UNIVERSITÀ
DEGLI STUDI
DI PADOVA

METODI STATISTICI PER LA BIOINGEGNERIA

A.A. 2024-2025

Prof. Alessandra Bertoldo
Ing. Mattia De Francisci, Ing. Claudia Tarricone

Dati a disposizione e obiettivo del Laboratorio:

- I dati a disposizione per il laboratorio sono stati raccolti per uno studio dell'andamento del battito cardiaco tramite segnale ECG ottenuto con dispositivi indossabili.

- L'obiettivo è trovare il numero di gruppi (o cluster) ottimo per suddividere i soggetti che abbiamo a disposizione.

- I dati sono contenuti nel file ***data_lab8.mat***, *sono presenti:*

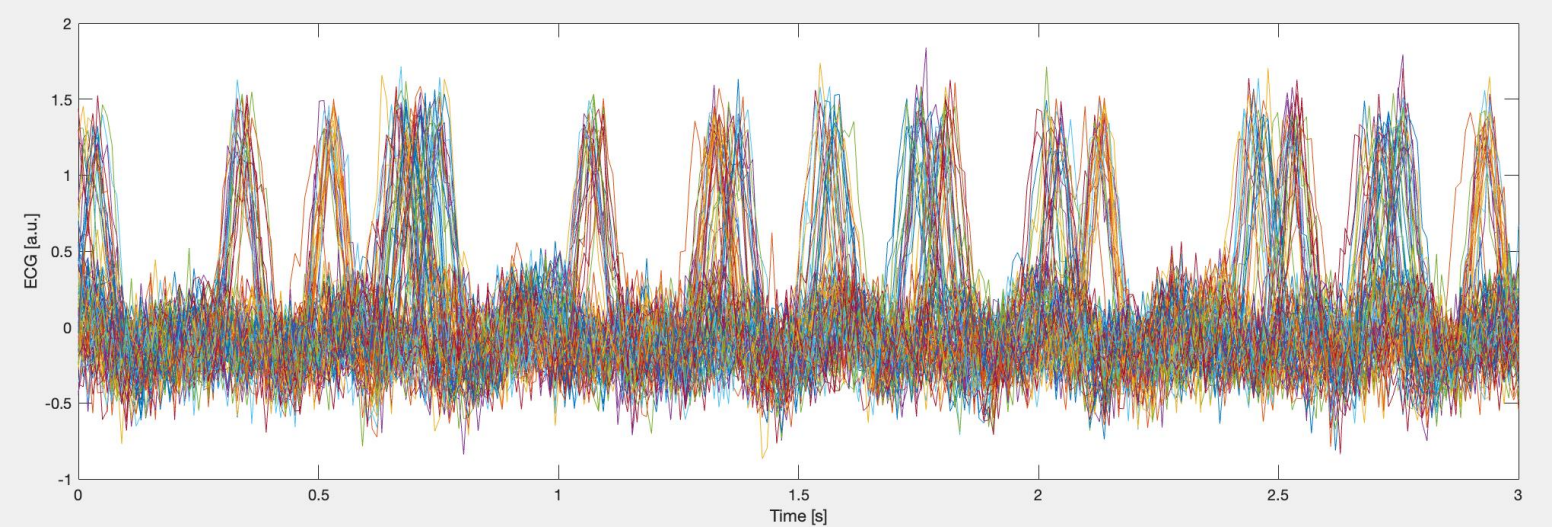
  ecg: Matrice 2D di dimensioni 100x300 (100 soggetti, 300 campioni)
  time: Vettore dei tempi in secondi (300x1)

Caricamento dei dati e osservazione

- Caricare i dati (***data_lab8.mat***).

  NOTA: non viene chiesta l'analisi per la ricerca di valori non utilizzabili (NaN,inf) perchè sono già stati testati e ripuliti. Ai dati è stata tolta la media quindi i valori negativi che vedete non sono da considerare non fisiologici. I dati sono quindi pronti all'uso.

- *Costruire un plot con i segnali a disposizione e osservarli: cosa notiamo?*



CHALLENGE

Obiettivo: trovare il raggruppamento ottimo dei tracciati ecg in K clusters, servendosi delle funzioni ***`kmeans`*** e ***`silhouette`***.

Per l'algoritmo K-means (***`kmeans`***), impostare i parametri come segue:

- *Numero di ripetizioni (**`replicates`**) del clustering, con diversi punti di partenza (casuali) per i centroidi, pari a 10*
- *Massimo numero di iterazioni (**`maxiter`**) pari a 200*
- *Distanza (**`distance`**): euclidea (**`sqeuclidean`**)*

Testare K = [2,…,10]. Per ciascuna configurazione (ciascun k), si calcoli il valore di silhouette per gli elementi del dataset (**`silhouette`**) e un valore di silhouette media

- In una figura si rappresenti il plot del valore della silhouette media in funzione di k
- In base al valore della silhouette media si identifichi il numero di clusters ottimo Kopt.

Qual è il numero ottimo di cluster?

Salvataggio dei risultati e calcolo delle matrici di similarità

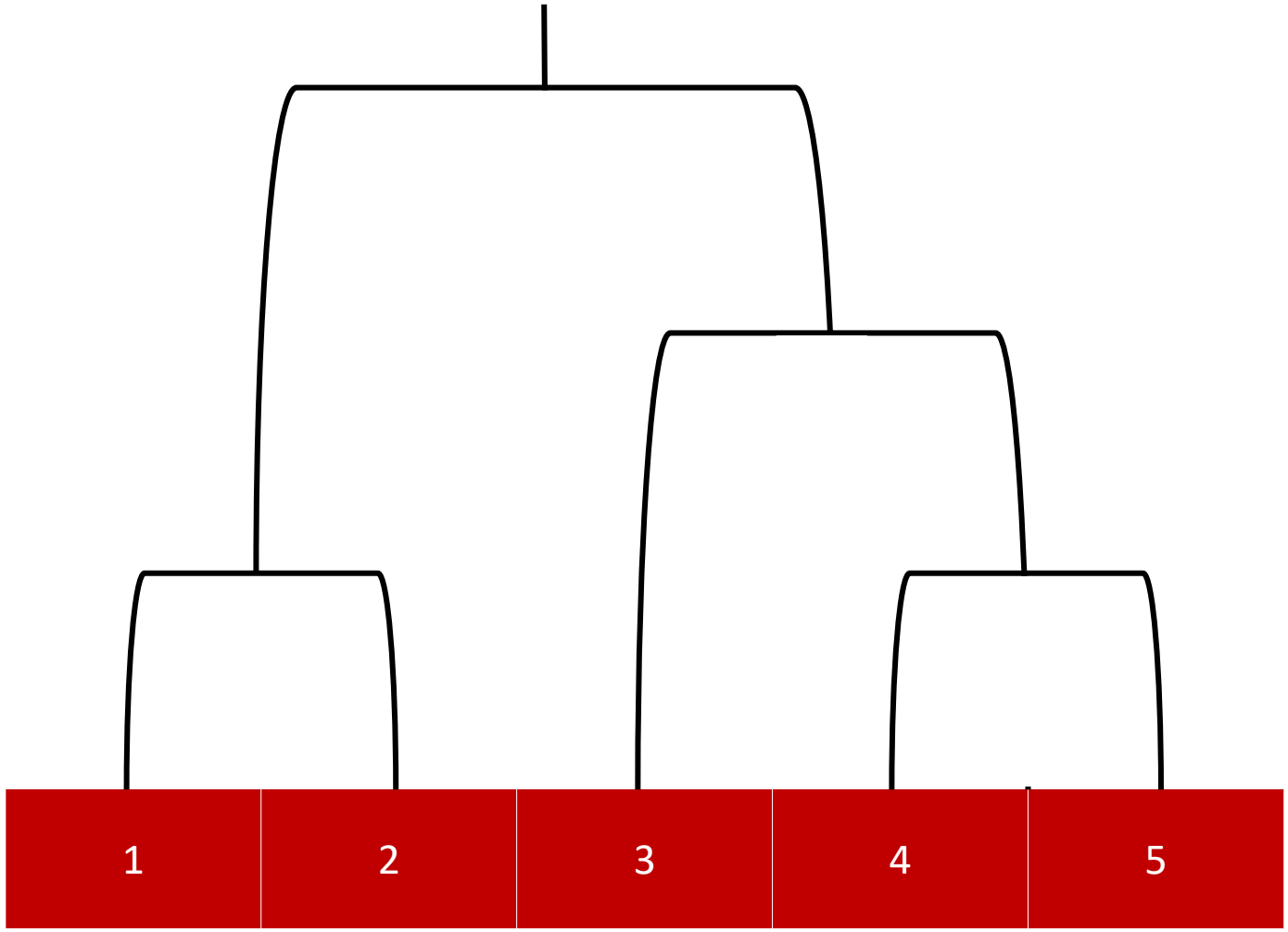
- Una volta scelto il numero ottimale di cluster (***Kopt***), si salvino nel vettore ***CLUST_kmeans*** le assegnazioni ai vari clusters ottenute e nella matrice ***Centroids_kmeans*** i centroidi relativi ai ***Kopt*** clusters ottenuti;
- Si rappresenti graficamente (vedere l'help della function ***silhouette***) i valori di silhouette ottenuti assumendo un numero di cluster pari a ***Kopt***: cosa si può osservare?
- Calcolare e raffigurare la matrice di similarità (***imagesc***) riportante le inverse delle distanze euclidee tra gli elementi. Servirsi delle funzioni ***pdist*** e ***squareform*** (consultare l'***help***).

OSSERVAZIONE: prima di procedere al calcolo delle distanze tramite pdist, riordinare le osservazioni sulla base del cluster di appartenenza.

Clustering gerarchico: come viene generata la variabile *tree*?

```
tree = linkage(x,method);

figure( )
cutoff = median([tree(end-(Kopt-1),3) ...
tree(end-(Kopt-2),3)]);
dendrogram(tree,'ColorThreshold',cutoff)
title('Hierarchical clustering tree')
```



tree =

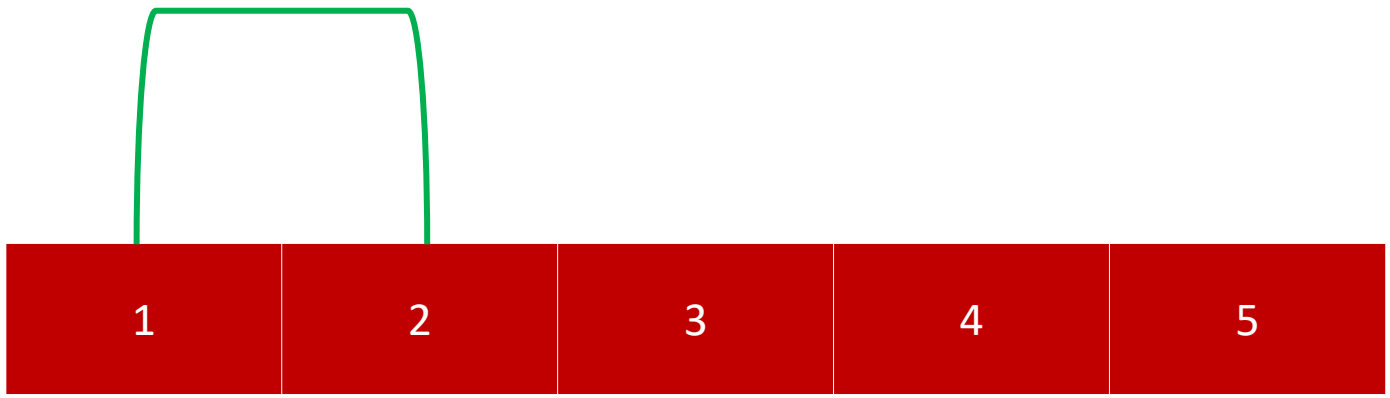
$$\begin{pmatrix} 1 & 2 & d_{1,2} \end{pmatrix}$$

size(tree) = (n-1) x 3

n = 5

1° link

n = 7

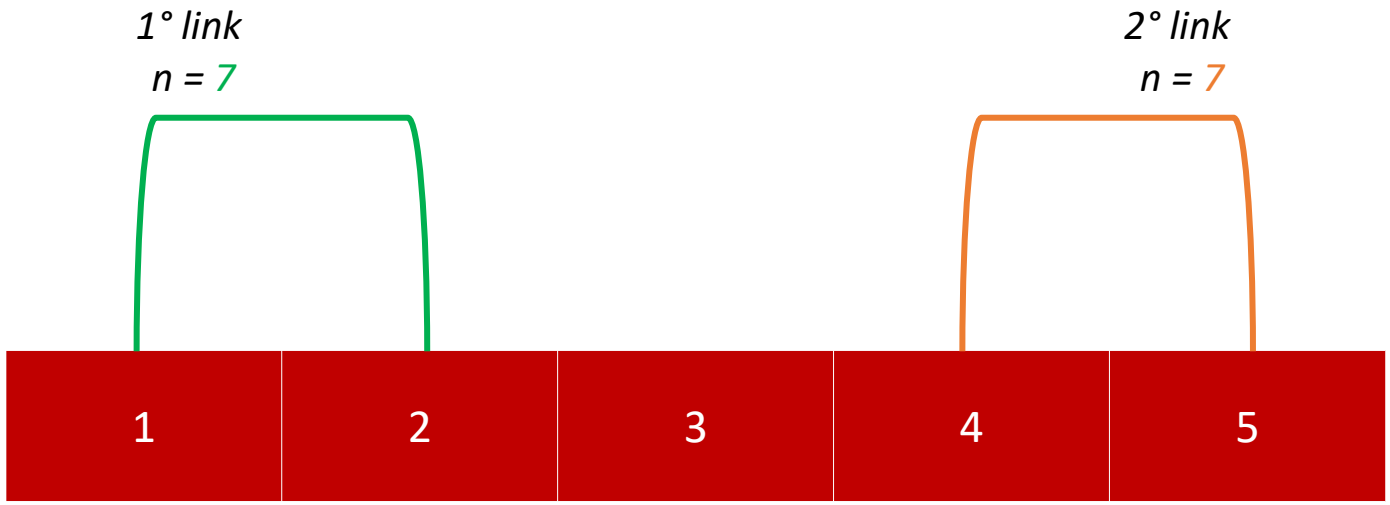


tree =

$$\begin{pmatrix} 1 & 2 & d_{1,2} \\ 4 & 5 & d_{4,5} \end{pmatrix}$$

size(tree) = (n-1) x 3

n = 5

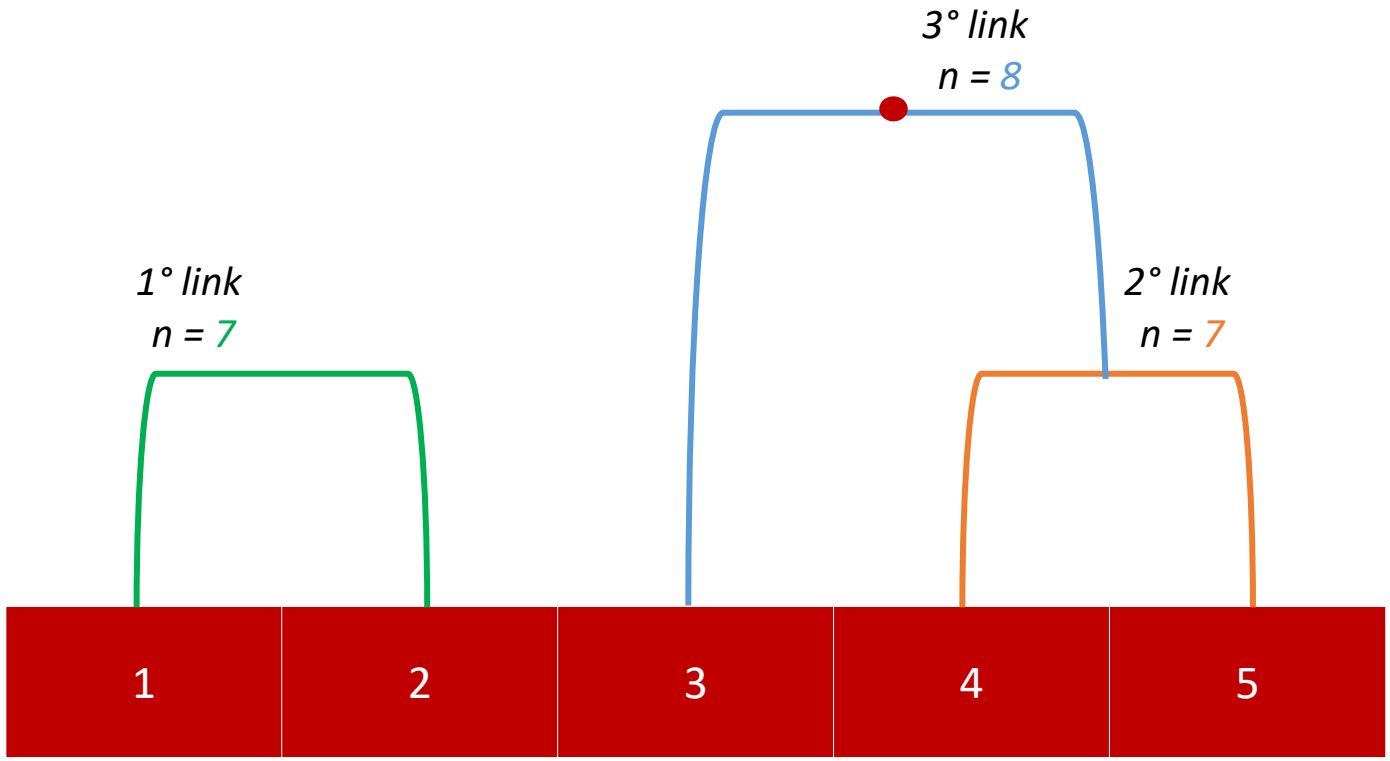


tree =

$$\begin{pmatrix} 1 & 2 & d_{1,2} \\ 4 & 5 & d_{4,5} \\ 3 & 7 & d_{3,7} \end{pmatrix}$$

size(tree) = (n-1) x 3

n = 5

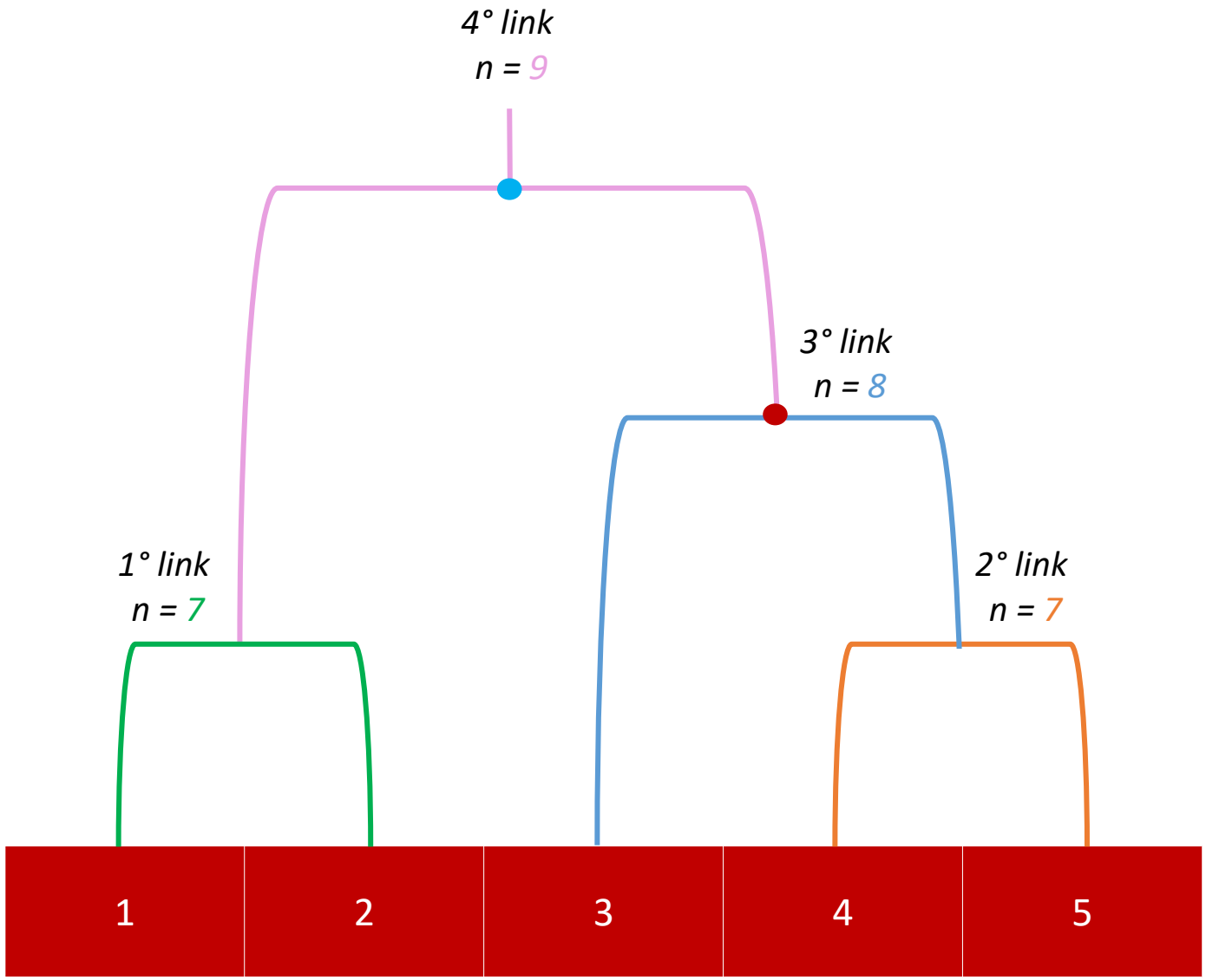


n = 5

tree =

$$\begin{pmatrix} 1 & 2 & d_{1,2} \\ 4 & 5 & d_{4,5} \\ 3 & 7 & d_{3,7} \\ 6 & 8 & d_{6,8} \end{pmatrix}$$

size(tree) = (n-1) x 3



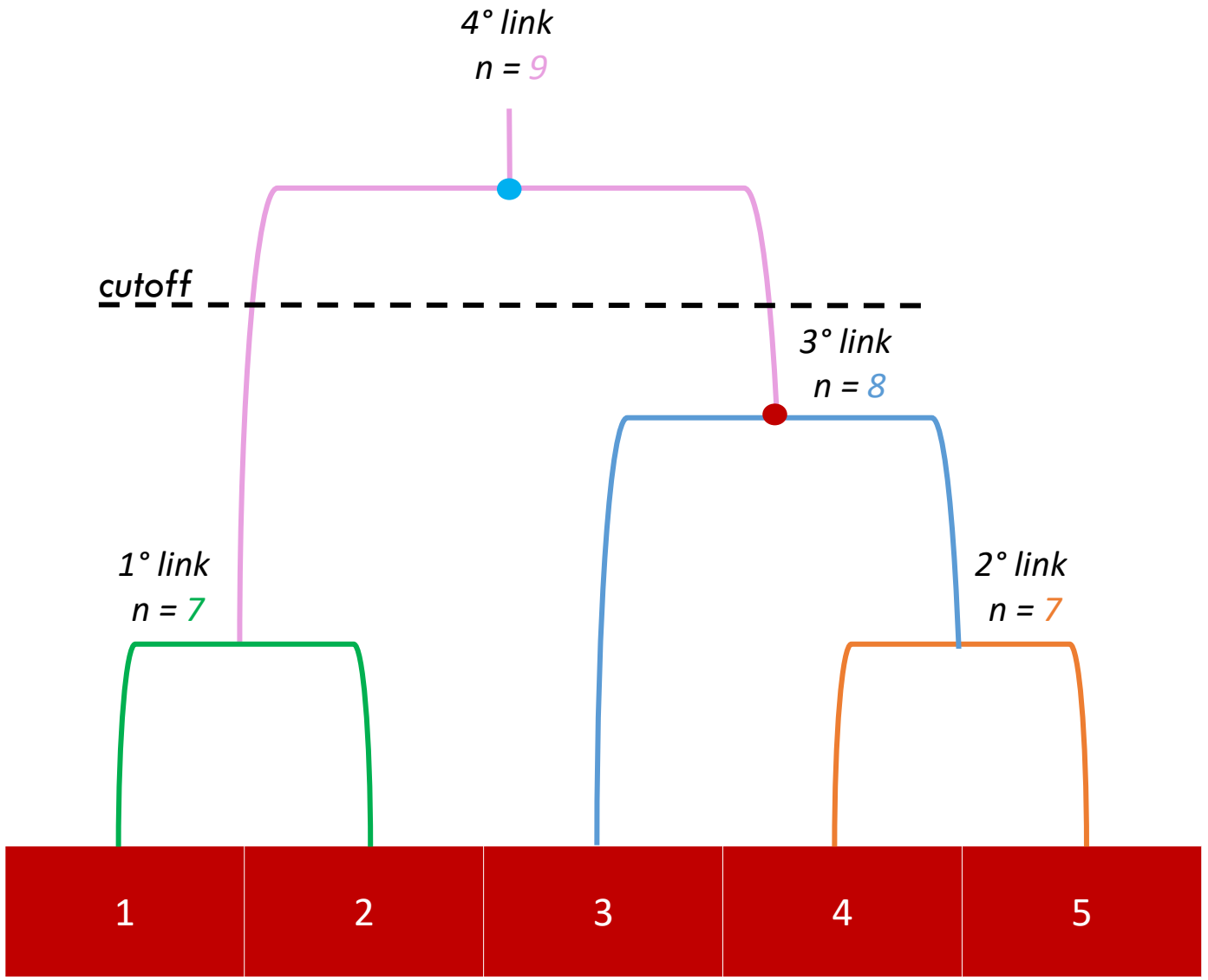
n = 5

tree =

1	2	$d_{1,2}$
4	5	$d_{4,5}$
3	7	$d_{3,7}$
6	8	$d_{6,8}$

size(tree) = (n-1) x 3

HP: Kopt = 2
cutoff = median([tree(end-(Kopt-1),3) …
tree(end-(Kopt-2),3)]);



Applicazione del Clustering Gerarchico

Si ripeta la classificazione dei soggetti applicando un clustering gerarchico:

- Si rappresenti in una figura l'albero gerarchico (si usi la function **`linkage`** con metrica *Ward* per il calcolo dell'albero e si utilizzi **`dendrogram`** per il plot): cosa si osserva dal dendrogramma?

- Si utilizzi un cutoff determinato da un numero di cluster pari al Kopt precedentemente ottenuto col k-means e la silhouette.

- Si ripetano i punti precedenti con la metrica *Complete*. Che cosa si può osservare confrontando i due alberi?

- Si verifichino dissimilarità (function **`cophenet`**) degli alberi gerarchici ottenuti con le due metriche. Qual è la migliore?
- A partire dall'albero con indice cofenetico migliore , si ricavi la classificazione dei soggetti scegliendo un numero di clusters pari al Kopt identificato con k-means (**`cluster`**). Si salvi l'assegnazione ottenuta nel vettore **`CLUST_tree.`**

- Si calcolino i centroidi relativi ai **`Kopt`** cluster ottenuti e si salvino i centroidi nella matrice **`Centroids_tree`**.

Confronto dei risultati ottenuti dall'applicazione dei due metodi di clustering:

- Si calcoli la correlazione tra centroidi rispettivi ottenuti con i due algoritmi di clustering (`corr`): le correlazioni sono statisticamente significative?

- Si calcoli la matrice delle distanze tra i centroidi (`pdist2`) e tramite ispezione visiva (`subplot`) si confrontino i centroidi stessi, individuando le coppie di cluster simili estratte dai due algoritmi di clustering. Verificare la somiglianza tra le coppie di centroidi plottandoli in sovrapposizione.

- [BONUS] automatizzare l'identificazione delle coppie di centroidi simili ottenute dai due algoritmi di clustering