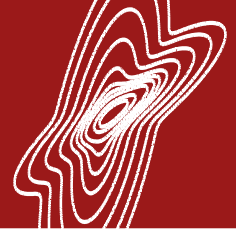


METODI STATISTICI PER LA BIOINGEGNERIA (B)

PARTE 16: CLUSTER ANALYSIS

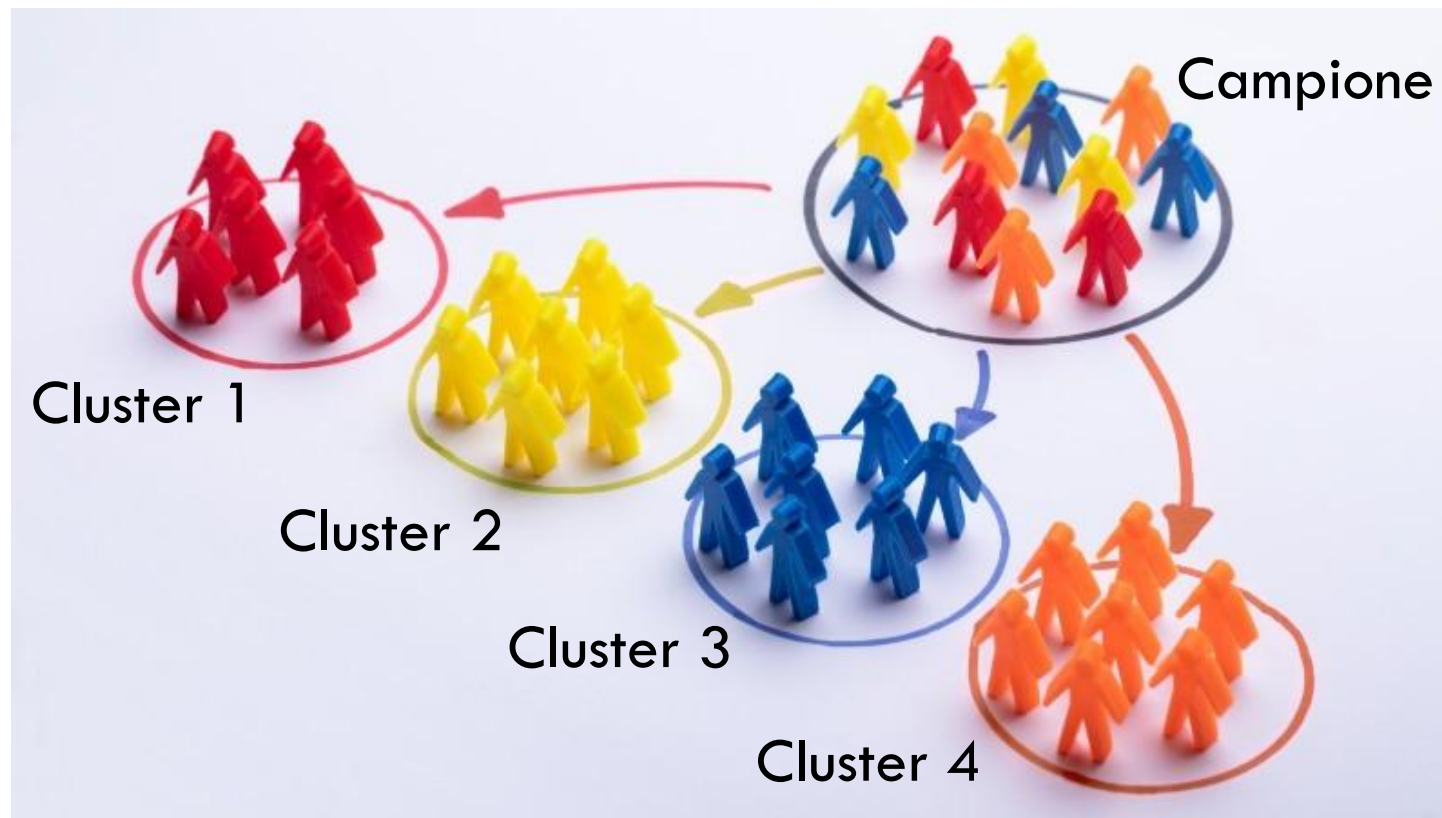
A.A. 2025-2026

Prof. Martina Vettoretti



CLUSTER ANALYSIS O CLUSTERING

- Obiettivo: determinare se il campione può essere suddiviso in **sottogruppi** relativamente distinti di osservazioni, detti **cluster**.



PRINCIPALI STEP DELLA CLUSTER ANALYSIS



Dataset

	X_1	X_2	...	X_m
Oss. 1				
Oss. 2				
...				
Oss. n				

Campione m-variato: n osservazioni (individui), m variabili (caratteristiche)

Distanze

	Oss. 1	Oss. 2	...	Oss. n
Oss. 1	0			
Oss. 2		0		
...			...	
Oss. n				0

Sulla base di una **misura di distanza**, si quantifica la distanza tra le osservazioni nel dataset

Algoritmo di clustering

Partizione del dataset

- Osservazioni sufficientemente simili tra loro si trovano nello stesso cluster
- Osservazioni sufficientemente diverse tra loro si trovano in cluster diversi

Cluster 1

Cluster 2

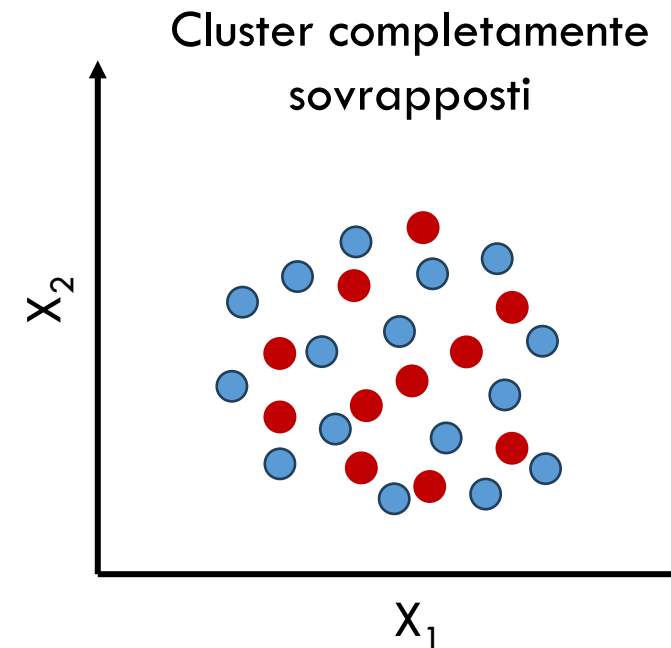
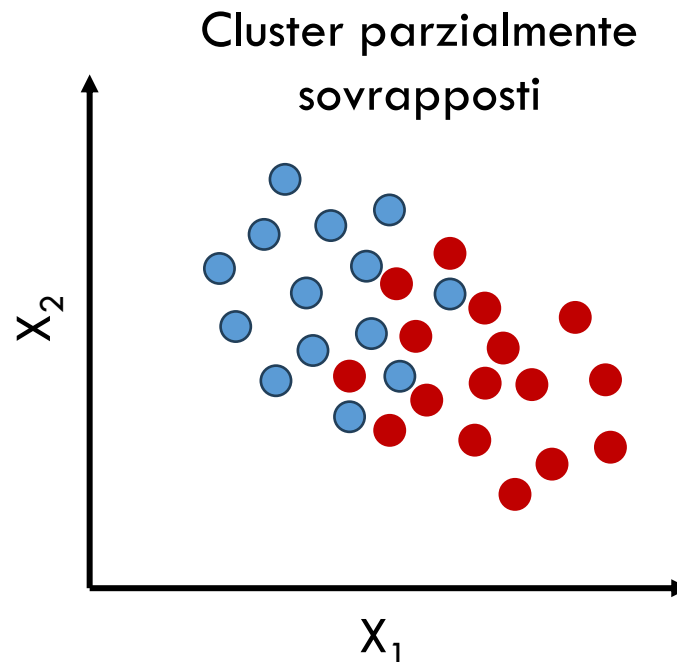
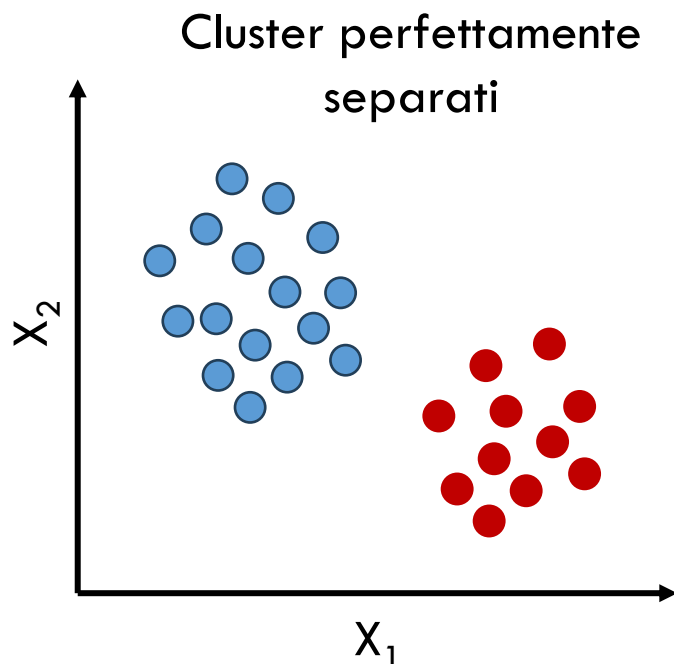
Cluster 3

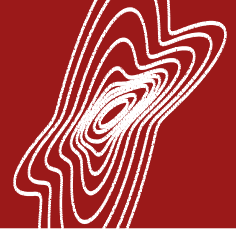


PRINCIPALE DIFFICOLTA' DELLA CLUSTER ANALYSIS



- I sottogruppi che stiamo cercando non sono noti a priori! Non sappiamo nemmeno se esistano...
- L'obiettivo della cluster analysis è proprio capire se le osservazioni si possono suddividere in sottogruppi sufficientemente omogenei e distinti tra loro. Dobbiamo mettere in conto che la risposta potrebbe essere no.





APPRENDIMENTO SUPERVISIONATO VS. NON SUPERVISIONATO



- Gli algoritmi di clustering fanno parte delle tecniche di apprendimento non supervisionato (*unsupervised learning*).
 - Non supervisionato perché nei dati che usiamo per la cluster analysis non abbiamo a disposizione un'etichetta (*label*) che ci indica il vero cluster di appartenenza di ciascuna osservazione.

- Le regressioni lineare, logistica e di Cox fanno parte delle tecniche di apprendimento supervisionato (*supervised learning*).
 - Supervisionato perché nei dati che usiamo per stimare i parametri del modello (training set) abbiamo una variabile che rappresenta l'outcome che vogliamo predire.

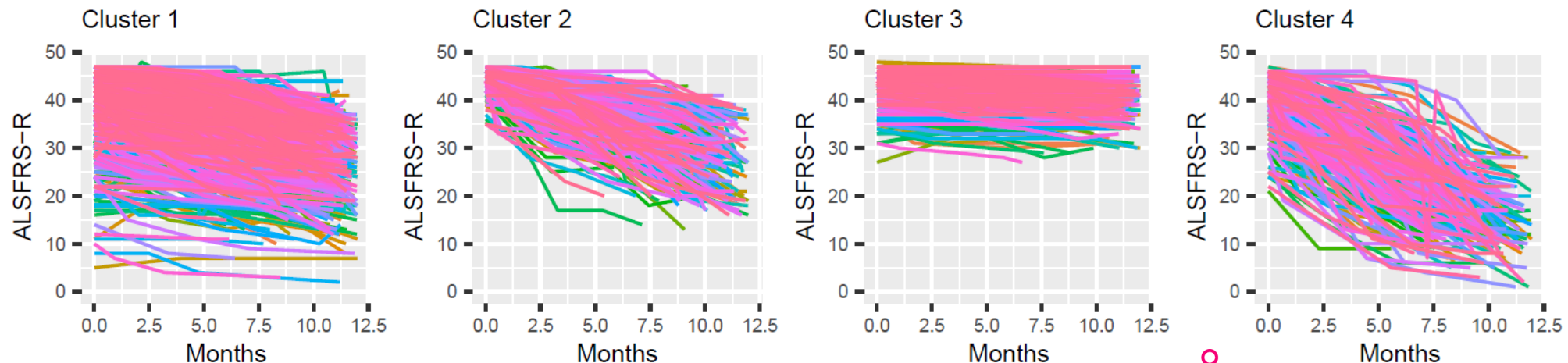
PERCHE' FARE CLUSTERING?

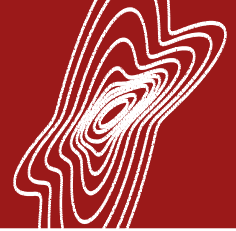


- **Ambito medico:** identificare sottogruppi di pazienti affetti da una stessa patologia che presentano diverse caratteristiche e possono quindi necessitare di trattamenti personalizzati.

Progetto BRAINTEASER (Horizon 2020)

- Stratificazione di pazienti affetti da SLA sulla base degli score funzionali misurati nel primo anno dopo la diagnosi → 4 diversi cluster di progressione
- Sviluppo di un modello predittivo che predice il cluster di progressione a cui appartiene un paziente usando le informazioni raccolte alla diagnosi

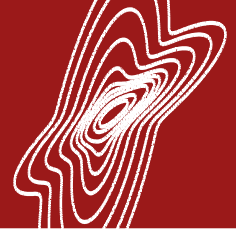




PERCHE' FARE CLUSTERING?



- **Imaging in ambito biomedico:** identificazione di parti di tessuto aventi proprietà alterate (es. tessuto danneggiato da una patologia).
- **Genomica:** identificazione di sottogruppi di geni sulla base della loro funzione o livello di espressione.
- **Rilevamento di anomalie:** identificazione di osservazioni che presentano caratteristiche anomale rispetto al resto della popolazione (es. rilevamento delle misure di un sensore affette da un artefatto).
- **Marketing:** identificazione di sottogruppi di clienti distinti per lo sviluppo di strategie di marketing mirate.

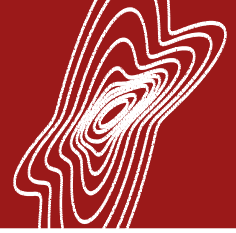


ALGORITMI DI CLUSTERING



➤ K-means

➤ Clustering gerarchico agglomerativo



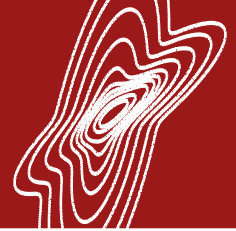
K-MEANS



- L'algoritmo K-means divide le osservazioni in **K cluster**, $C_i, i = 1, \dots, K$, dove K è un valore noto prestabilito:

$$C_1, C_2, \dots, C_K$$

- Ogni osservazione viene assegnata ad uno e un solo cluster.
- Idea: Si cerca la partizione che minimizza la **variabilità intra-cluster**.
- Questa è calcolata sulla base di una misura di distanza tra le osservazioni (da scegliere).
- Normalmente si usa la **distanza euclidea**.

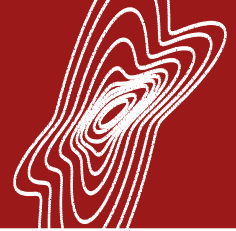


DISTANZA EUCLIDEA

- Osservazione i-esima: $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{im})$
Valori delle m variabili per
l'osservazione i-esima

- Distanza euclidea tra due osservazioni \mathbf{x}_i e \mathbf{x}_j :

$$d(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\| = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \dots + (x_{im} - x_{jm})^2}$$



FUNZIONE OBIETTIVO DEL K-MEANS



- **Variabilità intra-cluster** per il cluster k :

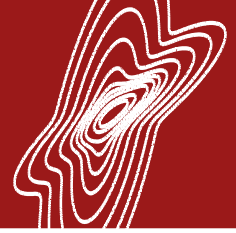
$$W(C_k) = \frac{1}{|C_k|} \sum_{i,j \in C_k} d(\mathbf{x}_i, \mathbf{x}_j)^2$$

dove $|C_k|$ è la cardinalità del cluster C_k .

Tanto più le osservazioni che appartengono al cluster C_k sono simili tra loro, tanto più piccola sarà $W(C_k)$.

- L'algoritmo K-means cerca la partizione, C_1, C_2, \dots, C_K , che minimizza la **somma delle variabilità intra-cluster**:

$$\operatorname{argmin}_{C_1, C_2, \dots, C_K} \sum_{k=1}^K W(C_k)$$



SOLUZIONE DEL PROBLEMA



➤ In pratica, la partizione ottima che minimizza la somma delle variabilità intra-cluster viene stimata mediante un algoritmo iterativo.

1. Ogni osservazione viene assegnata ad uno dei K cluster in modo casuale.

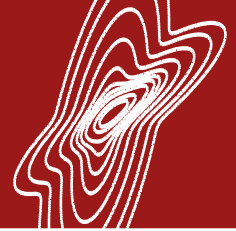
2. Si calcola il **centroide** μ_k per ciascun cluster mediando le variabili delle osservazioni che appartengono al cluster:

$$\mu_k = (\mu_{k1}, \mu_{k2}, \dots, \mu_{km})$$

$$\mu_{kj} = \frac{1}{|C_k|} \sum_{i \in C_k} x_{ij}$$

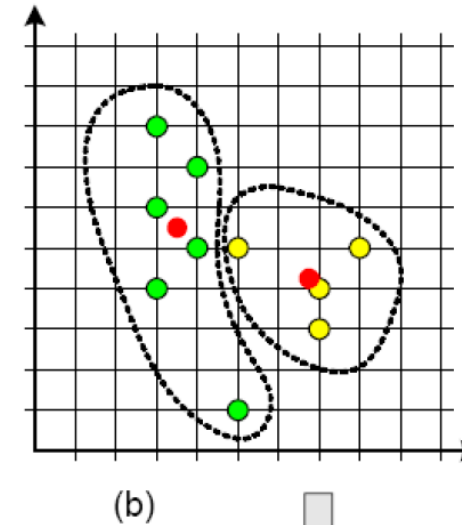
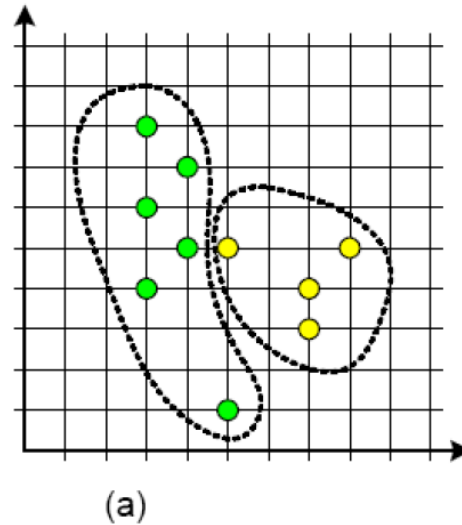
3. Per ogni osservazione si calcola la **distanza dai K centroidi**. Ogni osservazione viene assegnata al cluster corrispondente al centroide più vicino.

4. Si iterano i passi 2 e 3 finché i centroidi non cambiano più.

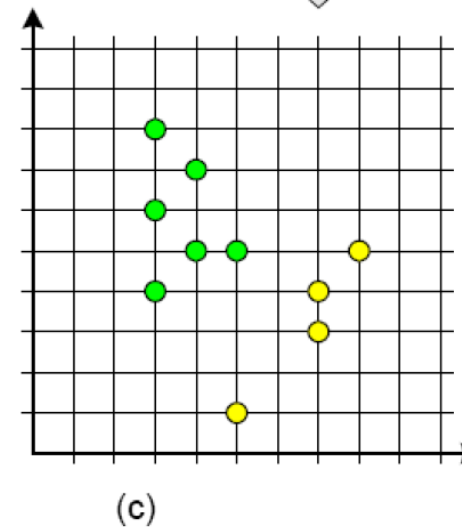


ESEMPIO

Cluster
all'iterazione i



Calcolo dei
centroidi per
ogni cluster

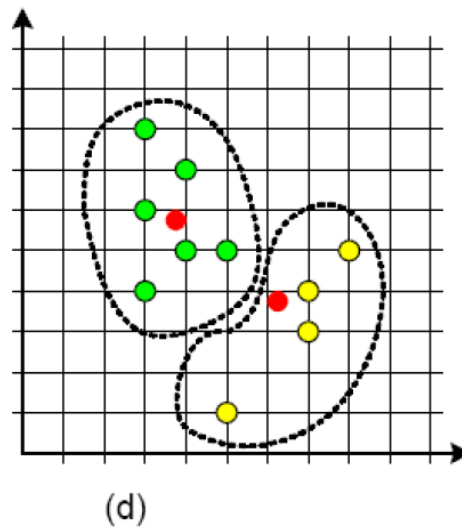


Assegnazione di ogni
osservazione al cluster
con centroide più vicino



Cluster all'iterazione
 $i+1$

Calcolo dei
centroidi per i
nuovi cluster



IL PROBLEMA DEI MINIMI LOCALI



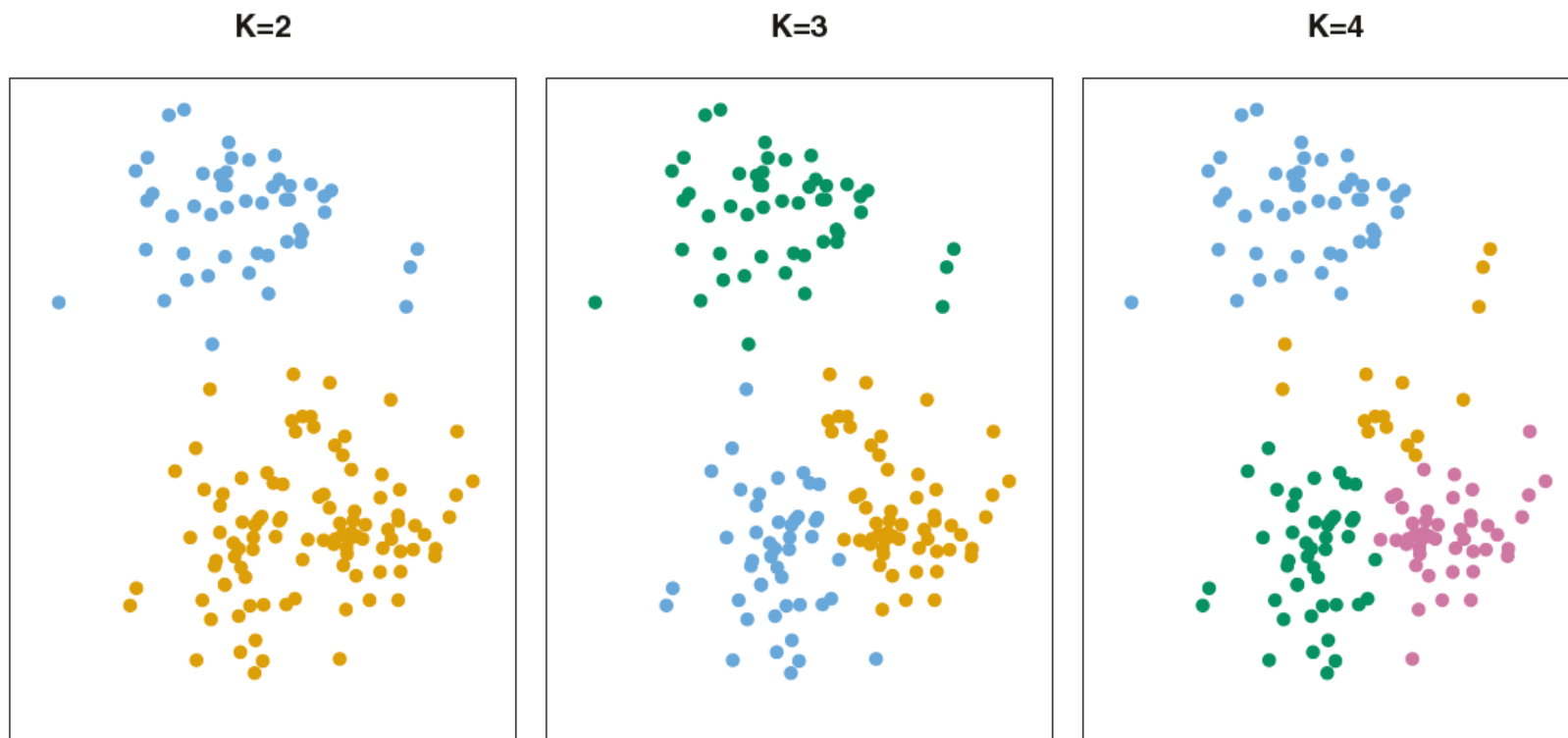
- L'algoritmo che ricerca la partizione ottima potrebbe fermarsi ad una suddivisione che rappresenta un minimo locale.
- Raccomandazione: eseguire l'algoritmo K-means più volte con diverse inizializzazioni dei cluster. Scegliere come partizione ottima quella avente somma delle variabilità intra-cluster minima.



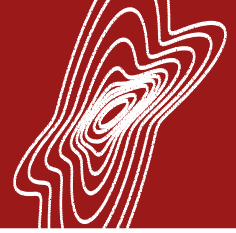
SCELTA DEL NUMERO DI CLUSTER



- Come scegliamo il numero di cluster K ?
- Generalmente si testano più valori di K e si sceglie quello per cui i risultati ci convincono di più.



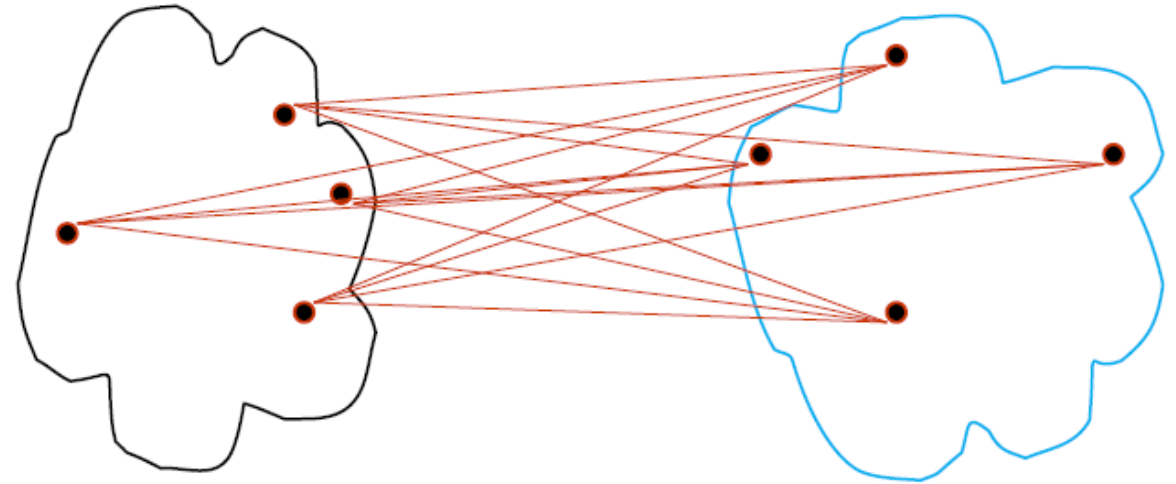
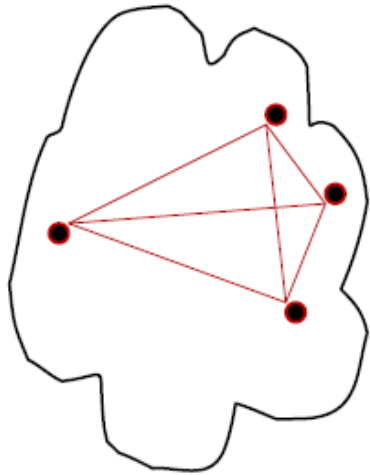
Nota: non c'è un ordine tra i cluster. I colori dei cluster sono assegnati in modo casuale.

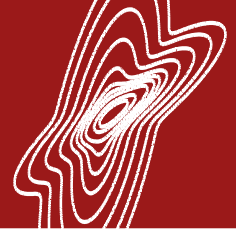


VALUTARE LA BONTA' DI UNA PARTIZIONE



- Esistono diverse metriche per valutare la bontà di una certa partizione.
- Queste in genere valutano due aspetti:
 - **Coesione dei cluster:** quanto sono vicine tra loro le osservazioni appartenenti allo stesso cluster.
 - **Separazione dei cluster:** quanto cluster diversi sono ben separati tra loro.





INDICE DI SILHOUETTE (1 / 2)

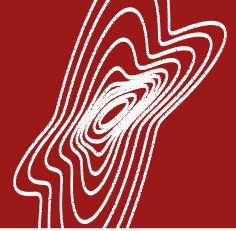


- L'**indice di silhouette** misura quanto ciascuna osservazione risulta simile alle osservazioni appartenenti allo stesso cluster (coesione), rispetto alle osservazioni degli altri cluster (separazione).
- Consideriamo l'*i*-esima osservazione, x_i .
- Distanza media di x_i dalle osservazioni appartenenti allo stesso cluster, C_I :

$$a(i) = \frac{1}{|C_I| - 1} \sum_{j \in C_I, i \neq j} d(x_i, x_j)$$

- Distanza media di x_i dalle osservazioni del cluster $C_J, J \neq I$:

$$b_J(i) = \frac{1}{|C_J|} \sum_{j \in C_J} d(x_i, x_j)$$



INDICE DI SILHOUETTE (2/2)

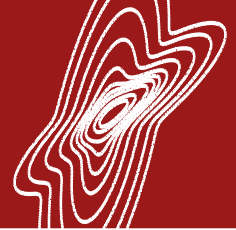


- Minima distanza media di x_i dalle osservazioni degli altri cluster:

$$b(i) = \min_{J \neq I} (b_J(i))$$

- Indice di silhouette per l'osservazione x_i :

$$s(i) = \begin{cases} \frac{b(i) - a(i)}{\max(a(i), b(i))} & \text{se } |C_I| > 1 \\ 0 & \text{se } |C_I| = 1 \end{cases}$$

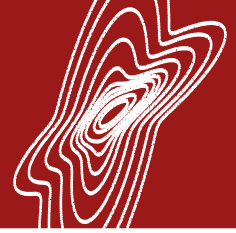


INTERPRETAZIONE DELL'INDICE DI SILHOUETTE



$$s(i) = \begin{cases} \frac{b(i) - a(i)}{\max(a(i), b(i))} & \text{se } |C_I| > 1 \\ 0 & \text{se } |C_I| = 1 \end{cases}$$

- $-1 \leq s(i) \leq 1$
- $s(i) = 1$ se $a(i) \ll b(i) \rightarrow x_i$ è mediamente molto meno distante dalle osservazioni del proprio cluster, C_I , rispetto a quelle degli altri cluster.
- $s(i) = -1$ se $b(i) \ll a(i) \rightarrow$ esiste un cluster C_J diverso da C_I per cui x_i risulta mediamente molto più vicina alle osservazioni di C_J piuttosto che a quelle di C_I .
- $s(i) = 0 \rightarrow$ l'osservazione i -esima è al confine tra due cluster.



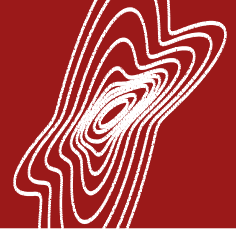
INDICE DI SILHOUETTE MEDIO



➤ **Indice di silhouette medio** sulle n osservazioni:

$$\bar{s} = \frac{1}{n} \sum_{i=1}^n s(i)$$

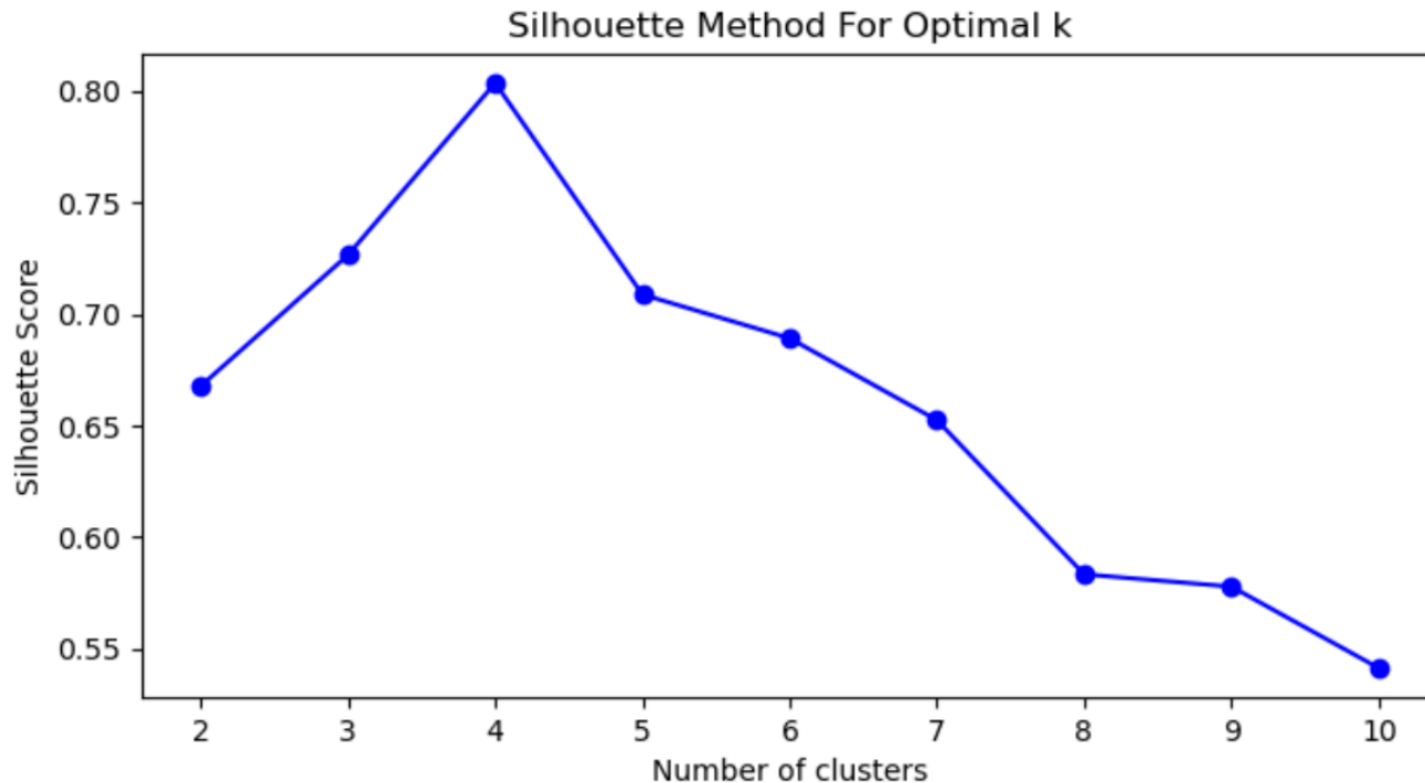
➤ Rappresenta una misura di quanto la partizione ottenuta sia buona.



SCELTA DEL NUMERO DI CLUSTER



- Possiamo provare diversi valori di K e scegliere quello che mi porta al massimo valore dell'indice di silhouette medio.



In questo caso che numero di cluster sceglieresti?

ALTRE POSSIBILI MISURE DI DISTANZA (1 / 3)



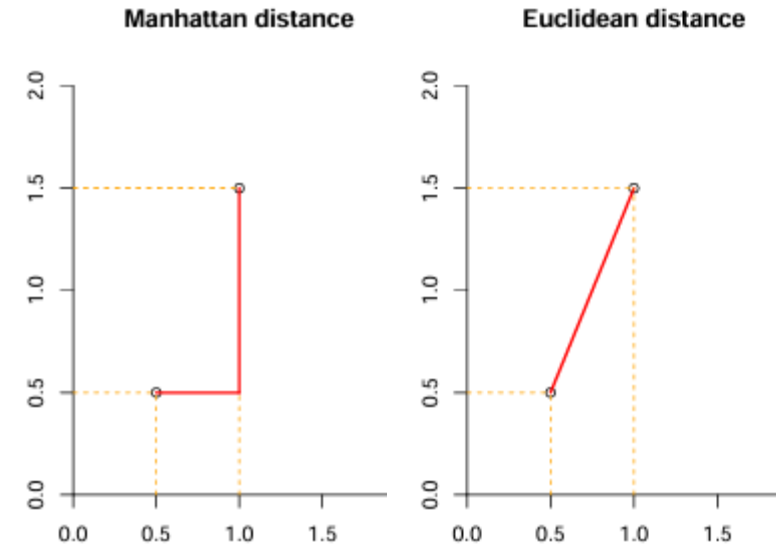
Per variabili quantitative:

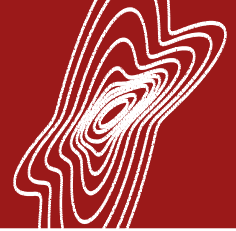
➤ **Distanza di Manhattan** o distanza «City Block»

$$d(x_i, x_j) = \sum_{k=1}^m |x_{ik} - x_{jk}|$$

➤ **Distanza di Minkowski**

$$d_p(x_i, x_j) = \sqrt[p]{\sum_{k=1}^m |x_{ik} - x_{jk}|^p}, \quad p \geq 1$$





ALTRE POSSIBILI MISURE DI DISTANZA (2/3)



Per variabili quantitative:

- **Distanza di Mahalanobis:** per calcolare la distanza di un'osservazione x_j dal centroide $\mu_k = (\mu_{k1}, \mu_{k2}, \dots, \mu_{km})$ di un insieme di osservazioni avente matrice di covarianza S .

$$d(x_j, \mu_k) = \sqrt{(x_j - \mu_k)^T S^{-1} (x_j - \mu_k)}$$

- Tiene conto della correlazione presente tra le variabili delle osservazioni che compongono il cluster.

Per variabili qualitative:

- **Distanza di Hamming:** quantifica il numero di variabili che presentano valori diversi nelle due osservazioni a confronto (x_i e x_j).

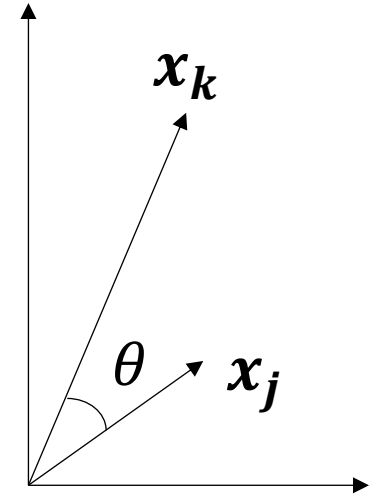
ALTRE POSSIBILI MISURE DI DISTANZA (3/3)



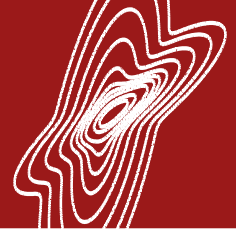
Per dati vettoriali:

➤ Distanza coseno:

$$\cos(\theta) = \frac{\mathbf{x}_i \cdot \mathbf{x}_j}{\|\mathbf{x}_i\| \cdot \|\mathbf{x}_j\|} = \frac{\sum_{k=1}^m x_{ik} \cdot x_{jk}}{\sqrt{\sum_{k=1}^m x_{ik}^2} \cdot \sqrt{\sum_{k=1}^m x_{jk}^2}}$$



- Misura la distanza tra le direzioni di due vettori, non importa quanto lunghi essi siano.
- Si usa nell'ambito dell'analisi di documenti di testo, per quantificare quanto due documenti siano distanti in termini di contenuto (orientamento dei vettori).



CLUSTERING K-MEANS: PRO E CONTRO

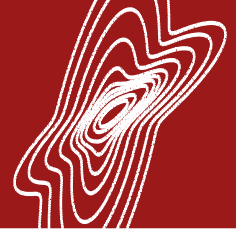


➤ Pro:

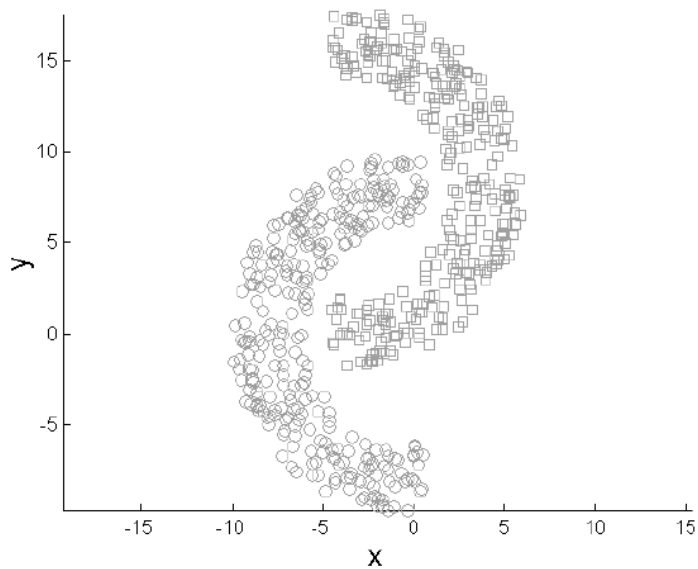
- Semplice e veloce

➤ Contro:

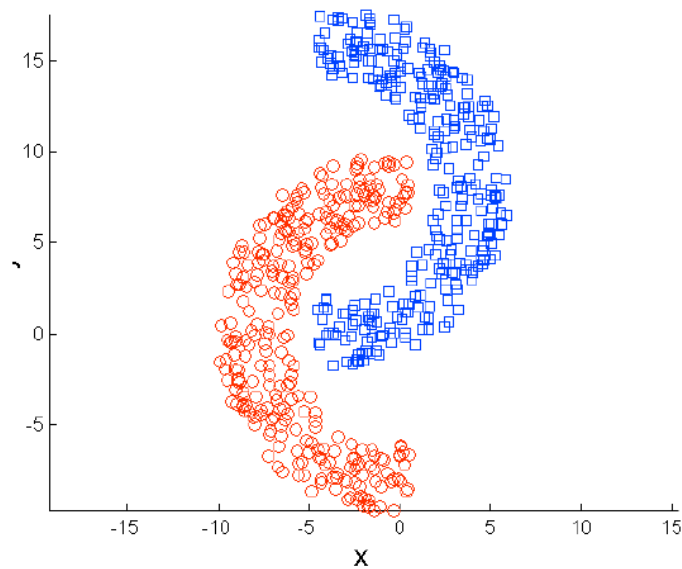
- Il numero di cluster va prespecificato
- Può convergere ad un minimo locale
- Il risultato può essere influenzato in modo importante da outlier
- Può creare solo cluster di forma globulare



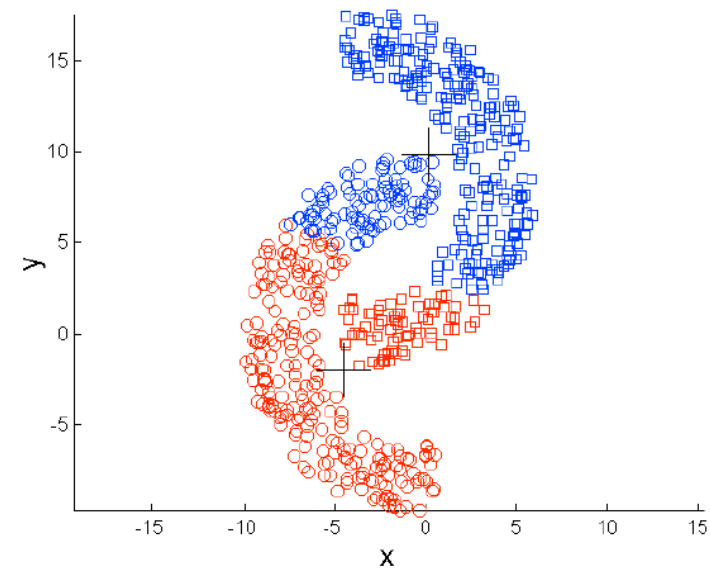
CLUSTER NON GLOBULARI



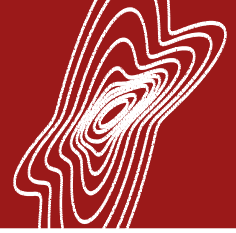
Dati non suddivisi



Possibile partizione in
due cluster

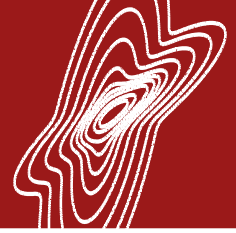


Suddivisione in due cluster
ottenuta con K-means



➤ K-means

➤ Clustering gerarchico agglomerativo



CLUSTERING GERARCHICO AGGLOMERATIVO



➤ **Clustering gerarchico agglomerativo:** algoritmo di clustering che raggruppa le osservazioni sfruttando un approccio gerarchico bottom-down.

1. Assegna ogni osservazione ad un cluster diverso → n cluster

➤ 2. Calcola le distanze tra tutte le coppie di cluster

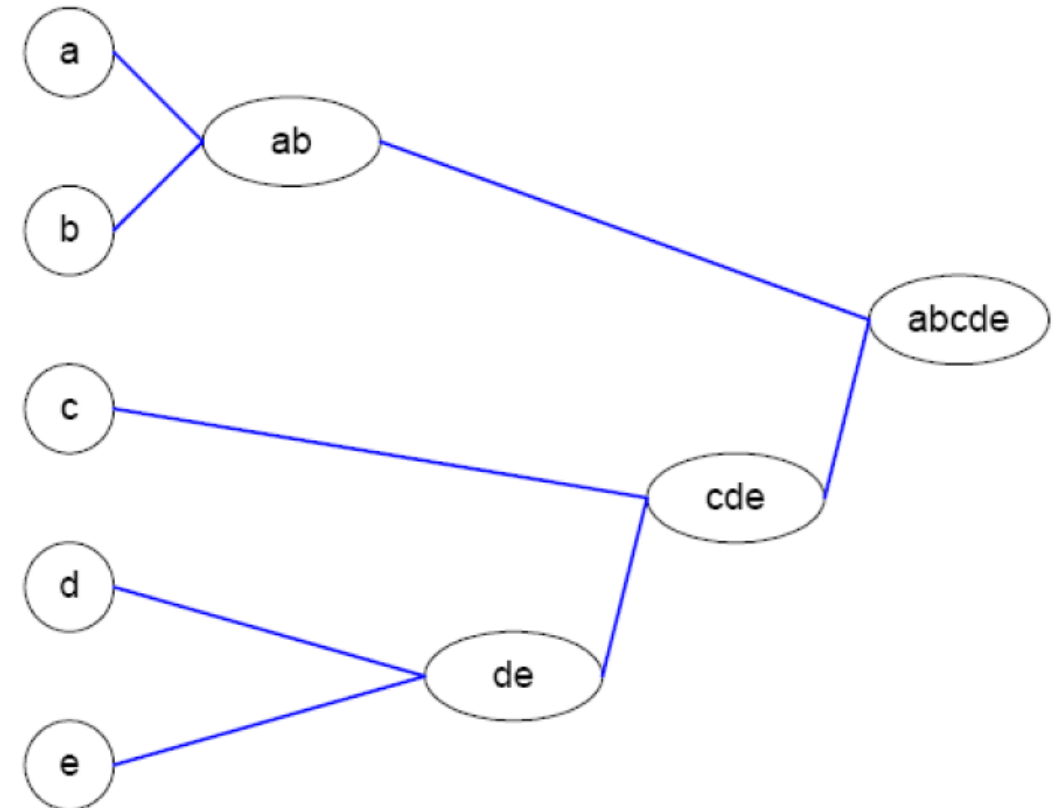
3. Fondi nello stesso cluster i due cluster più vicini tra loro

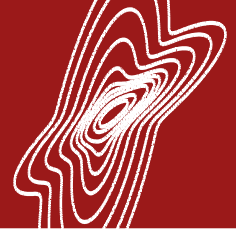
4. Ripeti gli step 2 e 3 finché tutte le osservazioni sono raggruppate in un unico cluster.

ESEMPIO



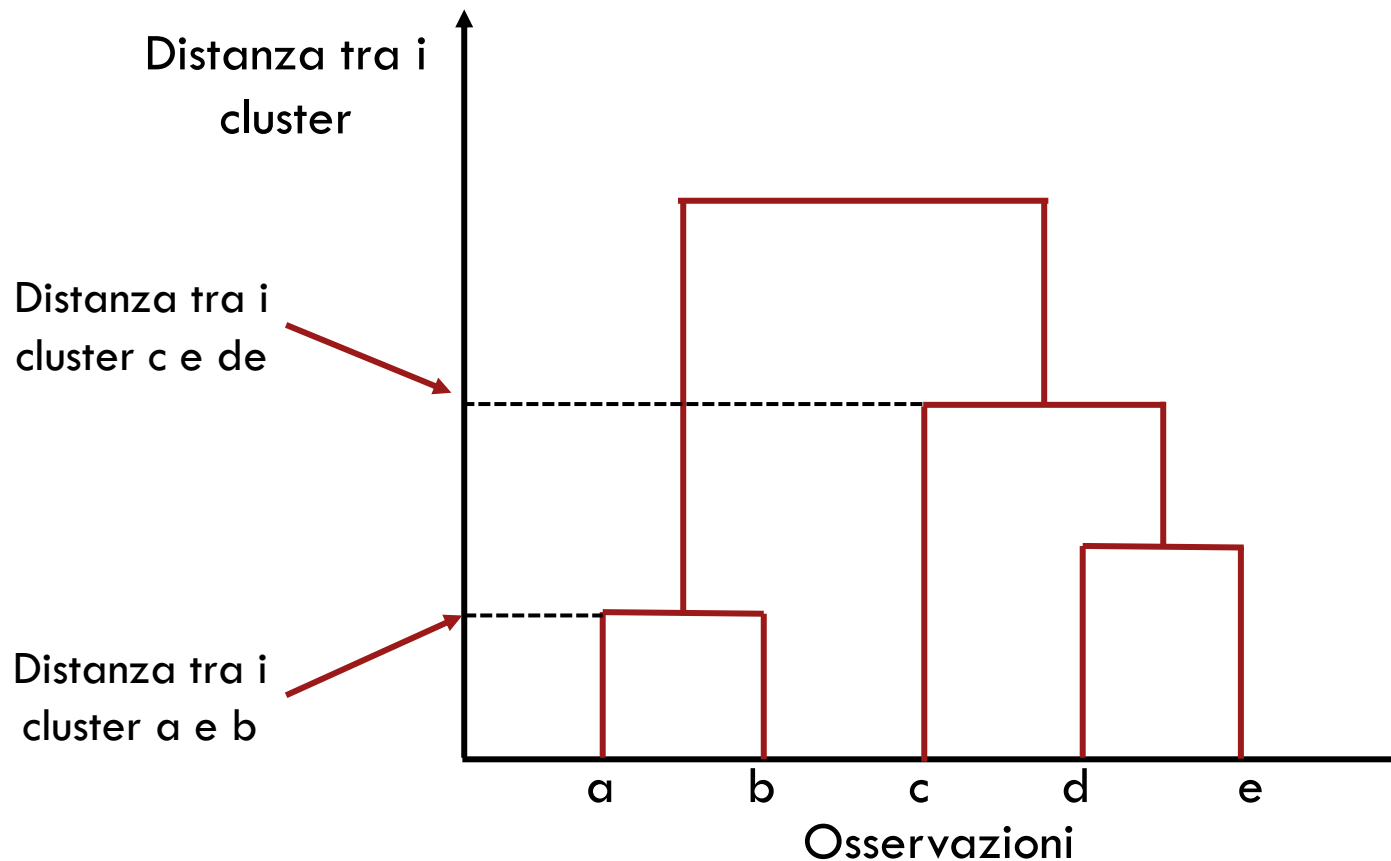
- 5 cluster: a, b, c, d, e
 - Distanze tra tutte le coppie di cluster: (a,b), (a,c), (a,d), (a,e), (b,c), (b,d), (b,e), (c,d), (c,e), (d,e).
 - Coppia a distanza minima: (a,b) → nuovo cluster ab
- 4 cluster: ab, c, d, e
 - Distanze tra tutte le coppie di cluster: (ab,c), (ab,d), (ab,e), (c,d), (c,e), (d,e).
 - Coppia a distanza minima: (d,e) → nuovo cluster de
- 3 cluster: ab, c, de
 - Distanze tra tutte le coppie di cluster: (ab,c), (ab,de), (c,de).
 - Coppia a distanza minima: (c,de) → nuovo cluster cde
- 2 cluster: ab, cde
 - Nuovo cluster abcde.





IL DENDROGRAMMA

➤ **Dendrogramma:** grafico ad albero che rappresenta i raggruppamenti realizzati dall'algoritmo di clustering gerarchico agglomerativo.

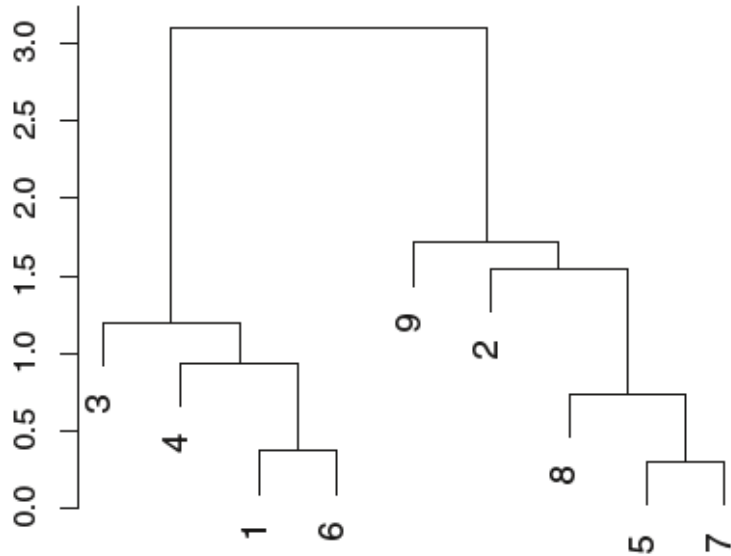


- Le foglie rappresentano le osservazioni.
- I nodi rappresentano le fusioni tra i cluster.
- L'altezza di ciascun nodo rappresenta la distanza tra i due cluster fusi in corrispondenza di quel nodo.
- L'ordine delle osservazioni sull'asse delle ascisse è «di comodo» per la rappresentazione. Due osservazioni vicine nell'asse delle ascisse non sono necessariamente simili tra loro.
- Gli stessi raggruppamenti si possono rappresentare con diversi dendrogrammi equivalenti.

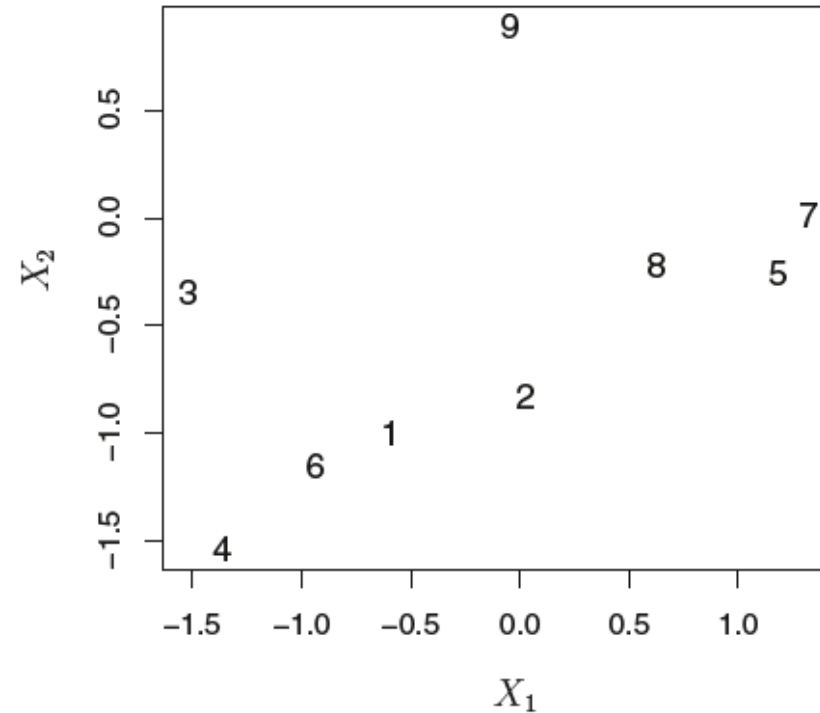
ESEMPIO



Dendrogramma

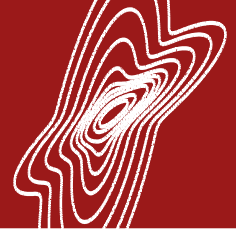


Rappresentazione delle osservazioni nello spazio delle variabili X_1 , X_2



Note:

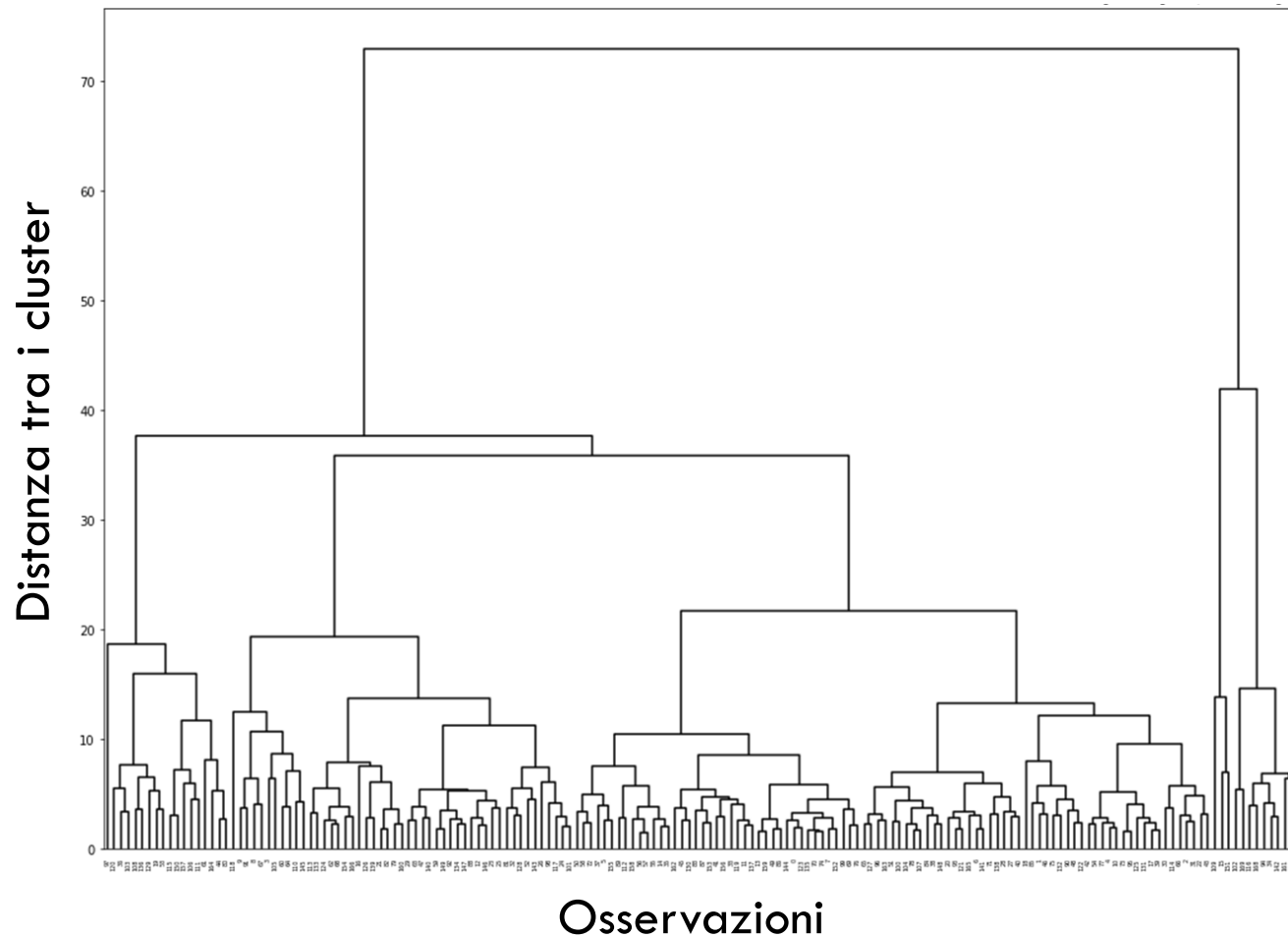
- Le osservazioni 9 e 2, seppur in posizioni vicine nel dendrogramma, sono distanti tra loro nel piano X_1, X_2 .
- Le osservazioni vicine tra loro vengono unite in punti a bassa altezza nel dendrogramma (come ad esempio le osservazioni 5 e 7).

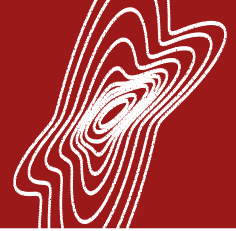


UN DENDROGRAMMA PIU' COMPLESSO



- Con maggiore numero di osservazioni il dendrogramma si fa più complesso.

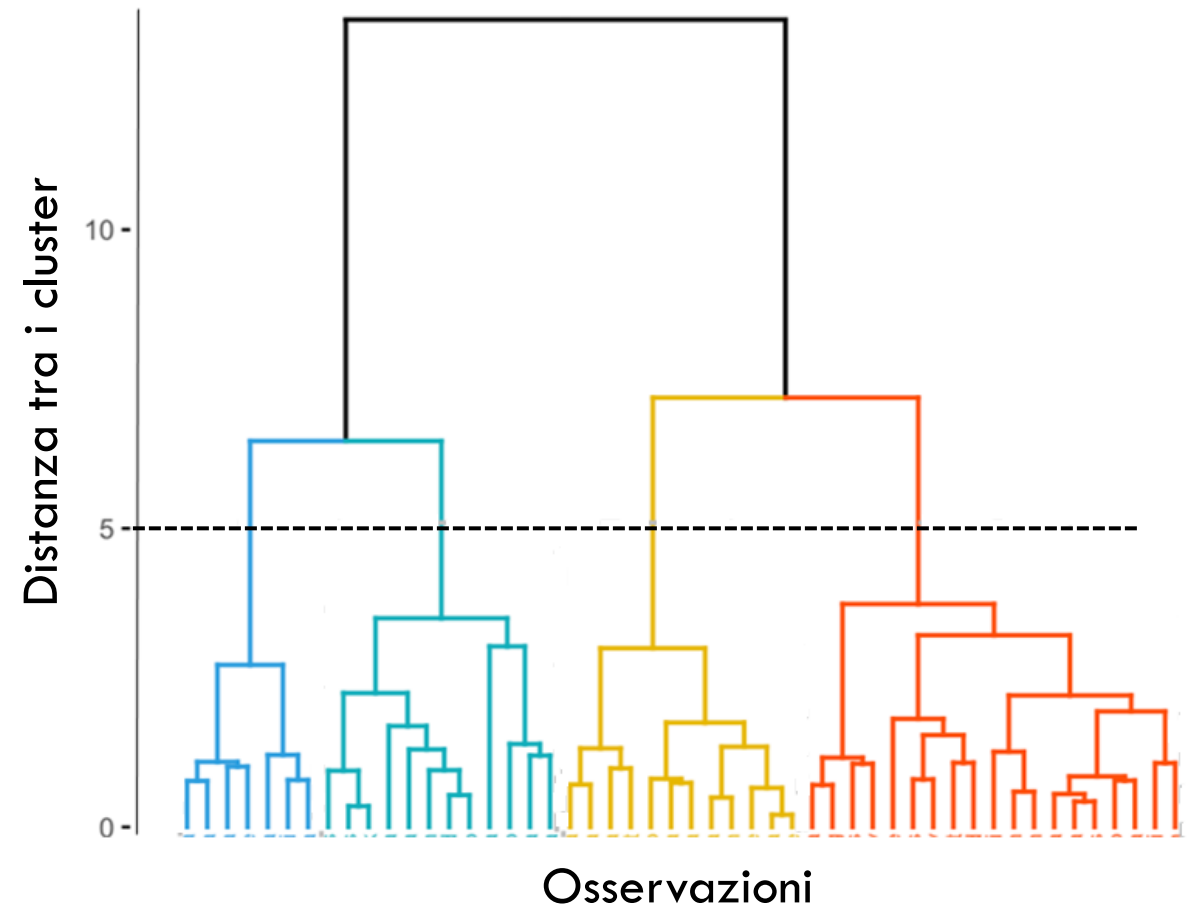




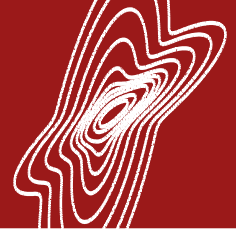
TAGLIO DEL DENDROGRAMMA



- Il numero di cluster non è stabilito a priori come nel K-means.
- Il clustering gerarchico agglomerativo realizza diversi raggruppamenti con un numero di cluster che varia tra n e 1.
- Per definire il numero di cluster finale dobbiamo tagliare il dendrogramma orizzontalmente ad una certa altezza.

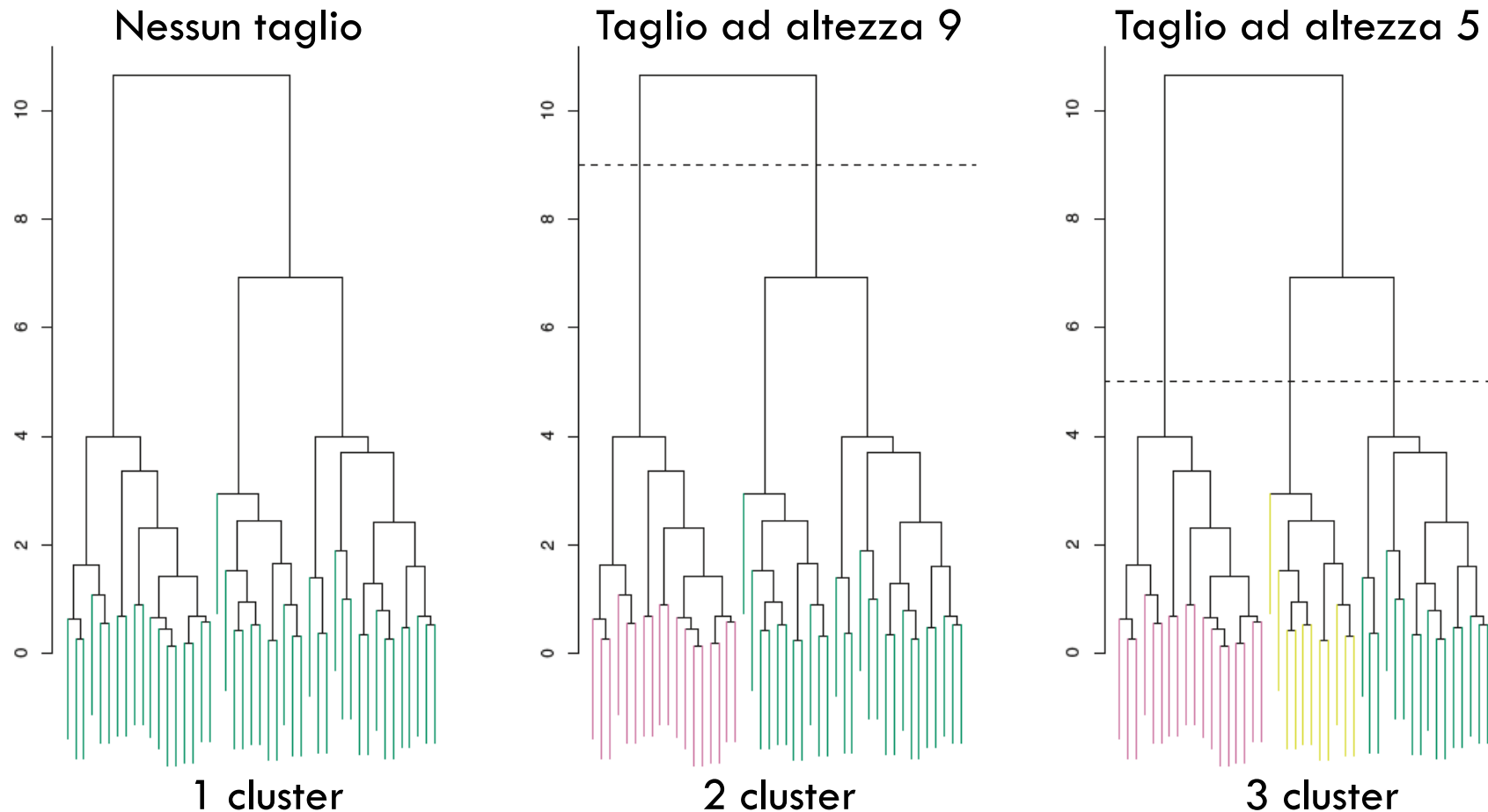


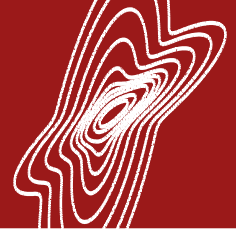
Tagliando il dendrogramma ad altezza 5 quanti cluster otteniamo?



L'ALTEZZA DI TAGLIO

- L'altezza a cui tagliamo il dendrogramma definisce il numero di cluster.

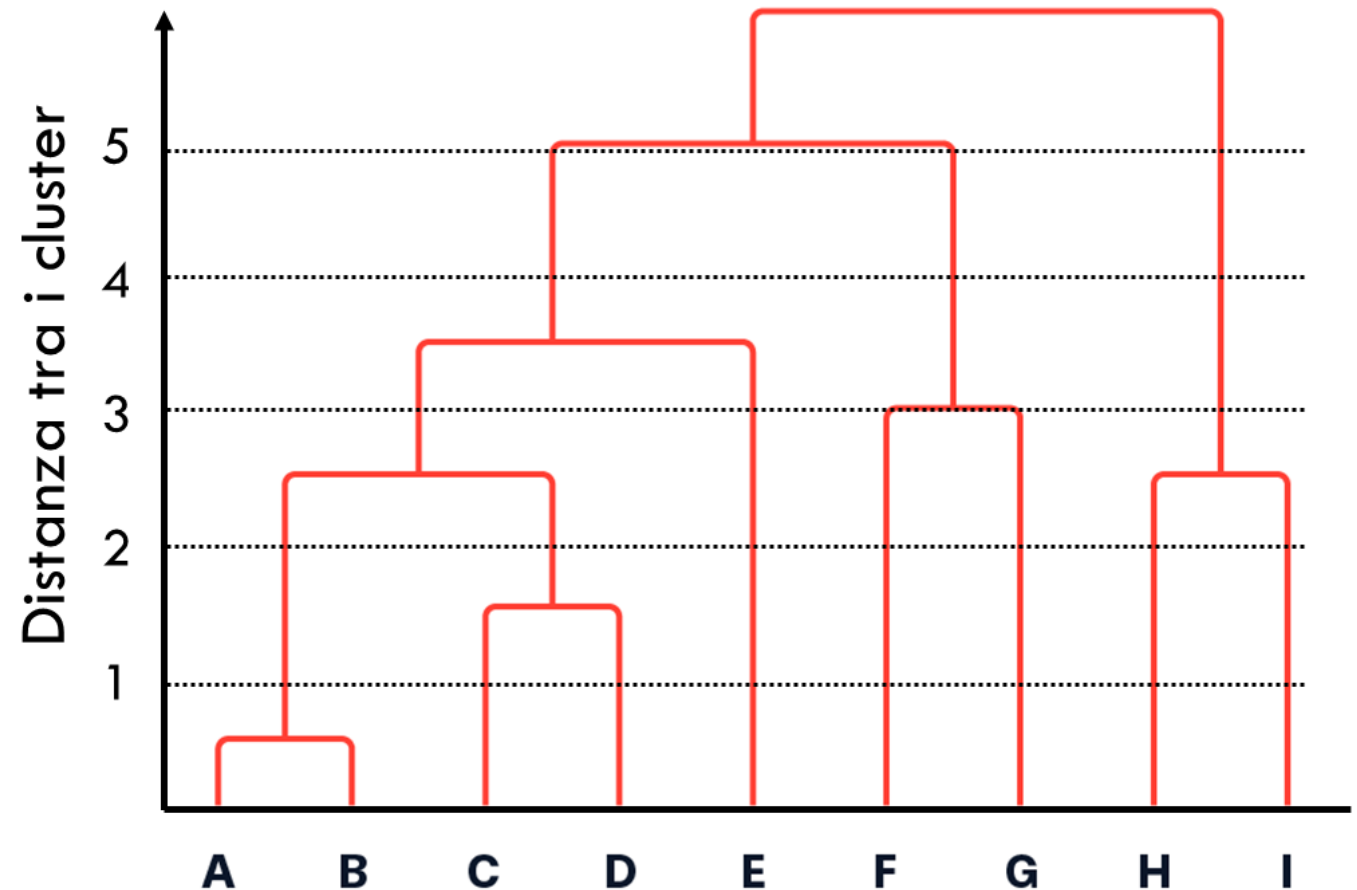


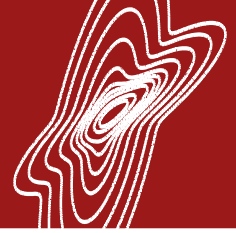


ESERCIZIO



1. Quali sono le due osservazioni più vicine tra loro?
2. Quanto misura la distanza tra le osservazioni F e G?
3. Se tagliamo il dendrogramma ad altezza 4 quanti cluster otteniamo?

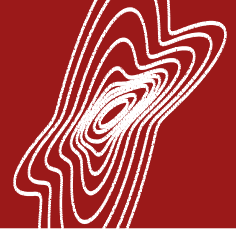




SCELTA DELL'ALTEZZA DI TAGLIO



- Scelta sulla base dell'ispezione visiva del dendrogramma.
- Scelta sulla base del numero desiderato di cluster.
- Scelta sulla base di un indice quantitativo.
 - Altezza di taglio che mi consente di massimizzare il valore dell'indice di silhouette medio.
 - Indice di inconsistenza.



INDICE DI INCONSISTENZA



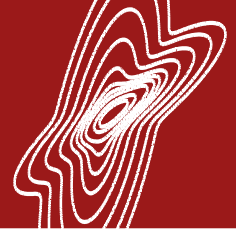
➤ **Indice di inconsistenza:** valuta quanto una fusione (nodo) in un dendrogramma sia coerente con le fusioni vicine.

➤ Indice di inconsistenza della fusione (nodo) k-esima:

$$I_k = \frac{h_k - \text{mean}(\{h\})}{\text{sd}(\{h\})}$$

- h_k : altezza della fusione (nodo) k-esima
- $\{h\}$: altezze delle fusioni (nodi) vicini al nodo k-esimo

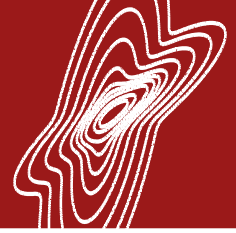
➤ Serve a individuare punti di fusione "anomali", dove due cluster vengono uniti a una distanza significativamente maggiore rispetto alle precedenti o a quelle di pari livello → informazione utile per decidere a che altezza tagliare il dendrogramma



SOGLIA SULL'INDICE DI INCONSISTENZA



- Valori comuni per l'indice di inconsistenza sono tra 0.5 e 2.
- Possiamo decidere di tagliare il dendrogramma quando la fusione di due cluster risulta in un indice di inconsistenza superiore a una certa soglia.
 - Soglia pari a 1 → buon punto di partenza che generalmente porta a cluster ben separati
 - Soglia > 1.5 → pochi cluster, più grandi e meno dettagliati
 - Soglia < 1.0 → tanti cluster, più piccoli e dettagliati → potrebbero essere difficili da interpretare per il rumore nei dati

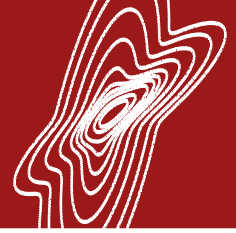


DISTANZA TRA I CLUSTER



Quando si implementa l'algoritmo di clustering gerarchico bisogna scegliere 2 misure di distanza:

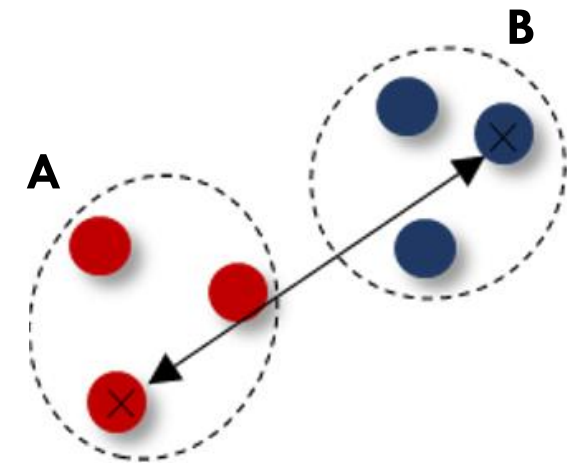
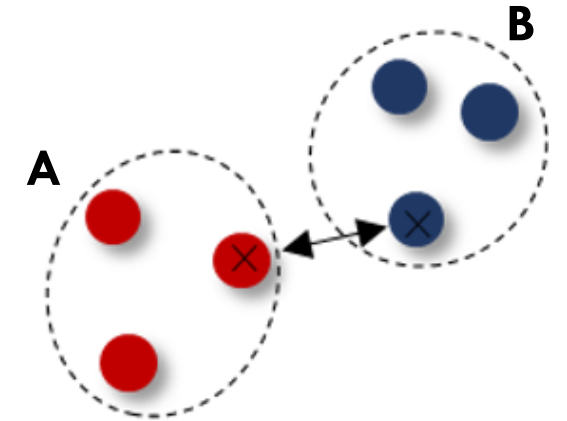
- Misura di **distanza tra due osservazioni** → tipicamente distanza euclidea, ma si possono usare anche altre distanze (Manhattan, Hamming ecc.)
- Criterio per stabilire la distanza tra due cluster (**criterio di linkage**):
 - Single linkage
 - Complete linkage
 - Average linkage
 - Centroid linkage
 - Ward's linkage

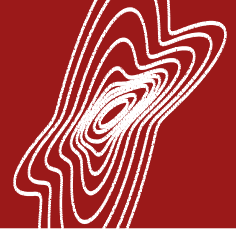


CRITERI DI LINKAGE (1 / 3)



- **Single linkage:** la distanza tra i cluster A e B è la distanza minima tra un'osservazione appartenente al cluster A e un'osservazione appartenente al cluster B.
- **Complete linkage:** la distanza tra i cluster A e B è la distanza massima tra un'osservazione appartenente al cluster A e un'osservazione appartenente al cluster B.

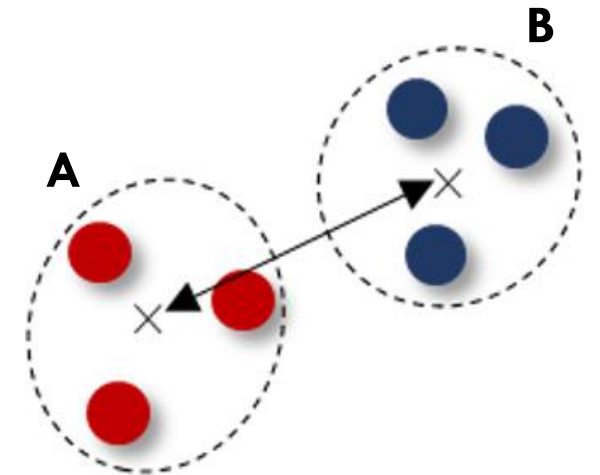
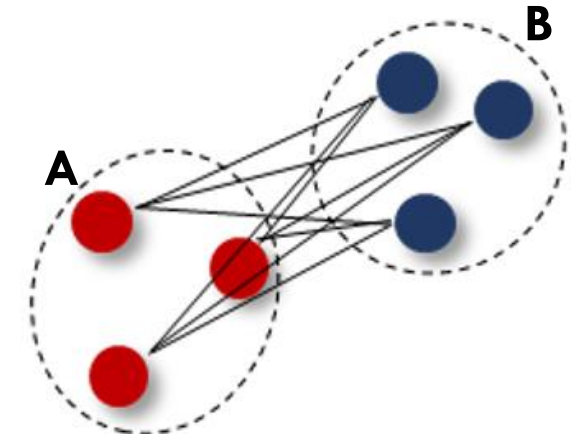


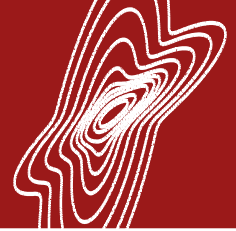


CRITERI DI LINKAGE (2/3)



- **Average linkage:** la distanza tra i cluster A e B è la distanza media tra un'osservazione appartenente al cluster A e un'osservazione appartenente al cluster B.
- **Centroid linkage:** la distanza tra i cluster A e B è la distanza tra i centroidi del cluster A e B.





CRITERI DI LINKAGE (3/3)



➤ **Ward's linkage:** criterio che cerca di minimizzare la varianza intra-cluster → tende a produrre cluster più omogenei.

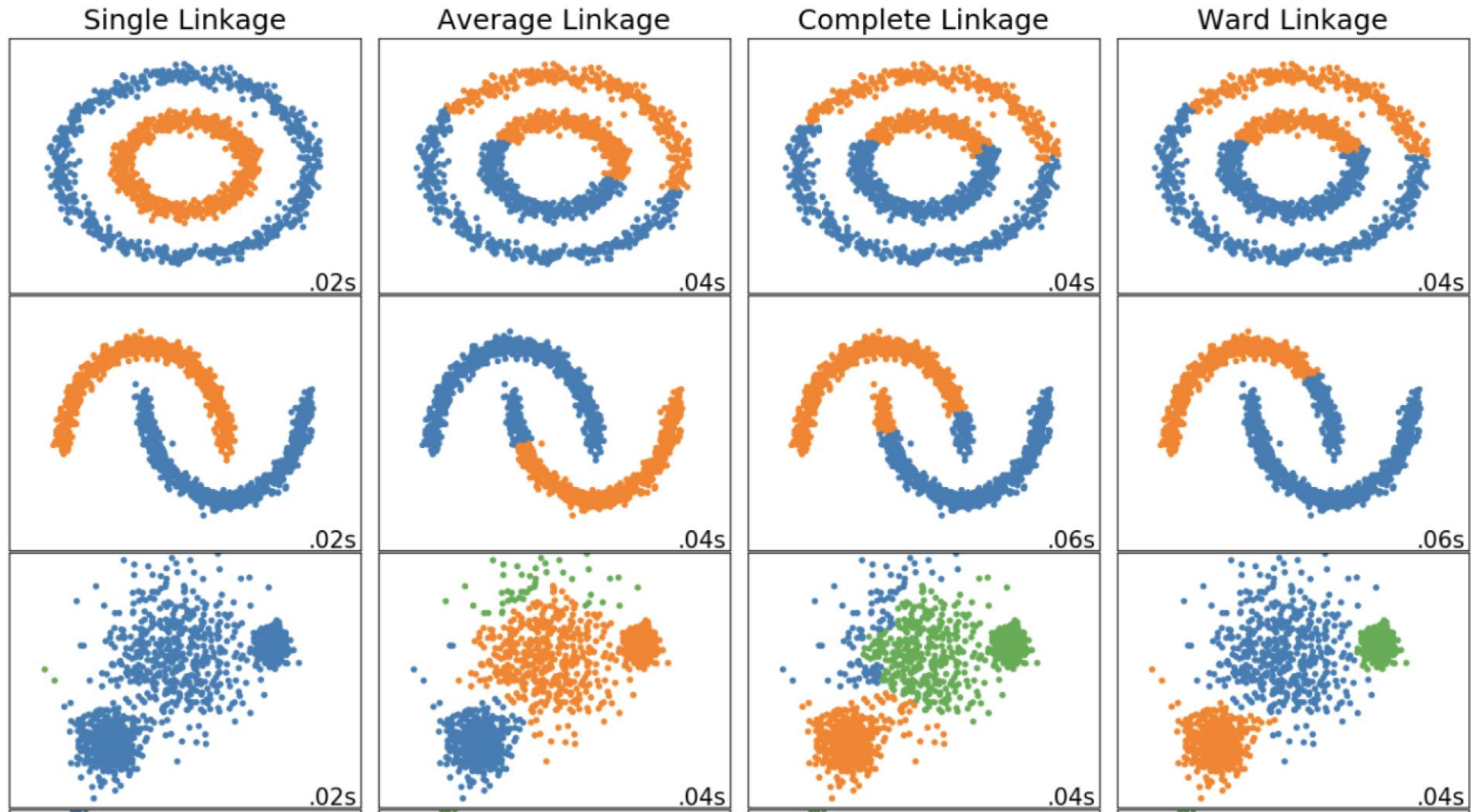
Dati due cluster A e B, di dimensioni n_A e n_B , aventi centroidi μ_A e μ_B , la distanza tra i due cluster è definita come:

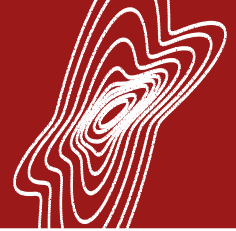
$$d(A, B) = \frac{n_A n_B}{n_A + n_B} \cdot \|\mu_A - \mu_B\|^2$$

Il valore di questa distanza rappresenta di quanto aumenta la somma delle varianze intra-cluster facendo il merge dei cluster A e B.



INFLUENZA DEL CRITERIO DI LINKAGE SULLE PARTIZIONI

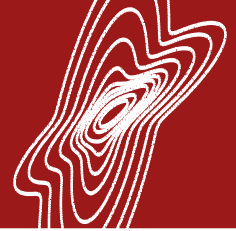




SCELTA DEL METODO DI LINKAGE



- Per scegliere il metodo di linkage più appropriato per descrivere i dati possiamo utilizzare l'indice cophenetic.
- **Indice cophenetic, c :** quantifica quanto fedelmente il dendrogramma preserva l'informazione sulle distanze tra coppie di osservazioni nel dataset.
 - $-1 \leq c \leq 1$
 - c vicino a 1 se osservazioni vicine (distanza bassa) vengono unite ad altezze basse nel dendrogramma.
- Possiamo provare diversi metodi di linkage e scegliere quello che massimizza l'indice cophenetic.



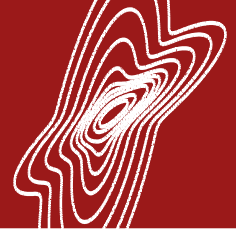
DEFINIZIONE DELL'INDICE COPENETICO



BONUS

$$c = \frac{\sum_{i < j} (d_{ij} - \bar{d})(t_{ij} - \bar{t})}{\sqrt{\sum_{i < j} (d_{ij} - \bar{d})^2 \sum_{i < j} (t_{ij} - \bar{t})^2}}$$

- d_{ij} : distanza tra le osservazioni x_i e x_j (es. distanza euclidea)
- t_{ij} : distanza copenetica tra le osservazioni x_i e $x_j \rightarrow$ altezza del punto nel dendrogramma in cui le due osservazioni vengono messe per la prima volta nello stesso cluster
- \bar{d} : media delle distanze d_{ij}
- \bar{t} : media delle distanze t_{ij}



NOTE SULL'INDICE COPENETICO

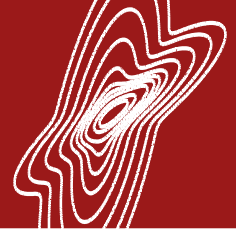


➤ Limiti dell'indice copenetico:

- Non valuta la qualità dei cluster finali (ma solo la fedeltà del dendrogramma alle distanze tra le osservazioni nel dataset).
- E' sensibile alla misura di distanza tra le osservazioni scelta.

➤ L'indice copenetico tende ad assumere valori elevati

- Solo valori molto alti (>0.9) possono essere considerati indice di un dendrogramma che effettivamente rappresenta fedelmente le distanze tra le osservazioni originali.
- Valori inferiori a 0.70 indicano che il dendrogramma ottenuto con clustering gerarchico non rappresenta bene la matrice delle distanze (forse i dati non sono adatti ad essere raggruppati con approccio gerarchico).



CLUSTERING GERARCHICO: PRO E CONTRO



➤ Pro:

- Non richiede di definire a priori il numero di cluster
- Genera una rappresentazione grafica dei raggruppamenti

➤ Contro:

- Richiede maggiore tempo computazionale
- Non tutti i problemi si prestano ad dei raggruppamenti di tipo gerarchico
- Il criterio di linkage scelto può influire in maniera drastica sui risultati