3 Luglio 2019 - Analisi Esplorativa

Cognome: .....

Nome: .....

Matricola: .....

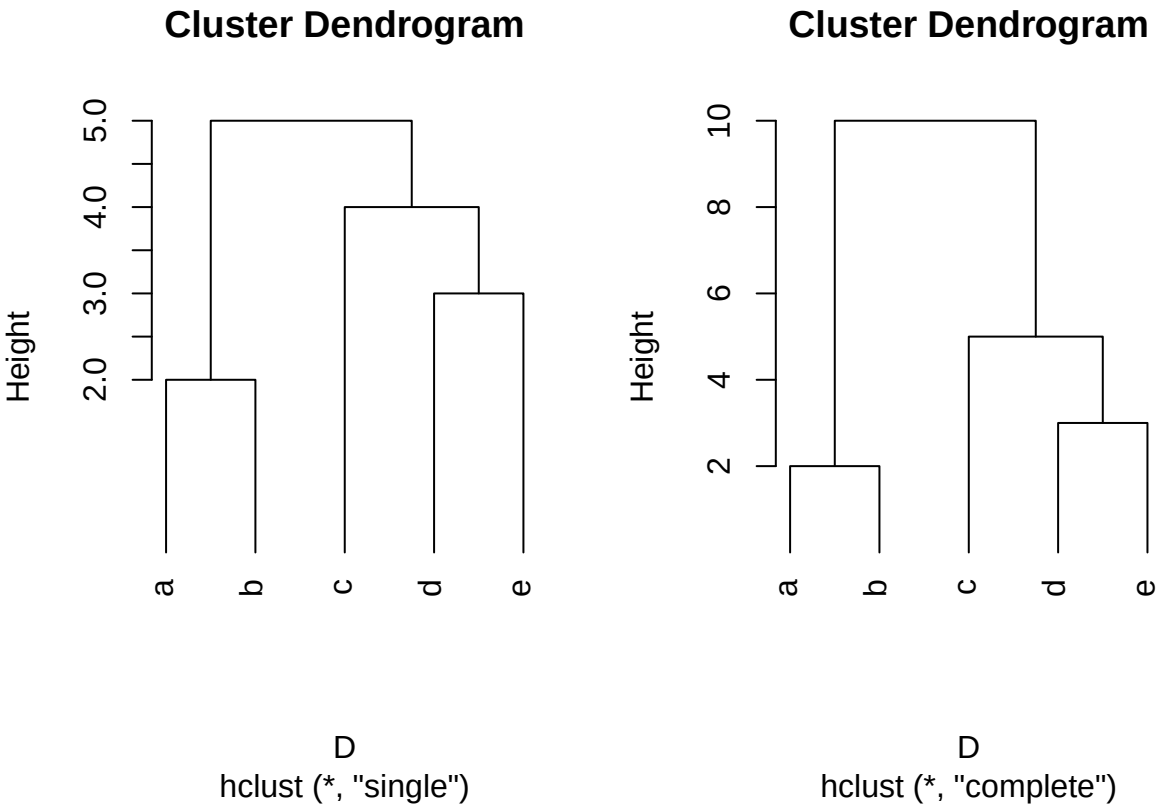
Tipologia d'esame: ☐ 12 CFU ☐ 15 CFU

Prova scritta - fila A

Si svolgano gli esercizi riportando il risultato dove indicato. Durata: 60 minuti

---

Esercizio 1 (Punti 3)



Sulla base dei due dendogrammi sopra riportati, completare la seguente matrice di distanze che gli ha generati.

	a	b	c	d	e
a	0				
b	...	0			
c	6	5	0		
d	...	9	4	0	
e	9	8	5	...	0

Esercizio 2 (Punti 3)

Quali dei seguenti vettori sono ortogonali?

$$x = \begin{pmatrix} 1 \\ -2 \\ 3 \\ -4 \end{pmatrix} \quad y = \begin{pmatrix} 6 \\ 7 \\ 1 \\ -2 \end{pmatrix} \quad z = \begin{pmatrix} 5 \\ -4 \\ 5 \\ 7 \end{pmatrix}$$

Riportare la versione normalizzata dei due vettori ortogonali (arrotondare al secondo decimale).

Esercizio 3 (Punti 9)

Si consideri il dataset *USArrests* presente nella libreria *datasets*. Per ciascuno dei 50 stati degli USA, l'insieme di dati contiene il numero di arresti per 100000 residenti per ognuno dei tre reati: Rapina (*Assault*), Omicidio (*Murder*) e Stupro (*Rape*). La variabile *UrbanPop* indica la percentuale di popolazione nelle aree urbane. Sia $\underset{50\times4}{X}$ la matrice dei dati corrispondente al dataset *USArrests*.

a. Si calcoli $d^2_M(x_i, \bar{x})$, il quadrato della distanza di Mahalanobis di ciascuna osservazione (ciascuna riga della matrice X) dal baricentro. Si riportino i valori di $d^2_M(x_i, \bar{x})$ solo se superano il valore 8, specificando anche il nome della riga di X (lo stato) a cui si fa riferimento.

b. Sulla base della matrice dei dati standardizzati $\underset{50\times4}{Z}$, applicare l'algoritmo delle K medie (`algorithm = "Hartigan-Wong"`) inizializzando i K centrodi utilizzando le prime K osservazioni (righe $1, \ldots, K$ della matrice Z). Arrotondando il risultato alla seconda cifra decimale, riportare per $K = 2, 3, \ldots, 7$
   - il valore dell'indice $\text{CH}(K) = \frac{B/(K-1)}{W/(n-K)}$ di Calinski and Harabasz
   - il valore medio della *silhouette* considerando come matrice delle distanze quella ottenuta con la metrica Euclidea basata su Z

K	2	3	4	5	6	7
$CH(K)$						
silhouette(K)						

c. Determinare l'appartenenza di ciascuna osservazione a $K = 2$ gruppi utilizzando

- il metodo gerarchico agglomerativo con funzione di legame completo considerando come matrice delle distanze quella ottenuta con la metrica di Manhattan basata su Z;
- il metodo delle delle K medie (`algorithm = "Hartigan-Wong"` inizializzando i K centrodi utilizzando le prime K osservazioni) applicato alla matrice dei punteggi (*scores*) $\underset{50\times2}{Y}$ ottenuta dalle prime due componenti principali di Z.

Riportare il numero delle osservazioni classificate nei cluster 1 e 2 secondo i due approcci (gerarchico e K-medie)

Approccio	n.ro osservazioni cluster 1	n.ro osservazioni cluster 2
Gerarchico	...	...
K-medie	...	...

Riportare i valori della tabella a doppia entrata che incrocia la classificazione ottenuta con l'approccio gerarchico e quello delle K-medie

Gerarchico / K-medie	cluster 1	cluster 2
cluster 1	...	...
cluster 2	...	...

Riportare il nome dell'unico stato classificato nel cluster 2 da entrambi gli approcci.

d. Stimare il modello fattoriale con $k = 1$ fattori con il metodo della massima verosimiglianza utilizzando i dati standardizzati Z e senza effettuare alcuna rotazione. Riportare il valore della statistica test rapporto di verosimiglianza $T = n \log\left(\frac{det(\hat{\Lambda}\hat{\Lambda}' + \hat{\Psi})}{det(R)}\right)$ (arrotondando al terzo decimale)

Esercizio 4 (Punti 3)

Si consideri il modello fattoriale con 1 fattore:

$$z_1 = \lambda_1 f + u_1$$
$$z_2 = \lambda_2 f + u_2$$
$$z_3 = \lambda_3 f + u_3$$

dove $\widehat{\mathbb{C}ov}(z) = \underset{3\times3}{R} = \begin{bmatrix} 1 & 0.25 & 0.25 \\ & 1 & 0.25 \\ & & 1 \end{bmatrix}$.

Riportare le stime $\hat{\Lambda}$ e $\hat{\Psi}$ utilizzando il metodo di stima *naive*.

Esercizio 5 (Punti 3)

Sulla base la matrice dei dati centrati $\underset{n\times 3}{\tilde{X}} = [\underset{n\times 1}{\tilde{x}_1}\ \underset{n\times 1}{\tilde{x}_2}\ \underset{n\times 1}{\tilde{x}_3}]$ è stata calcolata la seguente matrice di varianze/covarianze:

$$\underset{3\times 3}{S} = \begin{bmatrix} 2 & 0 & 0 \\ & 3 & 0 \\ & & 4 \end{bmatrix}$$

Si determinino i punteggi delle 3 componente principali:

$\underset{n\times 1}{y_1} =$

$\underset{n\times 1}{y_2} =$

$\underset{n\times 1}{y_3} =$

Esercizio 6 (Punti 5)

Dimostrare, esplicitando tutti i passaggi e le quantità coinvolte, che

a. $\det(S^Y) = \det(S)$ dove $Y = \tilde{X}V$ e le colonne di V sono gli autovettori normalizzati di S
b. nel modello fattoriale a k fattori, $\mathbb{E}(\underset{p\times 1}{x}\ \underset{1\times k}{f'}) = \underset{p\times k}{\Lambda}$