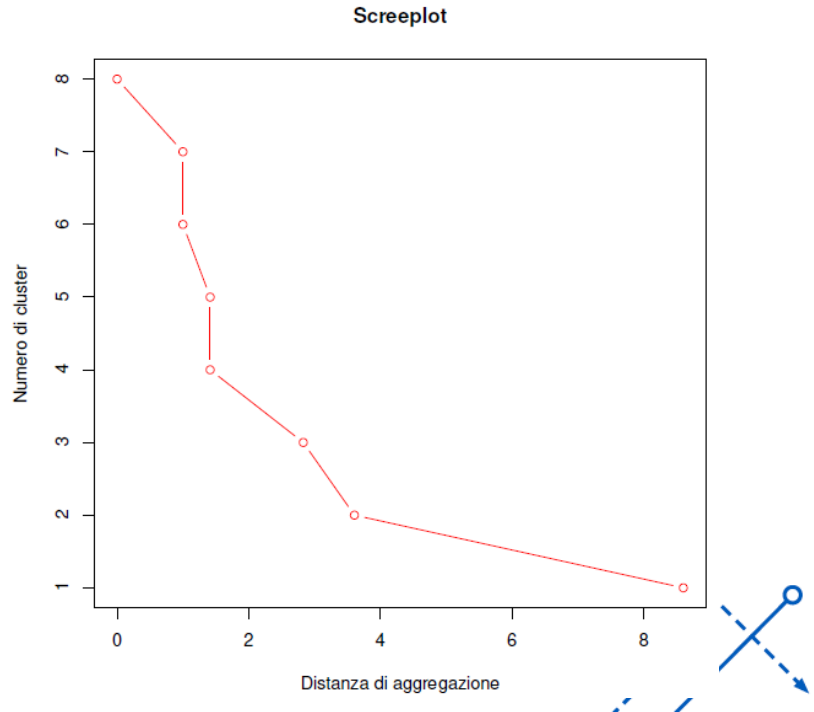
  - La funzione **c(0, hlc$height)** permette di concatenare 0 con il vettore **hlc$height** delle altezze a cui sono avvenute le successive agglomerazioni da cui si ottiene il seguente vettore di lunghezza 8:

$$0 \quad 1.000000 \quad 1.000000 \quad 1.414214 \quad 1.414214 \quad 2.828427 \quad 3.605551 \quad 8.602325$$

  che corrispondono alle aggregazioni in 8 gruppi, in 7 gruppi,…, 1 gruppo

  - La funzione **seq(8, 1)** permette di costruire il vettore contenente il numero di gruppi da 8 a 1
  - **type = ”b”** permette di connettere con delle linee i vari punti





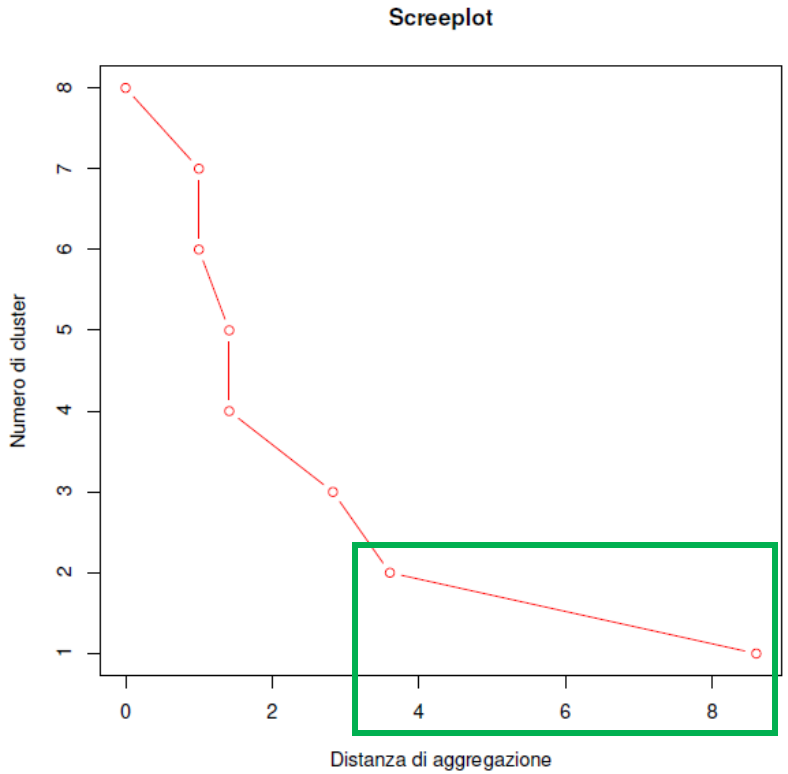
Screeplot

- Come possiamo lo Screeplot suggerisce di considerare una suddivisione in due gruppi
  - Nel passaggio da uno (altezza 8.602325) a due gruppi (altezza 3.605551) si registra un consistente incremento della distanza di aggregazione
- Inoltre, si ha:

$$\delta_2 = h_1 - h_2 = 8.602325 - 3.605551 = 4.996774$$
$$\delta_3 = h_2 - h_3 = 3.605551 - 2.828427 = 0.777124$$
$$\delta_4 = h_3 - h_4 = 2.828427 - 1.414214 = 1.414213$$
$$\delta_5 = h_4 - h_5 = 1.414214 - 1.414214 = 0$$
$$\delta_6 = h_5 - h_6 = 1.414214 - 1.000000 = 0.414214$$
$$\delta_7 = h_6 - h_7 = 1.000000 - 1.000000 = 0$$
$$\delta_8 = h_7 - h_8 = 1.000000 - 0.000000 = 1.$$

che mostra che il valore di *k* per il quale δ_k è massima è *k = 2*

- Cioè è preferibile considerare la divisione in 2 cluster

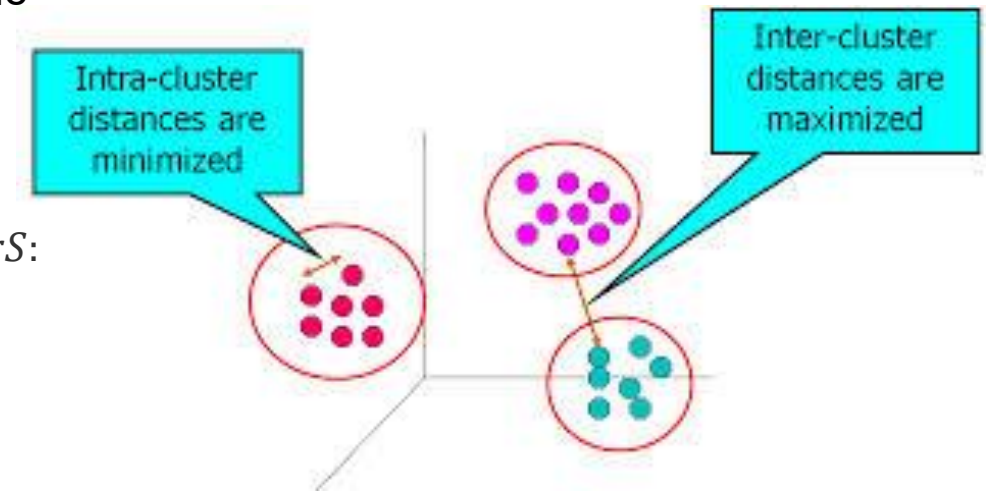


STATISTICA E ANALISI DEI DATI

Misure di non omogeneità statistiche

Misure di non omogeneità statistiche

- Le **misure di non omogeneità statistica** servono fondamentalmente a studiare **quanto i dati di un insieme siano "vari" o "diversi" tra loro**
  - In particolare, nel contesto del clustering, dove suddividiamo gli individui in gruppi (cluster), queste misure permettono di valutare **quanto bene** è stata effettuata la suddivisione
- **Valutare l'omogeneità interna dei cluster**
  - Quanto gli elementi dentro ogni cluster sono simili tra loro.
    - Questa è la **non omogeneità entro i cluster**, indicata come trS:
- **Valutare la separazione tra cluster**
  - Quanto i diversi cluster sono distanti tra loro.
    - Questa è la **non omogeneità tra cluster**, indicata come trB
- **Misurare la variabilità totale del dataset**
  - Quanta "dispersione" c'è tra tutti gli individui rispetto al centro dell'intero dataset
    - Questa è spesso indicata come trT



Misure di non omogeneità statistiche

- Siamo interessati a calcolare le misure di non omogeneità statistica relative:
  - all'insieme totale di individui (trT)
  - alla misura di omogeneità nei cluster (trS)
  - alla misura di omogeneità tra i cluster (trB)

$$trT = trS + trB \text{ equivalente a } 1 = \frac{trS}{trT} + \frac{trB}{trT}$$

- Poichè per ogni fissata matrice X dei dati si ha che la trT è fissata, i cluster dovrebbero essere individuati in modo da:
  - **minimizzare** la misura di non omogeneità statistica all'interno dei cluster (within)
  - **massimizzare** la misura di non omogeneità statistica tra i gruppi (between)
- Se, fissato il numero di cluster, due differenti metodi gerarchici conducono a due diverse partizioni, occorre scegliere quella partizione con misura di non omogeneità statistica all'interno dei cluster (trS) più piccola, che corrisponde a maggiore omogeneità interna

Misure di non omogeneità statistiche

- Esempio:
  - Riprendiamo il dataset X in cui i cluster individuati con il metodo del centroide sono:

$$G_1 = \{I_1, I_2, I_3\} \quad G_2 = \{I_4, I_5, I_6, I_7, I_8\}$$

$$X = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \\ I_6 \\ I_7 \\ I_8 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 0 & 0 \\ 1 & 1 \\ 2 & 2 \\ 4 & 3 \\ 5 & 3 \\ 4 & 4 \\ 6 & 5 \\ 7 & 5 \end{pmatrix} \end{array}$$

  - Sia $I = G_1 \cup G_2$ per l'insieme totale I si ha:

```
> n<-nrow(X)
# n>1
> trHI<-(n-1)*sum(apply(X,2,var))
> trHI# visualizza la misura di non omogeneita' totale
[1] 64.75
```

  - Cioè la misura di non omogeneità statistica totale è quindi $trH_I = 64.75$

Misure di non omogeneità statistiche

- Esempio:
  - Riprendiamo il dataset X in cui i cluster individuati con il metodo del centroide sono:

$$G_1 = \{I_1, I_2, I_3\} \quad G_2 = \{I_4, I_5, I_6, I_7, I_8\}$$

$$X = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \\ I_6 \\ I_7 \\ I_8 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 0 & 0 \\ 1 & 1 \\ 2 & 2 \\ 4 & 3 \\ 5 & 3 \\ 4 & 4 \\ 6 & 5 \\ 7 & 5 \end{pmatrix} \end{array}$$

  - Sia $I = G_1 \cup G_2$ per l'insieme totale I si ha:

```
> n<-nrow(X)
# n>1
> trHI<-(n-1)*sum(apply(X,2,var))
> trHI# visualizza la misura di non omogeneita' totale
[1] 64.75
```

Calcola la varianza per colonne del dataset

  - Cioè la misura di non omogeneità statistica totale è quindi $trH_I = 64.75$

$$\mathrm{tr}H_I = \sum_{r=1}^{p} h_{rr} = (n-1) \sum_{r=1}^{p} s_r^2$$

Misure di non omogeneità statistiche

- Calcoliamo ora le misure di non omogeneità statistiche dei due gruppi G_1 e G_2 seguendo due differenti metodi:
  - **Metodo 1**: occorre definire le matrici dei dati relative ai due gruppi e a partire da esse determinare le misure di non omogeneità statistiche

$$X = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \\ I_6 \\ I_7 \\ I_8 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 0 & 0 \\ 1 & 1 \\ 2 & 2 \\ 4 & 3 \\ 5 & 3 \\ 4 & 4 \\ 6 & 5 \\ 7 & 5 \end{pmatrix} \end{array}$$

```
> X1<-data.frame(c1=c(0,1,2),c2=c(0,1,2))
> rownames(X1)<-c("I1","I2","I3")
>  # n1>1
> n1<-nrow(X1)
> trH1<-(n1-1)*sum(apply(X1,2,var))
> trH1 # misura di non omogeneita' statistica del primo gruppo
[1] 4
```

Gruppo G_1

  - La traccia della matrice di non omogeneità statistica del primo gruppo G_1 è $trH_1 = 4$

```
> X2<-data.frame(c1=c(4,5,4,6,7),c2=c(3,3,4,5,5))
> rownames(X2)<-c("I4","I5","I6","I7","I8")
> # n2>1
> n2<-nrow(X2)
> trH2<-(n2-1)*sum(apply(X2,2,var))
> trH2 # misura di non omogeneita' statistica del secondo gruppo
[1] 10.8
```

Gruppo G_2

  - La traccia della matrice di non omogeneità statistica del primo gruppo G_2 è $trH_2 = 10.8$

Misure di non omogeneità statistiche

- **Metodo 2**: non occorre definire le matrici dei dati relative ai due gruppi e si possono ricavare le misure di non omogeneità statistica dei due gruppi utilizzando le funzioni **cutree**() e **aggregate**()

$$X = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \\ I_6 \\ I_7 \\ I_8 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 0 & 0 \\ 1 & 1 \\ 2 & 2 \\ 4 & 3 \\ 5 & 3 \\ 4 & 4 \\ 6 & 5 \\ 7 & 5 \end{pmatrix} \end{array}$$

```
> d<-dist(X,method="euclidean",diag=TRUE,upper=TRUE)
> d2<-d^2
> tree <- hclust(d2, method = "centroid")
>
> taglio<-cutree(tree, k =2, h = NULL)
> num<-table(taglio) #numero di elementi dei gruppi
> tagliolist<-list(taglio) #lista di indici per i gruppi
> agvar<- aggregate(X, tagliolist, var)[, -1]
> # n1>1
> trH1<-(num[[1]]-1) * sum(agvar[1, ])
> trH1 # visualizza la misura di non omogeneita' del primo gruppo
[1] 4
> # n2>1
> trH2<-(num[[2]] -1) *sum(agvar[2, ])
> trH2 # visualizza la misura di non omogeneita' del secondo gruppo
[1] 10.8
```

Matrice dei quadrati distanze euclidee

- La misura di non omogeneità del primo gruppo è calcolata moltiplicando $n_1 - 1$ per la somma degli elementi della prima riga della matrice ottenuta con **aggregate**(), ossia $2\,(1+1) = 4$

Misure di non omogeneità statistiche

- **Metodo 2**: non occorre definire le matrici dei dati relative ai due gruppi e si possono ricavare le misure di non omogeneità statistica dei due gruppi utilizzando le funzioni **cutree**() e **aggregate**()

$$X = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \\ I_6 \\ I_7 \\ I_8 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 0 & 0 \\ 1 & 1 \\ 2 & 2 \\ 4 & 3 \\ 5 & 3 \\ 4 & 4 \\ 6 & 5 \\ 7 & 5 \end{pmatrix} \end{array}$$

```
> d<-dist(X,method="euclidean",diag=TRUE,upper=TRUE)
> d2<-d^2
> tree <- hclust(d2, method = "centroid")
>
> taglio<-cutree(tree, k =2, h = NULL)
> num<-table(taglio) #numero di elementi dei gruppi
> tagliolist<-list(taglio) #lista di indici per i gruppi
> agvar<- aggregate(X, tagliolist, var)[, -1]
> # n1>1
> trH1<-(num[[1]]-1) * sum(agvar[1, ])
> trH1 # visualizza la misura di non omogeneita' del primo gruppo
[1] 4
> # n2>1
> trH2<-(num[[2]] -1) *sum(agvar[2, ])
> trH2 # visualizza la misura di non omogeneita' del secondo gruppo
[1] 10.8
```

Matrice dei quadrati distanze euclidee

- La misura di non omogeneità del secondo gruppo è calcolata moltiplicando $n_2 - 1$ per la somma degli elementi della prima riga della matrice ottenuta con **aggregate**(), ossia $4\ (1.7 + 1) = 10.8$

Misure di non omogeneità statistiche

- La misura di non omogeneità statistica totale è quindi $trH_I = 64.75$
  - La misura di non omogeneità statistica all'interno dei due gruppi (**within**):

$$trH_1 + trH_2 = 4 + 10.8 = 14.8$$

  - La misura di non omogeneità statistica tra i due gruppi (**between**):

$$trH(G_1 \cap G_2) = trH_I - trH_1 - trH_2 =$$
$$= 64.75 - 4 - 10.8 = 49.95$$

$$X = \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \\ I_6 \\ I_7 \\ I_8 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{pmatrix} 0 & 0 \\ 1 & 1 \\ 2 & 2 \\ 4 & 3 \\ 5 & 3 \\ 4 & 4 \\ 6 & 5 \\ 7 & 5 \end{pmatrix} \end{array}$$

- Si ha quindi che la misura di non omogeneità all'interno dei gruppi (within) è quindi piccola rispetto la misura di non omogeneità tra i cluster (between)