# CAPITOLO 8
# Paragrafo 8.5 Analisi multivariata degli indici di bilancio: analisi dei gruppi (*cluster* analysis)



PEARSON

# Analisi dei gruppi o *cluster analysis*

**Obiettivo principale**

Considerare un gran numero *n* di unità statistiche e creare un certo numero *g* di gruppi distinti che contengono unità simili, in base alle variabili considerate

**In questa presentazione** consideriamo solo variabili misurabili almeno su scala a intervallo anche se l'analisi cluster può agire pure su variabili ordinali e nominali.

# Analisi cluster dei ratio di bilancio

A fini comparativi sarà utile proporre dei confronti con unità comparabili e quindi sarà opportuno identificare **gruppi omogenei** di imprese (*righe della matrice dei dati*)

→ riduzione delle unità da *n* unità a *g* gruppi (*g*<*n*)

Matrice dei dati (indici di bilancio)

| Impresa | Indice 1 | Indice 2 | … | Indice *j* | … | Indice *p* |
|---|---|---|---|---|---|---|
| $U_1$ | $x_{11}$ | $x_{12}$ | … | $x_{1j}$ | … | $x_{1p}$ |
| $U_2$ | $x_{21}$ | $x_{22}$ | … | $x_{2j}$ | … | $x_{2p}$ |
| … | … | … | … | … | … | … |
| $U_i$ | $x_{i1}$ | $x_{i2}$ | … | $x_{ij}$ | … | $x_{ip}$ |
| … | … | … | … | … | … | … |
| $U_n$ | $x_{n1}$ | $x_{n2}$ | … | $x_{nj}$ | … | $x_{np}$ |



PEARSON

# La riduzione delle unità: analisi dei gruppi (analisi *cluster*)

**Obiettivo**

Identificare gruppi omogenei di imprese rispetto ai valori assunti da $p$ indici di bilancio per individuare gruppi con prestazioni e comportamenti gestionali simili e proporre confronti maggiormente significativi.

**Come sono fatti i gruppi individuati**

Si tratta di gruppi **esclusivi** (ogni unità appartiene a un solo gruppo) ed **esaustivi** (tutte le unità sono collocate nei gruppi)**.**

L'analisi è **esplorativa**: sono i dati e i metodi impiegati a formare i gruppi che vengono interpretati (nelle loro caratteristiche statistiche) a posteriori.

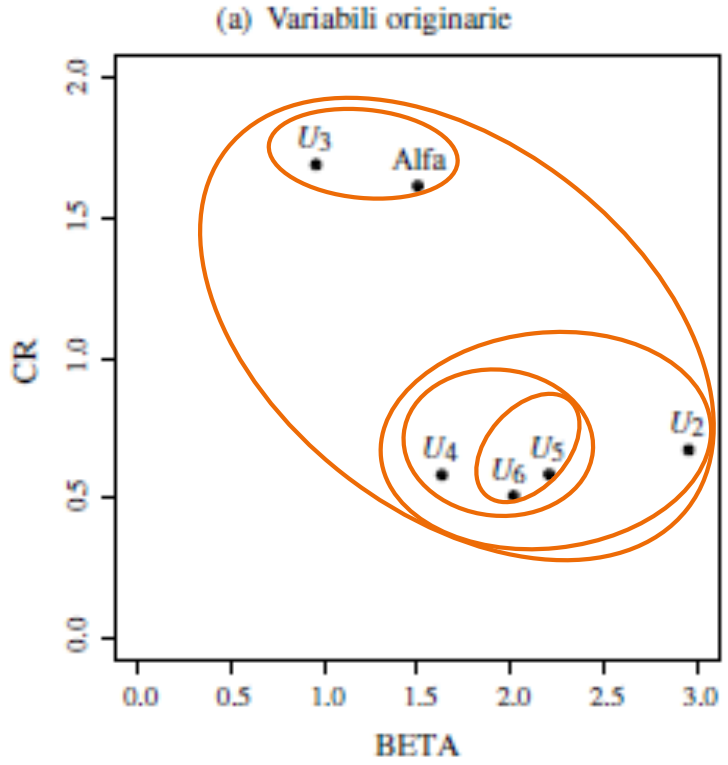
# Metodi esplorativi di identificazione dei gruppi

**Metodi gerarchici e non gerarchici**.
A seconda che i gruppi provengono dalla progressiva aggregazione o divisione di altri gruppi o no. I metodi gerarchici richiedono l'indicazione del numero *g* di gruppi cercati.

**Metodi gerarchici**

- **agglomerativi**: da *n* gruppi (quante sono le unità) si arriva progressivamente a raggruppare le unità *meno diverse* in sottogruppi, e a raggruppare i sottogruppi formati in gruppi, fino ad avere un solo gruppo che contiene tutte le unità
- **divisivi**: da un solo gruppo, che contiene tutte le unità, si procede alla progressiva suddivisione delle unità *più diverse* fino ad ottenere *n* gruppi ciascuno con una sola unità

# Gerarchico agglomerativo con *n*=6 e *p*=2




| Unità/gruppi formati | N. gruppi |
|---|---|
| $U_1$, $U_2$, $U_3$, $U_4$, $U_5$, $U_6$ | 6 |
| $G_1$:{ $U_5$, $U_6$}, $U_1$, $U_2$, $U_3$, $U_4$ | 5 |
| $G_1$:{ $U_5$, $U_6$}, $G_2$:{ $U_1$, $U_3$}, $U_2$, $U_4$ | 4 |
| $G_3$:{ $G_1$ $U_4$},$G_2$, $U_2$ | 3 |
| $G_2$, $G_4$:{$G_3$, $U_2$} | 2 |
| $G_5$:{$G_2$, $G_4$} | 1 |

Quando un'unità è entrata in un gruppo **non viene** da questo più rimossa

Il metodo gerarchico **non fornisce** il numero dei gruppi che deve essere scelto a posteriori esaminando i risultati



PEARSON

# Il punto di partenza dei metodi gerarchici agglomerativi: le distanze fra coppie di unità

- Per raggruppare le unità *meno diverse* è necessario stabilire una misura di **diversità** o **dissimilarità** sulla base delle *p* variabili
- Se le *p* variabili sono **quantitative** si parla di **distanze** o di **indici di distanza** fra unità

Gli indici di bilancio sono grandezze quantitative e indichiamo con:

$$d_{ij} = d\left(U_i, U_j\right) = d(\underline{x}_i, \underline{x}_j)$$

la **distanza** (o l'**indice di distanza**) fra le unità (imprese) $U_i$ e $U_j$ , dove $\underline{x}_i$ e $\underline{x}_j$ sono le righe *i* e *j* della matrice dei dati (i ratio).



PEARSON

# Distanze e indici di distanza: proprietà

$$d_{ij} = d(U_i, U_j) = d(\underline{x}_i, \underline{x}_j)$$

1. Identità: $d_{ij} = 0$ se e solo se $\underline{x}_i = \underline{x}_j$
2. Non negatività: $d_{ij} \geq 0$
3. Simmetria: $d_{ij} = d_{ji}$
4. Triangolarità: $d_{ij} \leq d_{ik} + d_{kj}$

Indice di distanza

Distanza

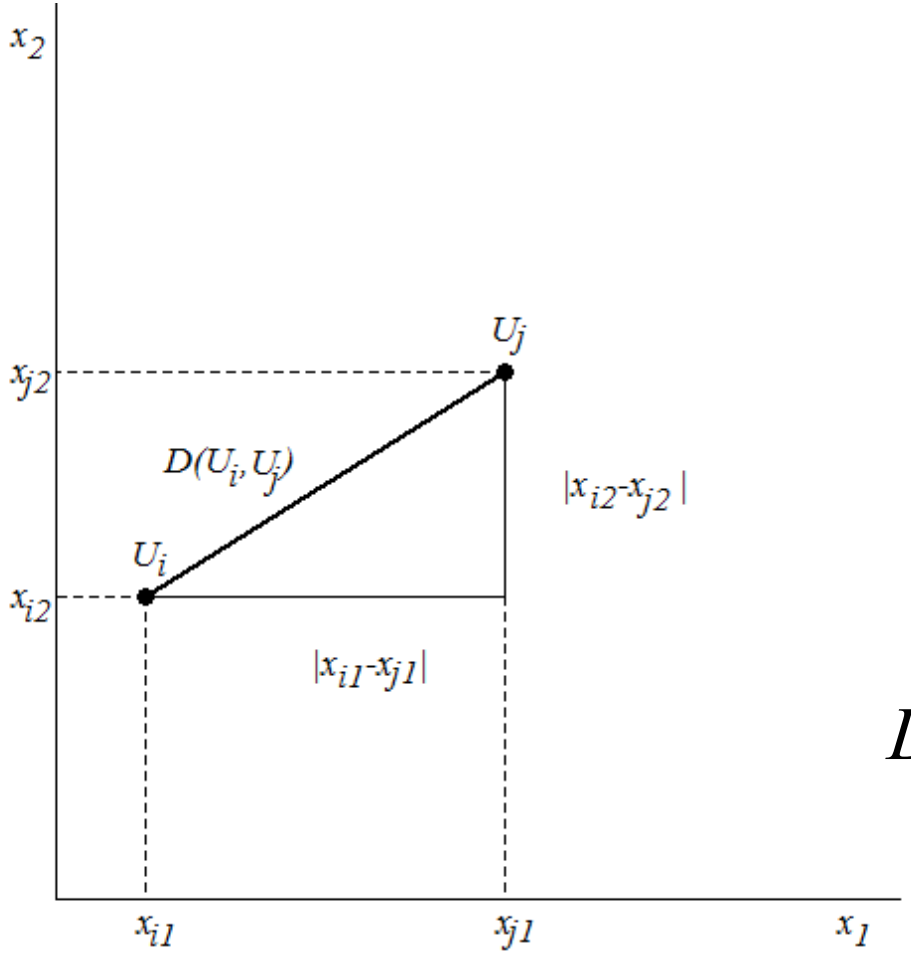
| | $U_1$ | $U_2$ | … | $U_j$ | … | $U_n$ |
|---|---|---|---|---|---|---|
| $U_1$ | 0 | $d_{12}$ | … | $d_{1j}$ | … | $d_{1n}$ |
| $U_2$ | | 0 | … | $d_{2j}$ | … | |
| … | … | … | … | … | … | … |
| $U_i$ | | | … | $d_{ij}$ | … | $d_{in}$ |
| … | | | | | … | … |
| $U_n$ | | | | | | 0 |

La matrice delle distanze (o degli indici di distanza) fra coppie di unità è **simmetrica**



PEARSON

# Distanza euclidea e distanza Manhattan (city block): rappresentazione grafica con *p*=2




Euclidea

$$D(U_i, U_j) = D(\underline{x}_i, \underline{x}_j) = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$$

Manhattan

$$D_M(U_i, U_j) = D_M(\underline{x}_i, \underline{x}_j) = \sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|$$



PEARSON

# Quadrato della distanza Euclidea

$$D^2\left(U_i, U_j\right) = D^2(\underline{x}_i, \underline{x}_j) = \sum_{k=1}^{p} (x_{ik} - x_{jk})^2$$

- a differenza della distanza Euclidea e Manhattan questo è un **indice di distanza** e non una distanza
- si presa a un più semplice trattamento algebrico rispetto alla distanza Euclidea (non ha la radice quadrata) e rispetto alla distanza Manhattan (non ha l'operatore di valore assoluto)
- esalta le distanze fra le unità più lontane



PEARSON

# Distanza euclidea, distanza Manhattan e quadrato della distanza Euclidea. Esempio.

$\underline{x}_i$:(5,4) $\underline{x}_j$:(1,3)

$$D(\underline{x}_i,\underline{x}_j)=\sqrt{\sum_{k=1}^{p}(x_{ik}-x_{jk})^2}=\sqrt{(5-1)^2+(4-3)^2}=\sqrt{4^2+1^2}=\sqrt{17}\approx 4.12$$

$$D_M(\underline{x}_i,\underline{x}_j)=\sum_{k=1}^{p}\left|x_{ik}-x_{jk}\right|=\left|5-1\right|+\left|4-3\right|=4+1=5$$

$$D^2(U_i,U_j)=D^2(\underline{x}_i,\underline{x}_j)=\sum_{k=1}^{p}(x_{ik}-x_{jk})^2=(5-1)^2+(4-3)^2=17$$



PEARSON

# Le fasi di un'analisi cluster gerarchica

**1.** scelta delle unità (le imprese);

**2.** scelta delle variabili (indici di bilancio) sulle quali basare il raggruppamento, considerando anche eventuali trasformazioni dei dati;

**3.** scelta dell'indice di distanza;

**4.** scelta del criterio (o algoritmo) gerarchico agglomerativo e applicazione;

**5.** interpretazione dei gruppi individuati



PEARSON

# 1. Scelta delle unità (le imprese)

La scelta delle unità è legata certamente all'obiettivo dello studio ma….
la conduzione di un'analisi cluster su dati di bilancio richiede comunque il rispetto dei requisiti di comparabilità che garantiscono la correttezza di una analisi statistica.


PEARSON

# 2. Scelta delle variabili (1/4)

La selezione delle variabili (nel nostro caso gli indici di bilancio) dovrà scaturire prima di tutto dagli obiettivi che ci si propone.
Ad esempio, potremo selezionare i ratio che esprimono le dimensioni di situazione e struttura finanziaria e di redditività.

Prima di calcolare le distanze occorre tenere presente che:

a) Le *p* variabili dovrebbero essere espresse nella **stessa** unità di misura.

b) Le variabili con **varianza più elevata** assumono **maggiore importanza** nell'indice di distanza e quindi nell'analisi cluster.

c) La correlazione fra variabili produce **ridondanza** che determina un maggior peso di alcune variabili rispetto ad altre.

# 2. Scelta delle variabili (2/4). Variabili con diversa unità di misura.

Su due individui abbiamo misurato peso (in Kg.) e altezza (in cm.)

$$\underline{x}_i:(50,160) \qquad x_j:(75,178)$$

$$D^2(\underline{x}_i,\underline{x}_j) = \underbrace{(50-75)^2}_{Kg^2} + \underbrace{(160-180)^2}_{cm^2} = \underbrace{1025}_{?}$$

**Possibile soluzione**: standardizzare i valori.

**ATTENZIONE**: la standardizzazione dei valori può produrre delle inversioni nelle distanze fra unità (v. Esempio 8.4 nel testo)

**Il problema della diversa unità di misura non si presenta coi ratio di bilancio perché sono numeri puri.**

# 2. Scelta delle variabili (3/4). Variabili con stessa unità di misura ma diversa scala.

Su due individui abbiamo misurato lunghezza del naso (in cm.) e altezza (in cm.)

$\underline{x}_i$:(3.7,160) $\qquad$ $x_j$:(4.1,180)

$$D^2(\underline{x}_i, \underline{x}_j) = (3.7 - 4.1)^2 + (160 - 180)^2 = 0.16 + 400$$

La distanza è determinata dall'altezza

**Possibile soluzione**: standardizzare i valori.

**ATTENZIONE**: la standardizzazione dei valori può produrre delle inversioni nelle distanze fra unità (v. Esempio 8.4 nel testo)



PEARSON

# 2. Scelta delle variabili (4/4). Variabili correlate

Se sono presenti variabili molto correlate (per esempio, *quick ratio* e *current ratio*), gli aspetti descritti da tali variabili (nella fattispecie: la stuazione finanziaria a breve termine) saranno *sovrarappresentati* nel calcolo della distanza e avranno quindi maggior importanza nell'analisi cluster.
La correlazione ha anche conseguenze sulla **forma dei gruppi** (v. oltre) che diventano difficili da individuare.

**Possibile soluzione**: ricorso all'ACP su CP standardizzate e impiego delle CP standardizzate


PEARSON

# 3. Scelta dell'indice di distanza

Di particolare interesse sono: la distanza **euclidea** il suo **quadrato**.

La **distanza euclidea** sui valori **standardizzati** delle componenti principali equivale ad applicare, sulle variabili originarie, una speciale distanza **generalizzata** che tiene conto della diversa varianza di queste e della loro correlazione.

Il **quadrato della distanza euclidea** può essere interpretata come **indice di variabilità** (ovvero: devianza; v. il metodo di Ward)

**ATTENZIONE**: l'indice di distanza interagisce con l'algoritmo di raggruppamento e quindi con la struttura di gruppi individuata.

# 4. Scelta dell'algoritmo: la logica del raggruppamento gerarchico (1/8)

Vediamo il funzionamento della cluster gerarchica a partire da una matrice delle distanza fra 6 unità (quadrato della distanza euclidea).

| | $U_1$ | $U_2$ | $U_3$ | $U_4$ | $U_5$ | $U_6$ |
|---|---|---|---|---|---|---|
| $U_1$ | 0 | 3,00 | 0,30 | 1,08 | 1,56 | 1,49 |
| $U_2$ | 3,00 | 0 | 5,02 | 1,76 | 0,57 | 0,91 |
| $U_3$ | 0,30 | 5,02 | 0 | 1,68 | 2,78 | 2,53 |
| $U_4$ | 1,08 | 1,76 | 1,68 | 0 | 0,33 | 0,15 |
| $U_5$ | 1,56 | 0,57 | 2,78 | 0,33 | 0 | **0,04** |
| $U_6$ | 1,49 | 0,91 | 2,53 | 0,15 | **0,04** | **0** |

**PASSO 1.**

Si uniscono le unità meno distanti: $U_5$ e $U_6$ formano il gruppo $G_1$



PEARSON

# 4. … la logica del raggruppamento gerarchico (2/8)

**PASSO 2.** Si aggiorna la matrice delle distanze che ora è 5×5 e si uniscono le unità/gruppi meno distanti.

Le unità $U_1$ e $U_3$ vanno a formare il gruppo $G_2$

| | $G_1$ | $U_1$ | $U_2$ | $U_3$ | $U_3$ |
|---|---|---|---|---|---|
| $G_1$ | 0 | | | | |
| $U_1$ | 1.56 | 0 | | | |
| $U_2$ | 0.91 | 3.00 | 0 | | |
| $U_3$ | 2.78 | **0.30** | 5.02 | 0 | |
| $U_4$ | 0.33 | 1.08 | 1.76 | 1.68 | 0 |

**Attenzione**: queste sono distanze fra unità e **gruppo**.



PEARSON

# 4. … la logica del raggruppamento gerarchico (3/8)

**PASSO 3.** Si aggiorna la matrice delle distanze che ora è 4×4 e si uniscono le unità/gruppi meno distanti.

L'unità $U_4$ si unisce al gruppo $G_1$ e forma il gruppo $G_3$

| | $G_1$ | $G_2$ | $U_2$ | $U_4$ |
|---|---|---|---|---|
| $G_1$ | 0 | | | |
| $G_2$ | 2,78 | 0 | | |
| $U_2$ | 0,91 | 5,02 | 0 | |
| $U_4$ | **0,33** | 1,68 | 1,76 | 0 |

**Attenzione**: in questa matrice ci sono distanze **fra gruppi** e distanze fra unità e gruppi.



PEARSON

# 4. … la logica del raggruppamento gerarchico (4/8)

**PASSO 4.** Si aggiorna la matrice delle distanze che ora è 3×3 e si uniscono le unità/gruppi meno distanti.

L'unità $U_2$ si unisce al gruppo $G_3$ e si forma il gruppo $G_4$

| | $U_2$ | $G_3$ | $G_2$ |
|---|---|---|---|
| $U_2$ | 0 | | |
| $G_3$ | **1,76** | 0 | |
| $G_2$ | 5,02 | 2,78 | 0 |



PEARSON

# 4. … la logica del raggruppamento gerarchico (5/8)

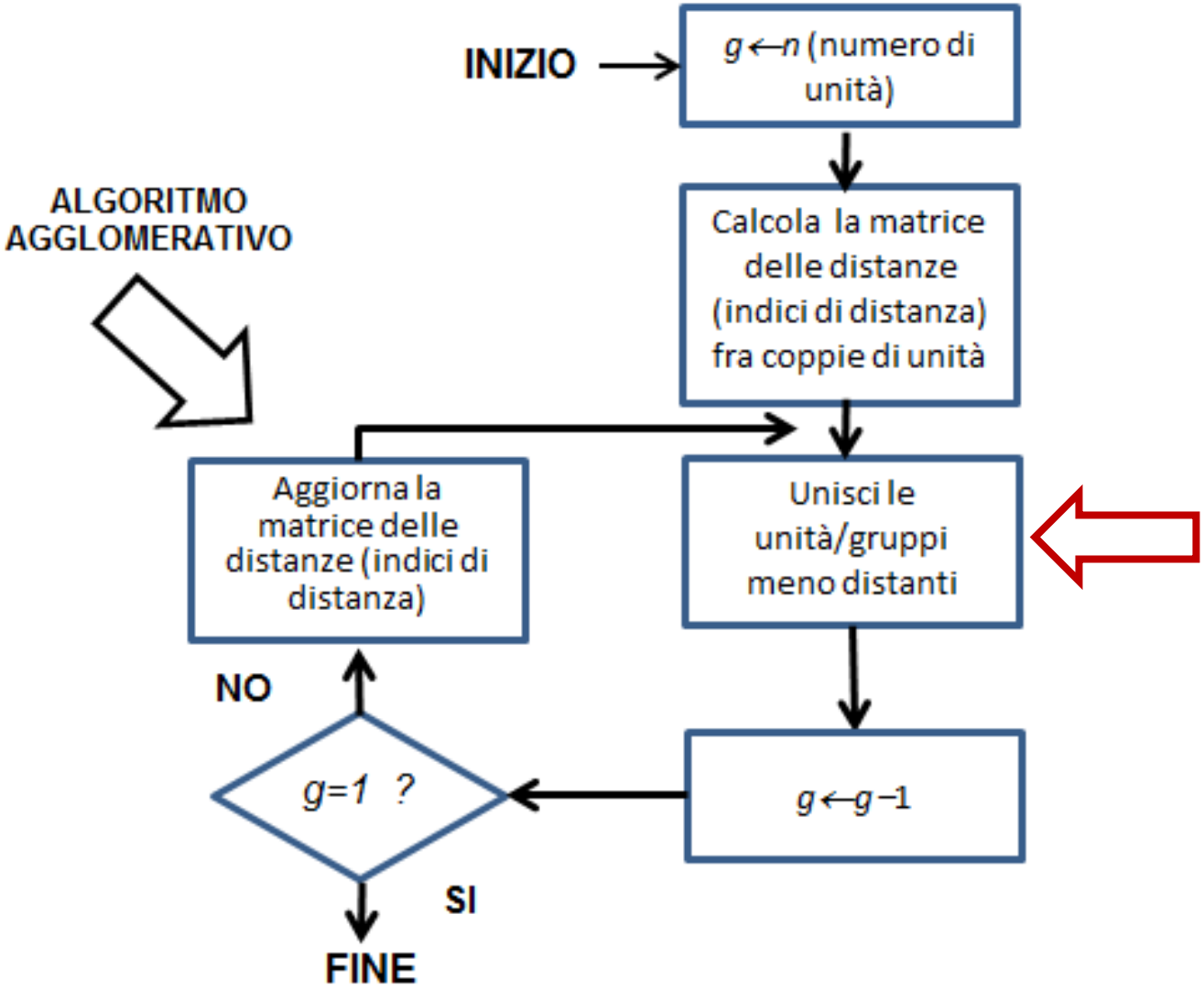
**PASSO 5.** Si aggiorna la matrice delle distanze che ora è 2×2 e si uniscono i due gruppi rimasti ottenendo un gruppo che contiene tutte le unità.

| | $G_4$ | $G_2$ |
|---|---|---|
| $G_4$ | 0 | |
| $G_2$ | **5,02** | 0 |



PEARSON

# 4. … la logica del raggruppamento gerarchico (6/8)



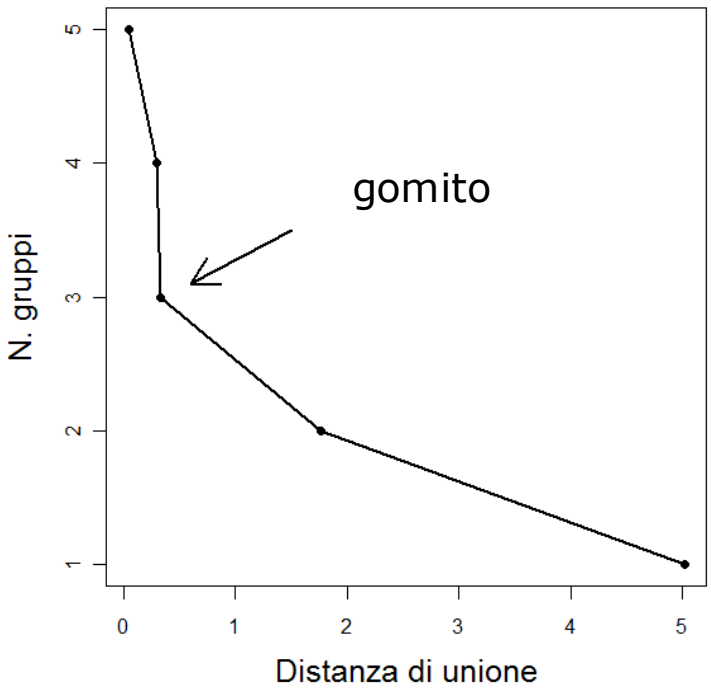

*In questa fase si va a cercare la distanza **minima** nella matrice delle distanze*



PEARSON

# 4. … la logica del raggruppamento gerarchico. Riepilogo dell'aggregazione: lo screeplot (7/8)

| Unità/gruppi formati | N. gruppi | Distanza di unione |
|---|---|---|
| $U_1, U_2, U_3, U_4, U_5, U_6$ | 6 | 0,00 |
| $G_1$:{ $U_5, U_6$}, $U_1, U_2, U_3, U_4$ | 5 | 0,04 |
| $G_1$:{ $U_5, U_6$}, $G_2$:{ $U_1, U_3$}, $U_2, U_4$ | 4 | 0,30 |
| $G_3$:{ $G_1$ $U_4$},$G_2$, $U_2$ | 3 | 0,33 |
| $G_2$, $G_4$:{$G_3$, $U_2$} | 2 | 1,76 |
| $G_5$:{$G_2$, $G_4$} | 1 | 5,02 |




Lo scree plot Suggerisce il numero dei gruppi in corrispondenza del **gomito.** In questo caso: **3** gruppi.

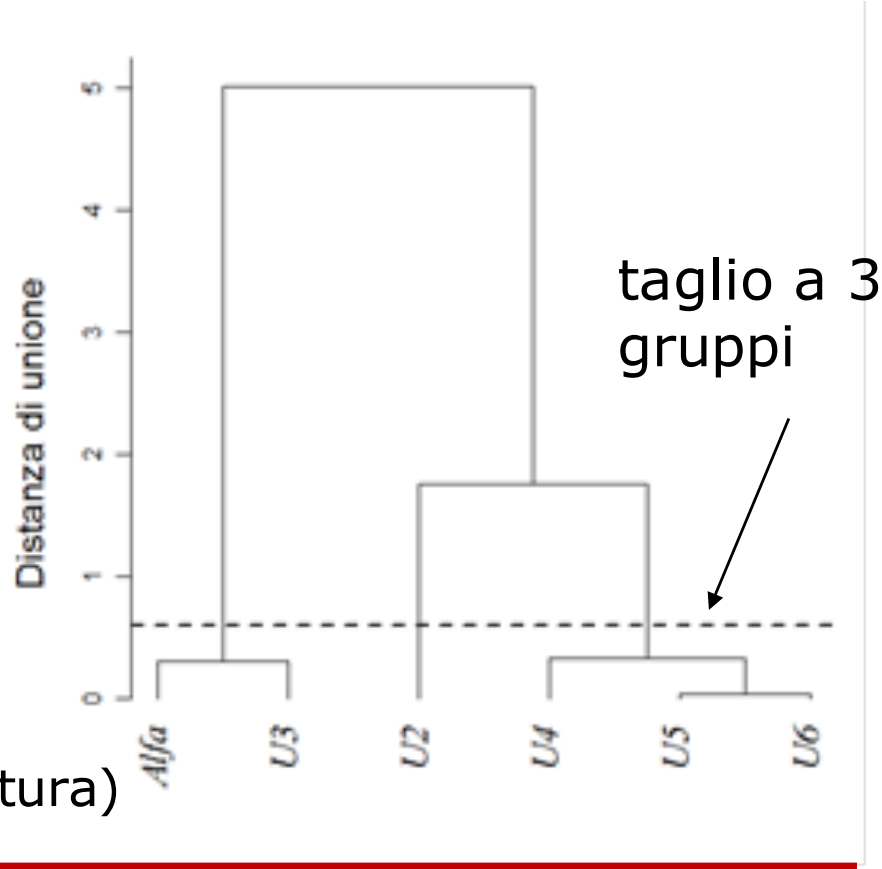
**N.B.** Più aggrego i gruppi e maggiore è la distanza di unione (cioè vado ad unire gruppi sempre più eterogenei fra loro e quindi formo gruppi sempre più eterogeni al loro interno)



PEARSON

# 4. … la logica del raggruppamento gerarchico. Riepilogo dell'aggregazione: il dendrogramma o diagramma ad albero (8/8)

| Unità/gruppi formati | N. gruppi | Distanza di unione |
|---|---|---|
| $U_1, U_2, U_3, U_4, U_5, U_6$ | 6 | 0,00 |
| $G_1$:{ $U_5, U_6$}, $U_1, U_2, U_3, U_4$ | 5 | 0,04 |
| $G_1$:{ $U_5, U_6$}, $G_2$:{ $U_1, U_3$}, $U_2, U_4$ | 4 | 0,30 |
| $G_3$:{ $G_1$ $U_4$},$G_2, U_2$ | 3 | 0,33 |
| $G_2, G_4$:{$G_3, U_2$} | 2 | 1,76 |
| $G_5$:{$G_2, G_4$} | 1 | 5,02 |




La cluster gerarchica non ci fornisce direttamente la composizione dei gruppi o il numero dei gruppi.
Si deve decidere dove fare il ***pruning*** (potatura)

**N.B.** Più taglio in alto (minore numero di gruppi) e più eterogenei al loro interno sono i gruppi che vado a formare


PEARSON

# 4. Scelta dell'algoritmo di raggruppamento (1/4)

L'algoritmo di raggruppamento stabilisce:

- come calcolare la distanza fra **singola unità** e **gruppo**
- come calcolare la distanza fra **due gruppi**

Esistono numerosi algoritmi gerarchici agglomerativi. Noi vediamo:

- **legame singolo**
- **legame completo**
- **distanza media**
- **metodo di Ward**



PEARSON

# 4. Scelta dell'algoritmo di raggruppamento (2/4)

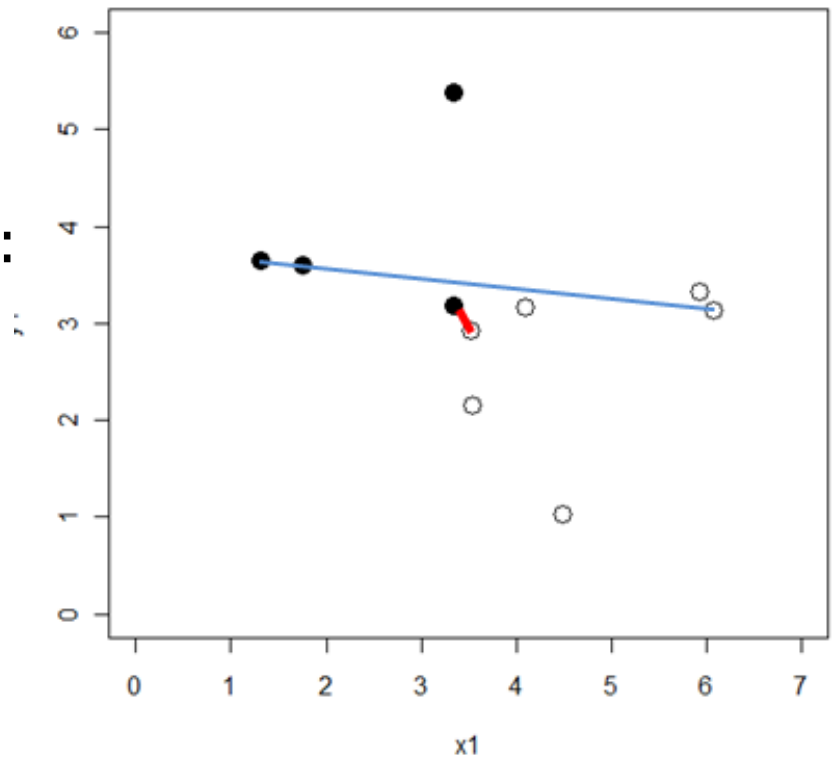
Abbiamo 2 gruppi ($G_1$: punti neri, $G_2$: punti bianchi).

**Legame singolo:** la distanza fra i 2 gruppi è la minima distanza fra due unità: una di un gruppo e l'altra dell'altro gruppo (linea rossa)

$$d(G_1, G_2) = \min_{i \in G_1, j \in G_2} d(U_i, U_j)$$

**Legame completo:** la distanza fra i 2 gruppi è la massima distanza fra due unità: una di un gruppo e l'altra dell'altro gruppo (linea azzurra)

$$d(G_1, G_2) = \max_{i \in G_1, j \in G_2} d(U_i, U_j)$$






PEARSON

# Legame singolo: esempio

Matrice delle distanza fra unità

| | $U_1$ | $U_2$ | $U_3$ | $U_4$ | $U_5$ | $U_6$ |
|---|---|---|---|---|---|---|
| $U_1$ | 0 | 3.00 | 0.30 | 1.08 | 1.56 | 1.49 |
| $U_2$ | 3.00 | 0 | 5.02 | 1.76 | 0.57 | 0.91 |
| $U_3$ | 0.30 | 5.02 | 0 | 1.68 | 2.78 | 2.53 |
| $U_4$ | 1.08 | 1.76 | 1.68 | 0 | 0.33 | 0.15 |
| $U_5$ | 1.56 | 0.57 | 2.78 | 0.33 | 0 | 0.04 |
| $U_6$ | 1.49 | 0.91 | 2.53 | 0.15 | 0.04 | 0 |

Calcoliamo la distanza fra: $G_1$:$\{U_5, U_6\}$ e $G_2$:$\{U_1, U_3\}$

$min(d(U_5, U_1), d(U_5, U_3), d(U_6, U_1), d(U_6, U_3))=$

$min(1.56, 2.78, 1.49, 2.53)=$**1.49**



PEARSON

# Legame completo: esempio

Matrice delle distanza fra unità

| | $U_1$ | $U_2$ | $U_3$ | $U_4$ | $U_5$ | $U_6$ |
|---|---|---|---|---|---|---|
| $U_1$ | 0 | 3.00 | 0.30 | 1.08 | 1.56 | 1.49 |
| $U_2$ | 3.00 | 0 | 5.02 | 1.76 | 0.57 | 0.91 |
| $U_3$ | 0.30 | 5.02 | 0 | 1.68 | 2.78 | 2.53 |
| $U_4$ | 1.08 | 1.76 | 1.68 | 0 | 0.33 | 0.15 |
| $U_5$ | 1.56 | 0.57 | 2.78 | 0.33 | 0 | 0.04 |
| $U_6$ | 1.49 | 0.91 | 2.53 | 0.15 | 0.04 | 0 |

Calcoliamo la distanza fra: $G_1$:{$U_5$, $U_6$} e $G_2$:{ $U_1$, $U_3$}

$max(d(U_5, U_1), d(U_5, U_3), d(U_6, U_1), d(U_6, U_3))=$

$max(1.56, 2.78, 1.49, 2.53)=$**2.78**



PEARSON

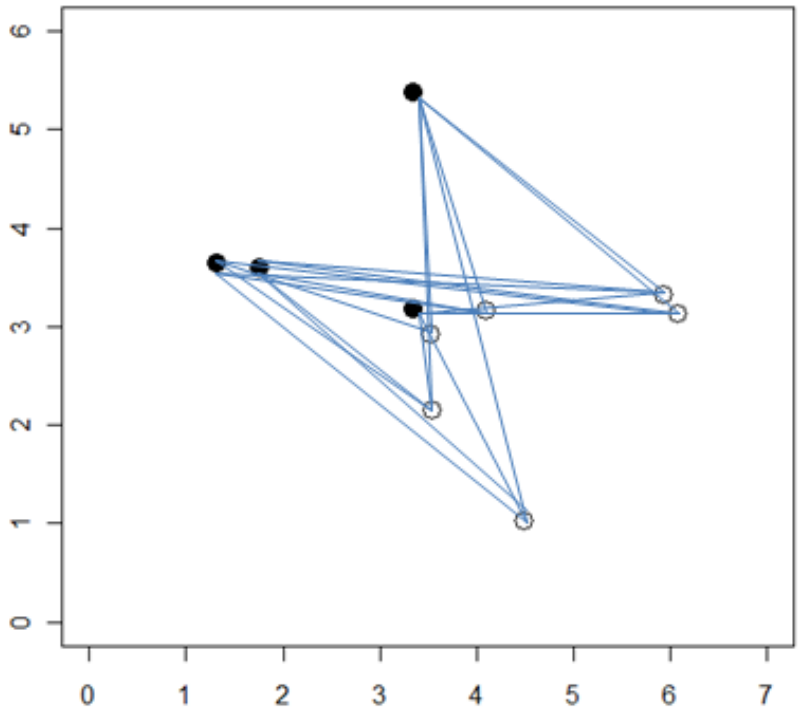
# 4. Scelta dell'algoritmo di raggruppamento (3/4)

**Distanza media**

la distanza fra i 2 gruppi è la media delle distanze fra ogni coppia di unità una che appartiene ad un gruppo e l'altra che appartiene all'altro gruppo.

$$d(G_1, G_2) = \frac{1}{n_1 n_2} \sum_{i \in G_1} \sum_{j \in G_2} d(U_i, U_j)$$

numerosità dei due gruppi






PEARSON

# Distanza media: esempio

Matrice delle distanza fra unità

| | $U_1$ | $U_2$ | $U_3$ | $U_4$ | $U_5$ | $U_6$ |
|---|---|---|---|---|---|---|
| $U_1$ | 0 | 3.00 | 0.30 | 1.08 | 1.56 | 1.49 |
| $U_2$ | 3.00 | 0 | 5.02 | 1.76 | 0.57 | 0.91 |
| $U_3$ | 0.30 | 5.02 | 0 | 1.68 | 2.78 | 2.53 |
| $U_4$ | 1.08 | 1.76 | 1.68 | 0 | 0.33 | 0.15 |
| $U_5$ | 1.56 | 0.57 | 2.78 | 0.33 | 0 | 0.04 |
| $U_6$ | 1.49 | 0.91 | 2.53 | 0.15 | 0.04 | 0 |

Calcoliamo la distanza fra: $G_1$:{$U_5$, $U_6$} e $G_2$:{ $U_1$, $U_3$}

$(d(U_5, U_1), d(U_5, U_3), d(U_6, U_1), d(U_6, U_3))/4=$

$(1.56 + 2.78 + 1.49 + 2.53)/4=$**2.09**



PEARSON

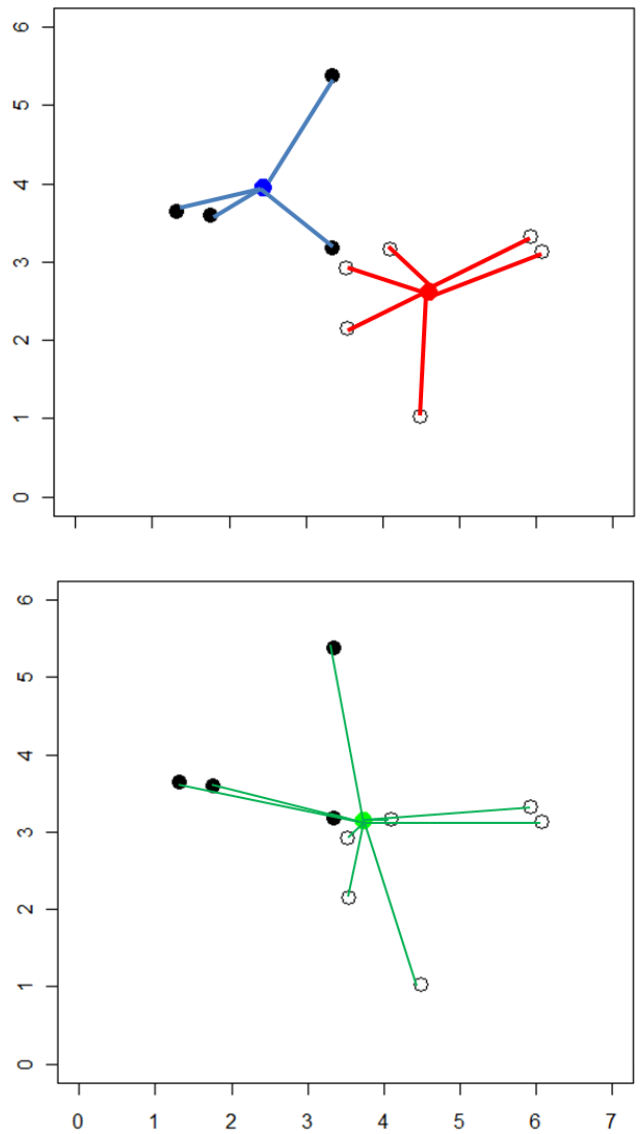
# 4. Scelta dell'algoritmo di raggruppamento ... (4/4)

**Metodo di Ward**

La distanza fra due gruppi è pari all' aumento di devianza interna dovuto all'unione dei 2 gruppi.

Tale aumento di devianza interna è:

$$\underbrace{\frac{n_1 n_2}{n_1 + n_2}}_{\text{numerosità dei gruppi}} \underbrace{\sum_{k=1}^{p} (\bar{x}_{G_1 k} - \bar{x}_{G_2 k})^2}_{\text{devianza fra le } p \text{ medie dei 2 gruppi}}$$






PEARSON

# Metodo di Ward: esempio

Calcoliamo la distanza fra*:* $G_1$:{$U_5$, $U_6$} e $G_2$*:*{ $U_1$, $U_3$} partendo dai valori originari delle variabili.

1) Calcolo dei centroidi dei due gruppi $G_1$ :(2.111; 0.547)
$G_2$ :(1.223; 1.655) vedi Tabella 8.7

2) $$D_{Ward}(G_1, G_2) = \frac{2+2}{2\times 2}\left((2.111-1.223)^2 + (0.547-1655)^2\right) = 2.00$$

Si parte, in genere, dalla matrice del **quadrato delle distanze euclidee** perché:

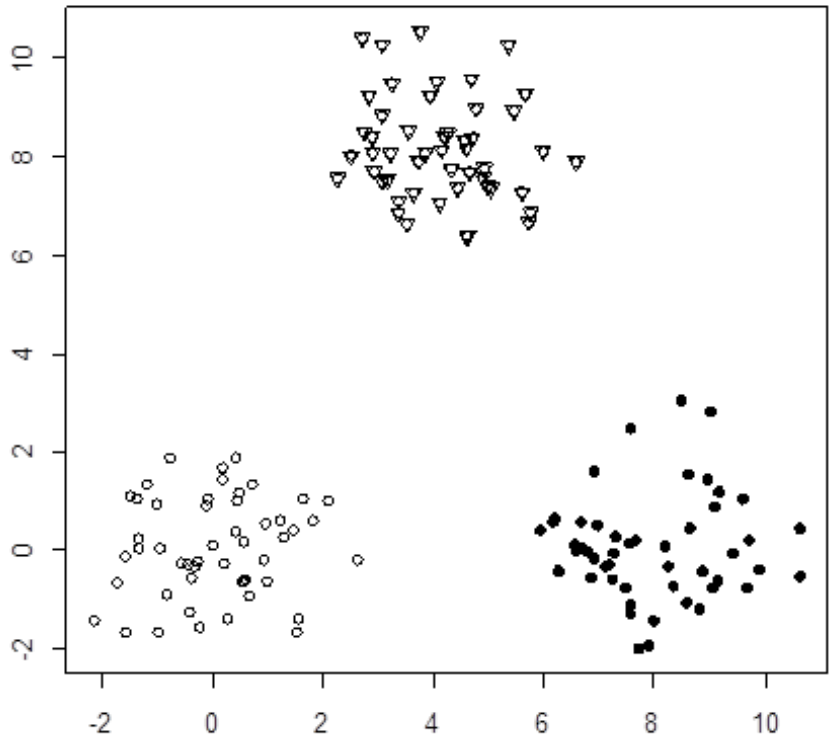
$$D_{Ward}(U_1, U_2) = \frac{1}{2} D^2(U_1, U_2)$$

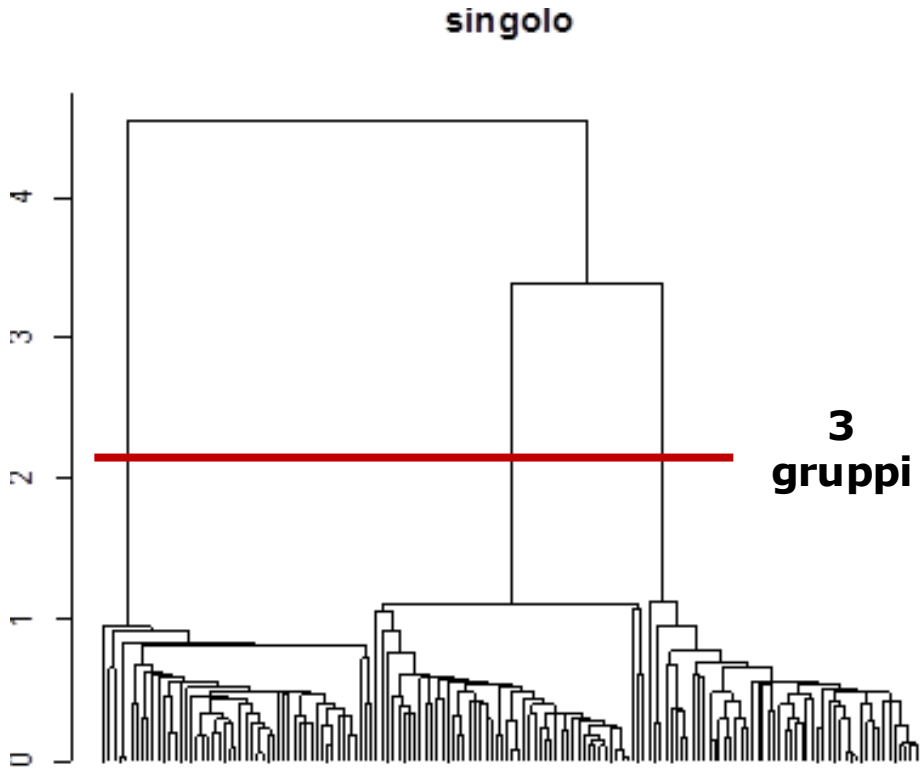


PEARSON

# Funzionamento dei metodi gerarchici: alcuni confronti (1/7) – 150 unità, 3 gruppi {50, 50, 50}

**Caso di gruppi ben distinti**: tutti i metodi riescono perfettamente ad individuarli, anche il legame singolo (esempio con distanza euclidea)



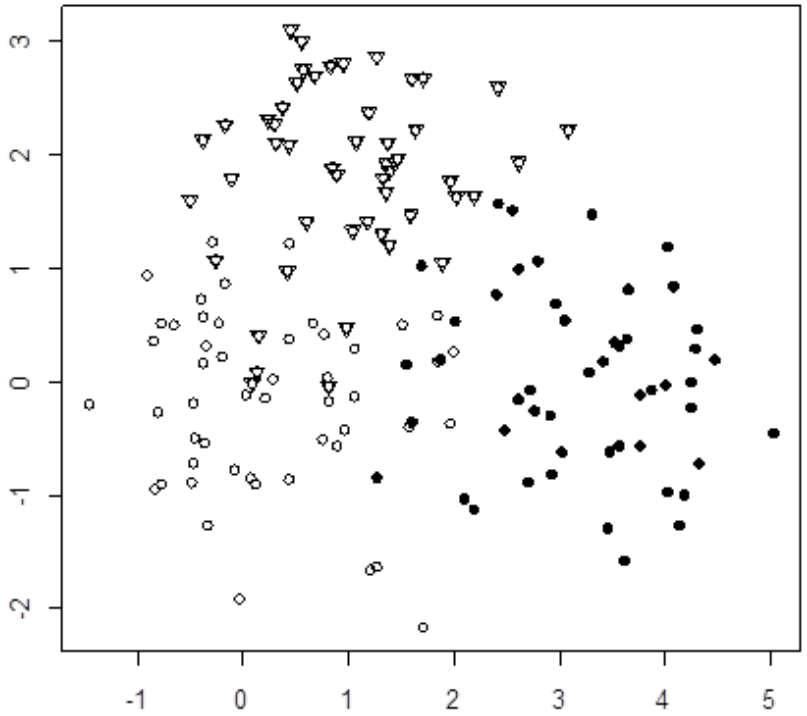



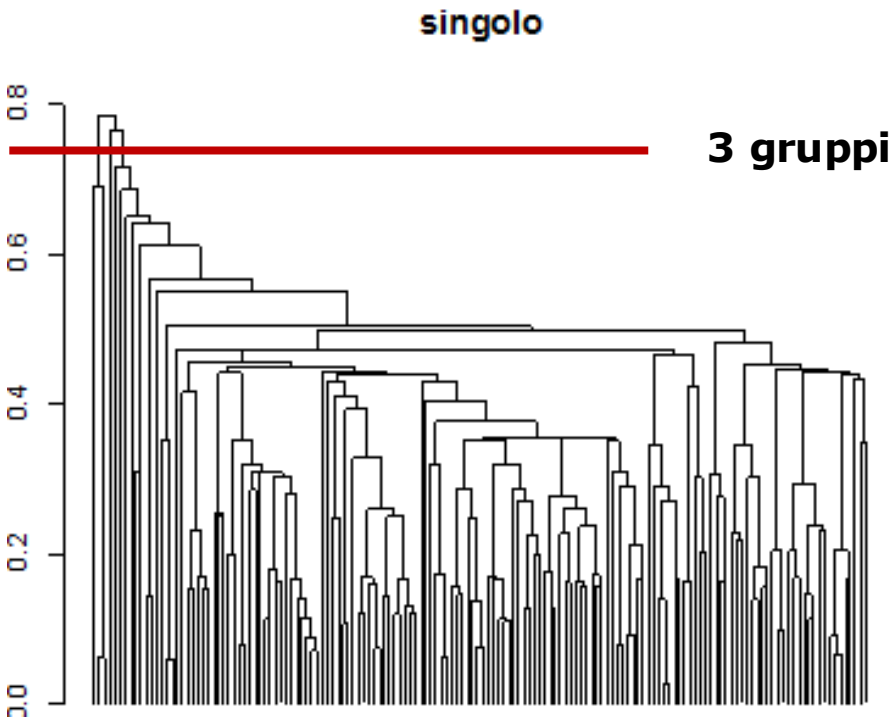



PEARSON

# Funzionamento dei metodi gerarchici: alcuni confronti (2/7) – 150 unità, 3 gruppi {50, 50, 50}

**Caso di gruppi ravvicinati ma di forma sferica**: il legame singolo fallisce completamente (esempio con distanza euclidea)



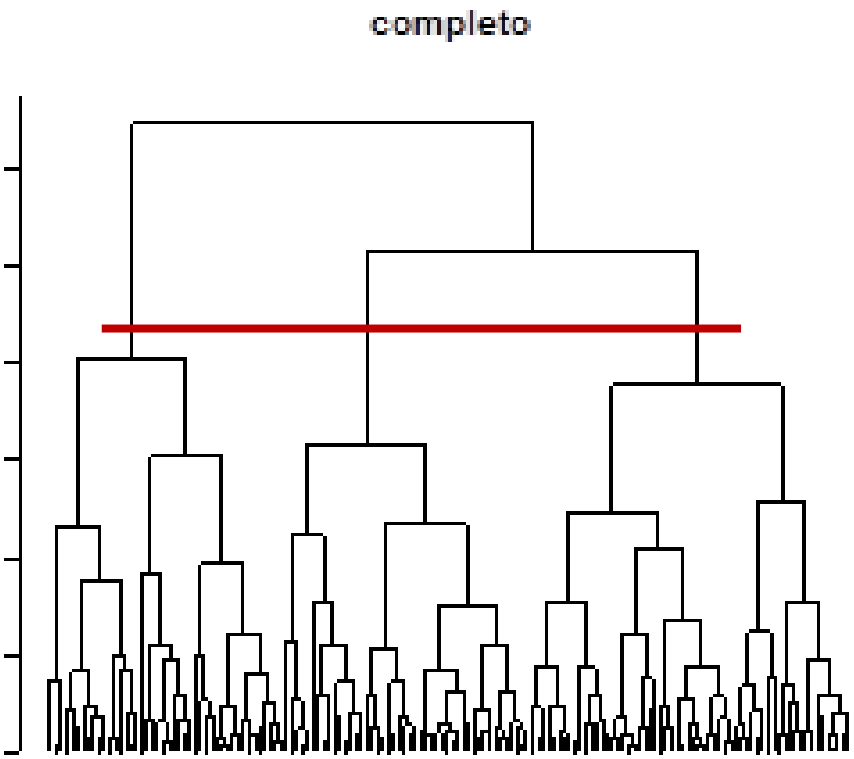



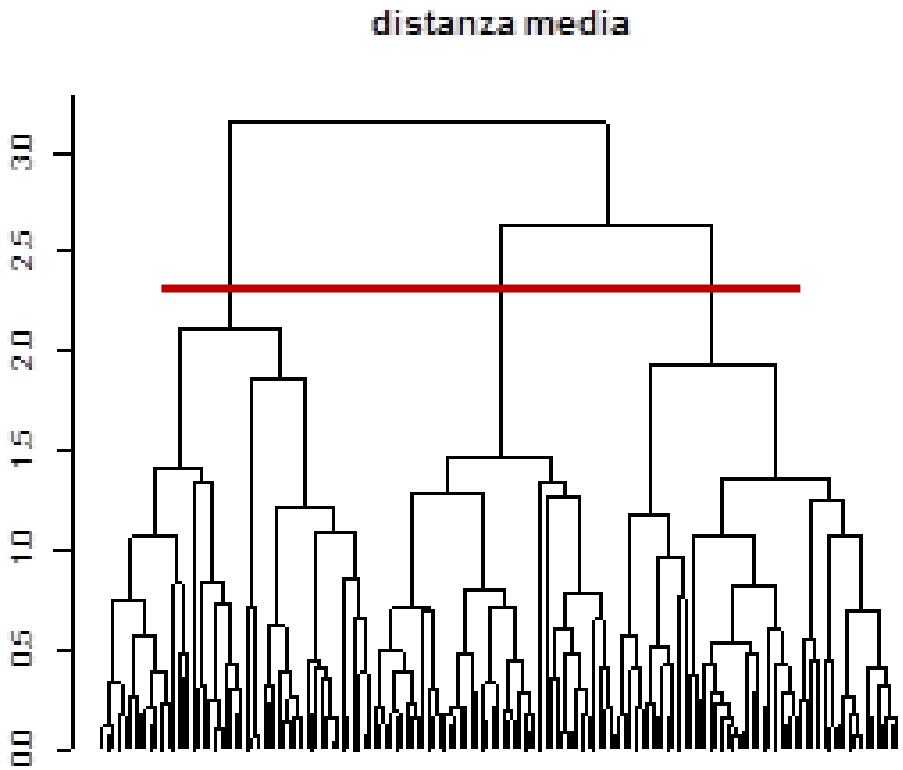



PEARSON

# Funzionamento dei metodi gerarchici: alcuni confronti (3/7) – 150 unità, 3 gruppi {50, 50, 50}

**Caso di gruppi ravvicinati ma di forma sferica**:
metodi del legame completo e della distanza media (con distanza euclidea)



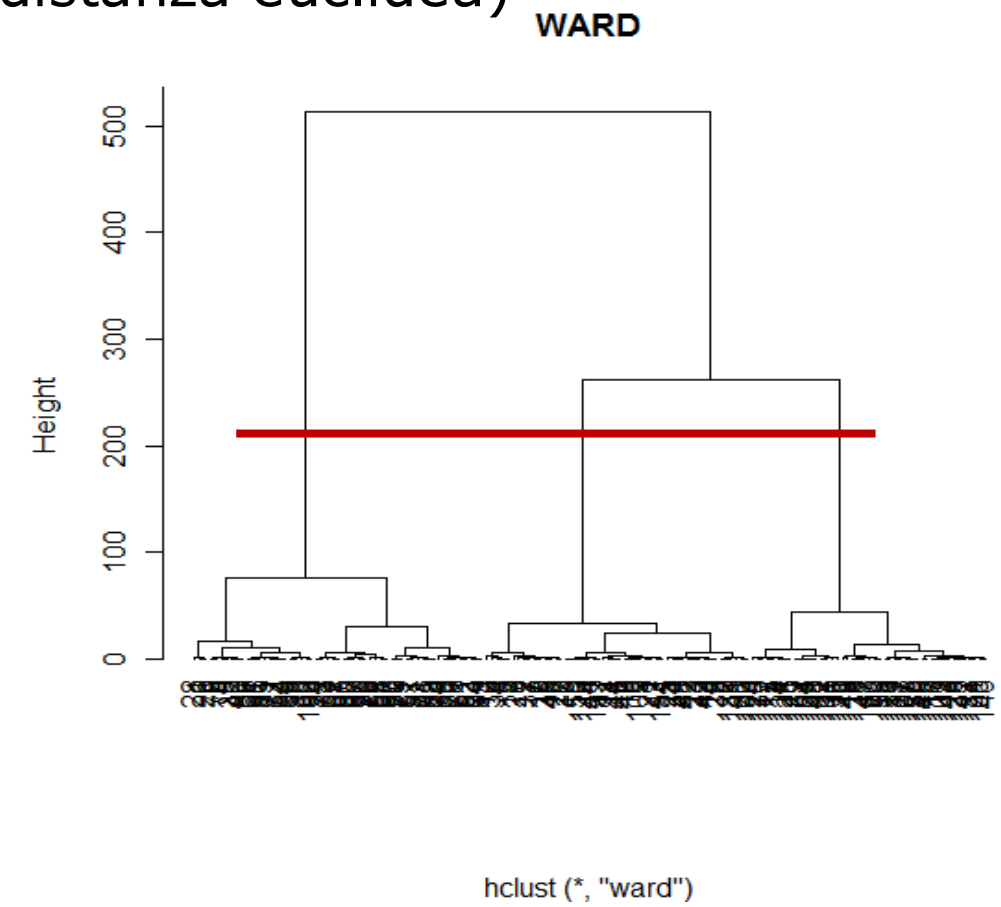






PEARSON

# Funzionamento dei metodi gerarchici: alcuni confronti (4/7) – 150 unità, 3 gruppi {50, 50, 50}

**Caso di gruppi ravvicinati ma di forma sferica**: metodo di Ward (quadrato della distanza euclidea)






PEARSON

# Funzionamento dei metodi gerarchici: alcuni confronti (5/7) – Gruppi ravvicinati di forma sferica

Legame singolo

| **Gruppo originario** | **Gruppo individuato** | | | **Totale** |
|---|---|---|---|---|
| | G1 | G2 | G3 | |
| G1 | 3 | **47** | 0 | 50 |
| G2 | 0 | **49** | 1 | 50 |
| G3 | 0 | **50** | 0 | 50 |
| Totale | 3 | 146 | 1 | 150 |

Legame completo

| **Gruppo originario** | **Gruppo individuato** | | | **Totale** |
|---|---|---|---|---|
| | G1 | G2 | G3 | |
| G1 | 5 | **40** | 5 | 50 |
| G2 | **39** | 0 | 11 | 50 |
| G3 | 0 | 6 | **44** | 50 |
| Totale | 44 | 46 | 60 | 150 |

Distanza media

| **Gruppo originario** | **Gruppo individuato** | | | **Totale** |
|---|---|---|---|---|
| | G1 | G2 | G3 | |
| G1 | 9 | **40** | 1 | 50 |
| G2 | **42** | 0 | 8 | 50 |
| G3 | 0 | 6 | **44** | 50 |
| Totale | 51 | 46 | 53 | 150 |

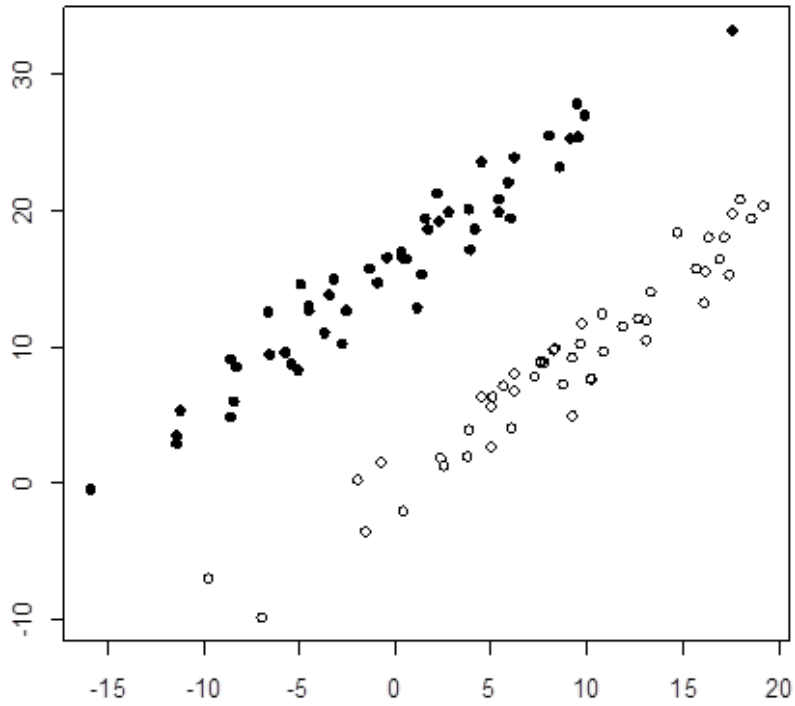
Ward

| **Gruppo originario** | **Gruppo individuato** | | | **Totale** |
|---|---|---|---|---|
| | G1 | G2 | G3 | |
| G1 | 9 | **41** | 0 | 50 |
| G2 | **46** | 1 | 3 | 50 |
| G3 | 0 | 8 | **42** | 50 |
| Totale | 55 | 50 | 45 | 150 |

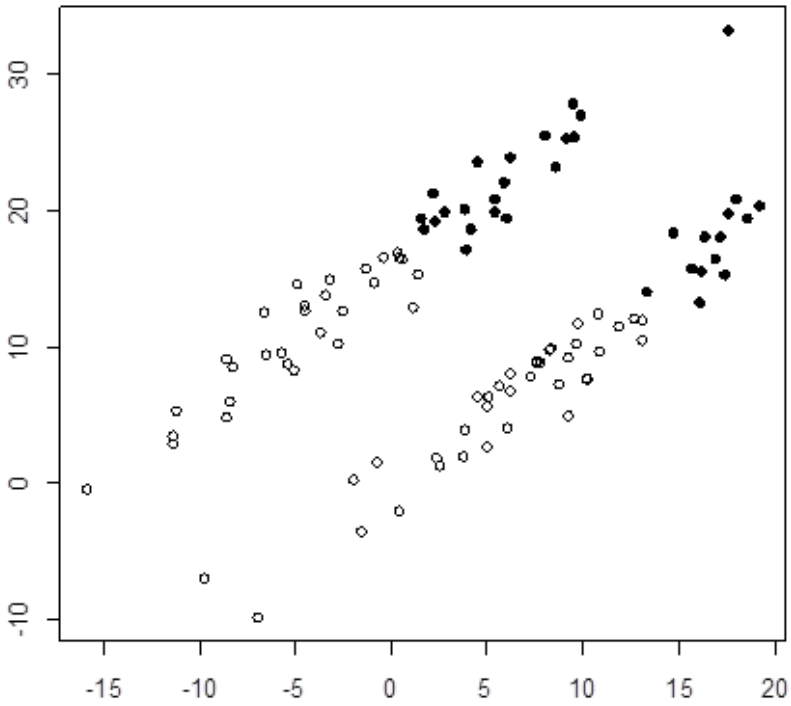
# Funzionamento dei metodi gerarchici: alcuni confronti (6/7) – 100 unità, 2 gruppi {50 50}

**Caso di gruppi allungati con variabili correlate:** nessuno di questi metodi riesce a individuarli

Gruppi originari




Gruppi individuati col metodo del legame completo

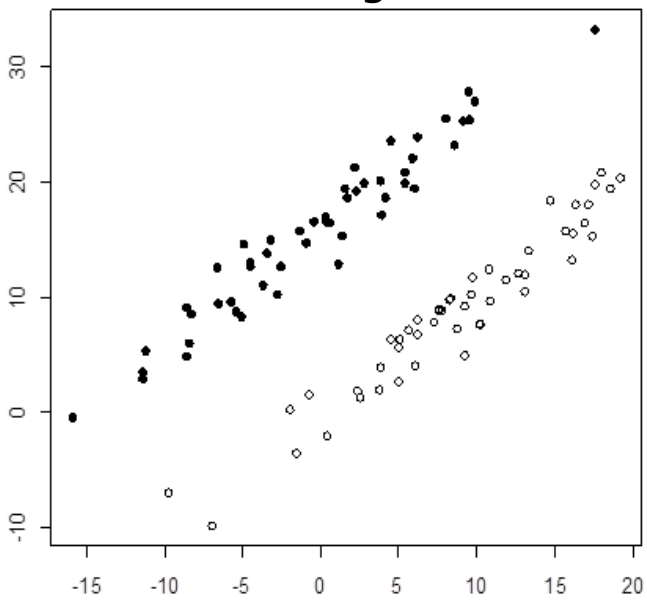




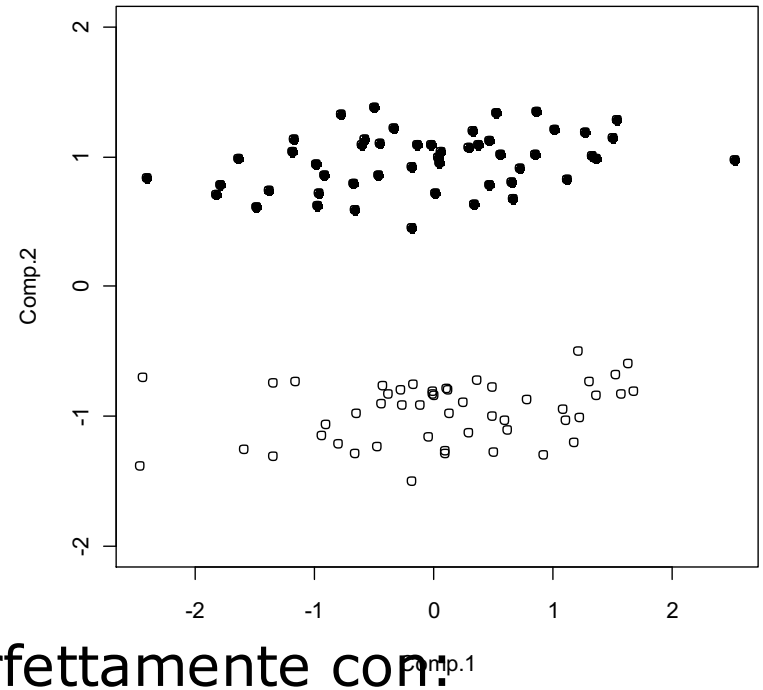

PEARSON

# Funzionamento dei metodi gerarchici: alcuni confronti (7/7) – 100 unità, 2 gruppi {50 50}

Assi originari




CP standardizzate e metodo di WARD




Si riesce ad individuarli perfettamente con:

- ACP su valori standardizzati
- standardizzazione delle componenti principali
- metodo di Ward sulle CP standardizzate



PEARSON

# 5. Interpretazione dei risultati

**1) Capire se i gruppi sono fra loro diversi rispetto alle variabili di cluster**

- scomposizione della varianza delle variabili di cluster
- box-plot per gruppo delle variabili di cluster

**2) Interpretazione dei gruppi**

Se i gruppi sono effettivamente diversi fra loro, interpretiamo queste differenze e possiamo anche analizzare eventuali variabili **esterne** (che non hanno partecipato all'analisi cluster)

**3) Comparazione a fini di benchmarking**

L'impresa individuerà la sua posizione in base a quella del gruppo di appartenenza e si confronterà con le unità del proprio gruppo.



PEARSON

# Esempio

- **30** imprese tessili: tessiture (**Ateco 13.2**)
- **6** indici di bilancio

$$\text{Indice di rigidità (RIG)} = \frac{\text{Attivo immobilizzato}}{\text{Totale attivo}}$$

$$\text{Current ratio (CR)} = \frac{\text{attivo circolante}}{\text{passività correnti}}$$

$$\text{BETA} = \frac{\text{totale debiti}}{\text{patrimonio netto}}$$

$$\text{ROA} = \frac{\text{reddito operativo}}{\text{totale attivo}}$$

$$\text{ROS} = \frac{\text{reddito operativo}}{\text{valore della produzione}}$$

$$\text{ROE} = \frac{\text{risultato prima delle imposte}}{\text{patrimonio netto}}$$



PEARSON

# Controllo varianza e correlazioni delle variabili

Varianza

| RIG | CR | BETA | ROA | ROE | ROS |
|---|---|---|---|---|---|
| 0.0236 | 0.1924 | 7.6998 | 0.0021 | 0.0287 | 0.0035 |

Si decide di standardizzare le variabili

Matrice di correlazione

| | RIG | CR | BETA | ROA | ROE | ROS |
|---|---|---|---|---|---|---|
| RIG | 1.000 | -0.463 | 0.233 | -0.002 | -0.043 | 0.162 |
| CR | -0.463 | 1.000 | -0.434 | -0.057 | -0.098 | -0.050 |
| BETA | 0.233 | -0.434 | 1.000 | -0.066 | 0.000 | -0.299 |
| ROA | -0.002 | -0.057 | -0.066 | 1.000 | 0.297 | 0.602 |
| ROE | -0.043 | -0.098 | 0.000 | 0.297 | 1.000 | 0.595 |
| ROS | 0.162 | -0.050 | -0.299 | 0.602 | 0.595 | 1.000 |

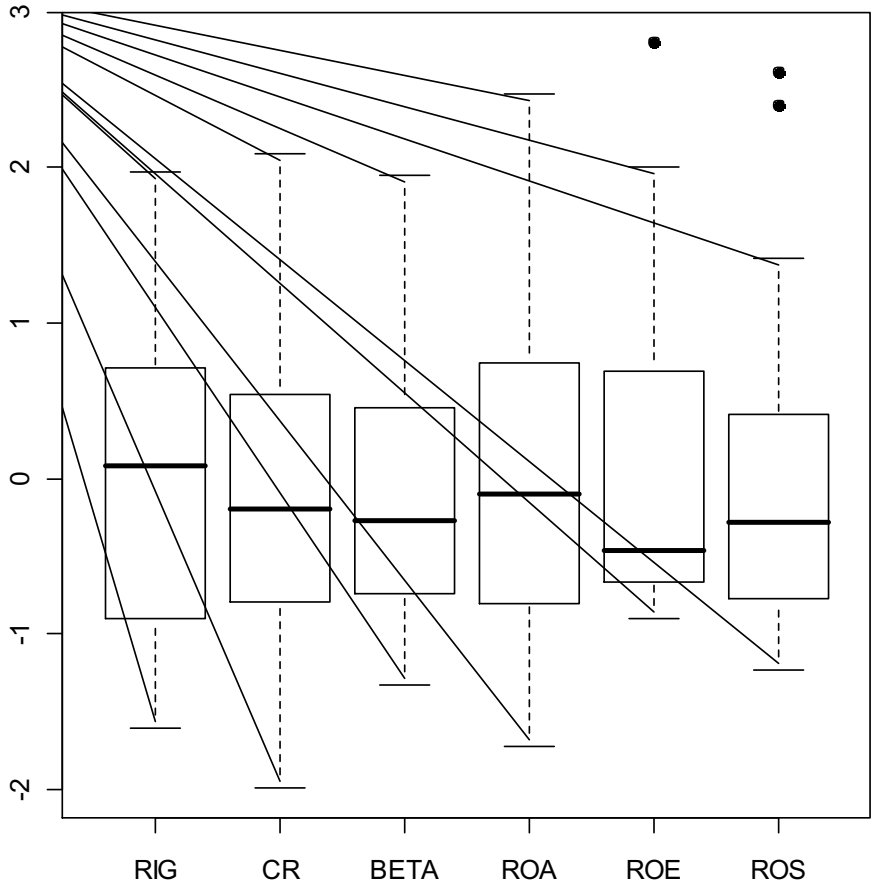
# Analisi descrittiva delle variabili standardizzate e scelta della distanza

Per il ROE e ROS (standardizzati) si osservano dei valori eccezionali.

Si tratta di imprese con una redditività superiore alle altre.

A fini esemplificativi si conduce l'analisi su tutti i dati.

Applichiamo il metodo di Ward col quadrato della distanza euclidea.

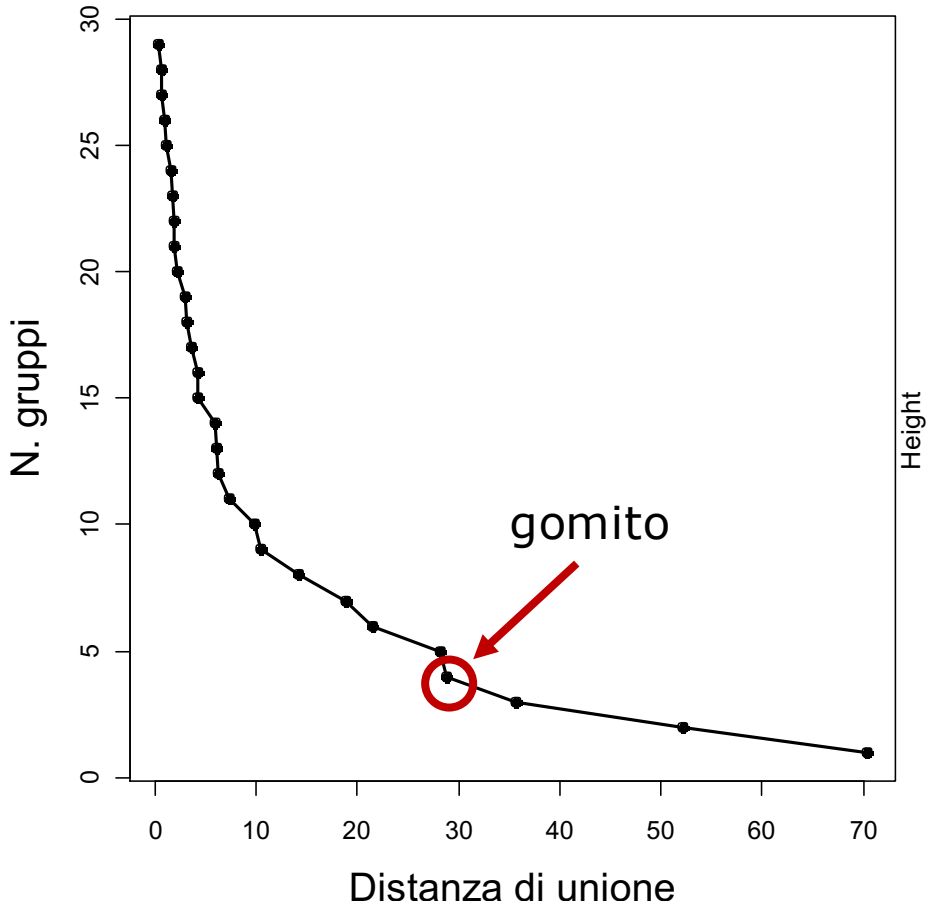


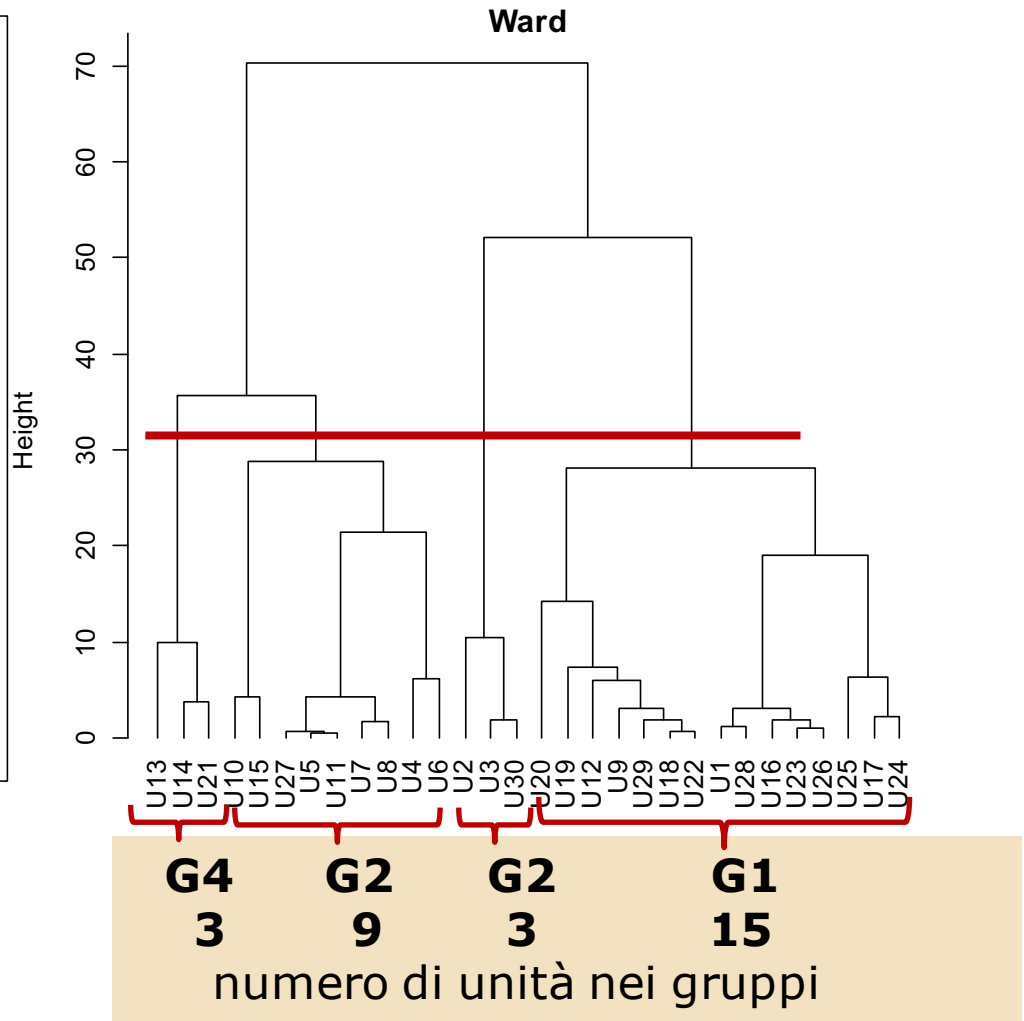

**N.B.** Possiamo fare un unico grafico perché le variabili sono standardizzate



PEARSON

# Metodo di Ward → 4 gruppi






PEARSON

# I gruppi sono ben differenziati ?

| DEVIANZA | RIG | CR | BETA | ROA | ROE | ROS |
|---|---|---|---|---|---|---|
| **Between** | 0.364 | 3.863 | 81.327 | 0.009 | 0.449 | 0.047 |
| **Interna** | 0.319 | 1.715 | 141.966 | 0.052 | 0.382 | 0.056 |
| **Totale** | 0.683 | 5.578 | 223.293 | 0.061 | 0.831 | 0.103 |
| **R-quadro** | 0.533 | 0.693 | 0.364 | 0.148 | 0.540 | 0.455 |

Dalla scomposizione della devianza si nota che i gruppi sono differenziati rispetto a RIG, CR e ROE

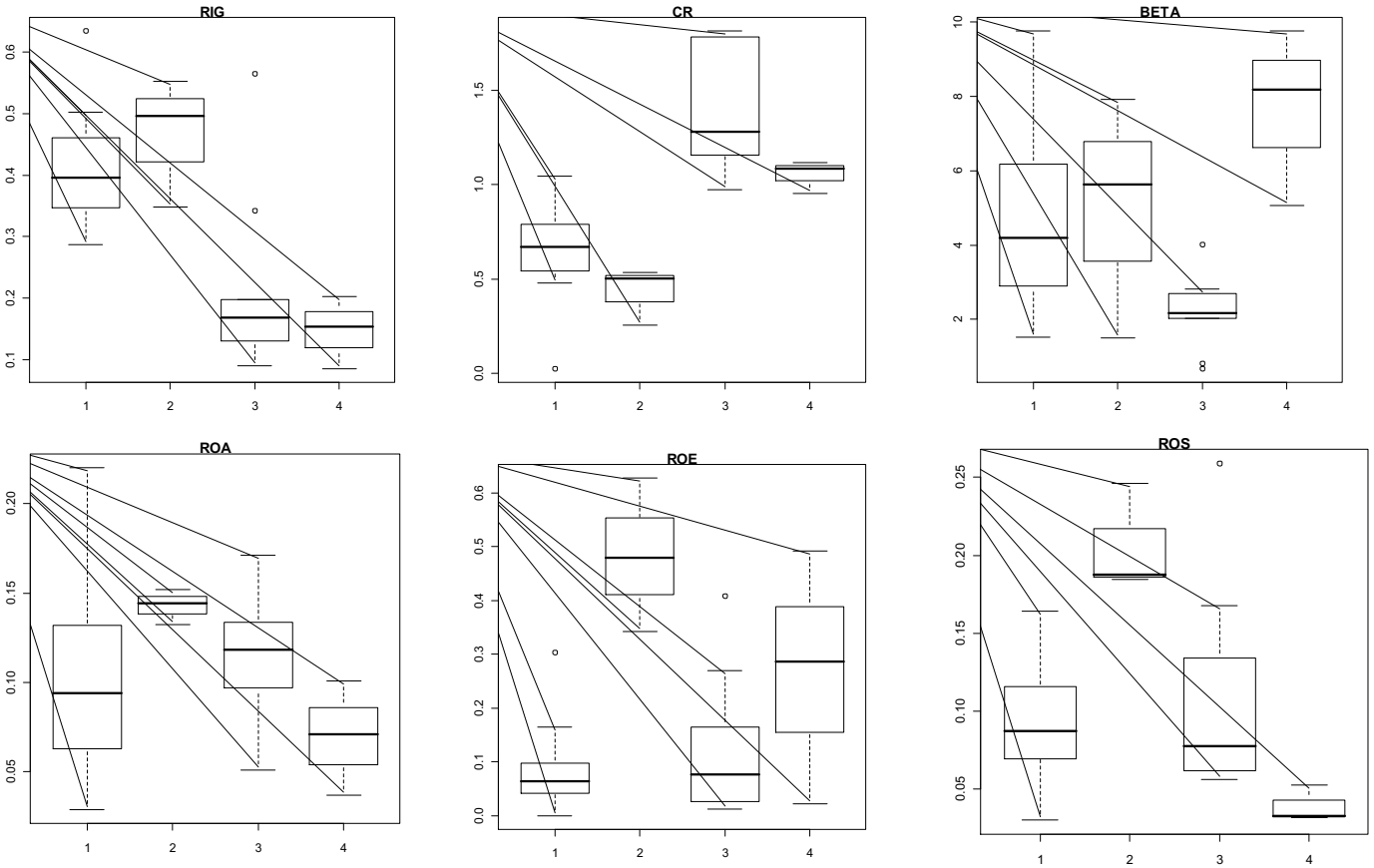
Lo sono in misura minore rispetto a BETA e ROA



PEARSON

# I gruppi sono ben differenziati ?

Un tipo di analisi più dettagliata è il boxplot separato per ogni variabile (lo facciamo, anche se con pochi dati non ha molto senso)






PEARSON

# Alcune interpretazioni

Il gruppo **G1** è piuttosto eterogeneo ma, rispetto agli altri, ha un livello di rigidità mediamente più elevato e un ROE più basso

Il gruppo **G2** ha una redditività (ROA, ROE, ROS) migliore degli altri, uno squilibrio fra attività e passività correnti (CR basso).

Il gruppo **G3** ha una più bassa rigidità gestionale, un elevato CR ma anche un elevato indebitamento (BETA): sembra che le imprese del gruppo finanziano passività a lungo termine con attività a breve. Questo sembra influenzare la redditività globale (ROE) che è più bassa di quella degli altri gruppi, a fronte di una redditività operativa (ROA) nella media.

Il gruppo **G4** ha la più bassa redditività operativa (ROA, ROS) e la più alta esposizione verso terzi (BETA elevato) dovuta a impegni a lungo termine essendo CR in media intorno a 1.

# Che cosa abbiamo imparato

- La logica esplorativa dell'analisi cluster gerarchica
- La matrice delle distanze fra unità e i diversi tipi di distanza
- I diversi algoritmi agglomerativi
- I problemi legati all'unità di misura e alla scala delle variabili
- Gli strumenti per l'interpretazione della procedura (scree plot e dendrogramma)
- Come procedere nell'interpretazione dei risultati



PEARSON