7 Febbraio 2020 - Analisi Esplorativa (Analisi Statistica Multivariata)

Cognome: ....

Nome: ....

Matricola: ....

Prova scritta - fila A

Si svolgano gli esercizi riportando il risultato dove indicato. Durata: 90 minuti

---

Esercizio 1 (Punti 2)

Si dimostri che $\underset{n\times n}{J} = I - H$ è una matrice idempotente, dove $\underset{n\times n}{I}$ è la matrice identità e $\underset{n\times n}{H}$ è la matrice di centramento, giustificando tutti i passaggi.

Esercizio 2 (Punti 4)

Si consideri la seguente matrice $\underset{n\times p}{X} = \begin{bmatrix} 3 & 1 & 0 \\ 6 & 4 & 6 \\ 4 & 2 & 2 \\ 7 & 0 & 3 \\ 5 & 3 & 4 \end{bmatrix}$

a. (Punti 1) Calcolare la matrice dei dati centrati $\tilde{X}$

b. (Punti 2) Verificare che le colonne di $\tilde{X}$ sono linearmente dipendenti, determinando un vettore non-nullo $\underset{p\times 1}{c}$ tale che $\tilde{X}c = \underset{n\times 1}{0}$

c. (Punti 1) Calcolare la matrice di varianze e covarianze S e la varianza generalizzata

$\tilde{X} = \begin{bmatrix} \quad \end{bmatrix}$ $c = \begin{bmatrix} \quad \end{bmatrix}$ $S = \begin{bmatrix} \quad \end{bmatrix}$ $det(S) =$

Esercizio 3 (Punti 4)

Alla matrice di varianze/covarianze $\underset{p\times p}{S}$ sono associati i seguenti autovalori $\lambda_1 = 6, \lambda_2 = 4$ e autovettori normalizzati $v_1 = \begin{bmatrix} 1/\sqrt{5} \\ 2/\sqrt{5} \end{bmatrix}$, $v_2 = \begin{bmatrix} 2/\sqrt{5} \\ -1/\sqrt{5} \end{bmatrix}$.

a. (Punti 2) Riportare la matrice di correlazione $\underset{p\times p}{R}$ e la matrice $\underset{p\times p}{S^{2/3}}$ (arrotondando i risultati al secondo decimale):

$$R = \begin{bmatrix} \quad & \quad \\ \quad & \quad \end{bmatrix} \qquad S^{2/3} = \begin{bmatrix} \quad & \quad \\ \quad & \quad \end{bmatrix}$$

b. (Punti 2) Calcolare la correlazione tra la prima colonna $\underset{n\times 1}{\tilde{x}_1}$ di $\underset{n\times p}{\tilde{X}}$ e i punteggi $\underset{n\times 1}{y_1}$ della prima componente principale, arrotondando il risultato al secondo decimale:

Esercizio 4 (Punti 3)

Si consideri la seguente matrice di distanza:

	1	2	3	4
1	0			
2	3	0		
3	7	9	0	
4	8	6	5	0

a. (Punti 2) Si utilizzi il metodo gerarchico agglomerativo con il legame completo, riportando ad ogni passo dell'algoritmo la matrice delle distanze tra gruppi

passo 1)

	......	......	......
......	0		
......	......	0	
......	......	......	0

passo 2)

	......	......
......	0	
......	......	0

b. (Punti 1) Riportare il dendogramma corrispondente al punto a.

Esercizio 5 (Punti 9)

Si considereri la seguente matrice di correlazione calcolata sulla base di $n = 50$ osservazioni:

	Murder	Assault	UrbanPop	Rape
Murder	1.0	0.8	0.1	0.6
Assault	0.8	1.0	0.3	0.7
UrbanPop	0.1	0.3	1.0	0.4
Rape	0.6	0.7	0.4	1.0

a. (Punti 1) Sulla base dalla matrice di correlazione, si stimi il modello fattoriale con $k = 1$ fattore utilizzando il metodo della massima verosimiglianza senza effettuare alcuna rotazione. Arrontondando al terzo decimale, si riportino le stime dei pesi fattoriali:

$\hat{\lambda}_1 = \ldots\ldots\ldots\ldots\ldots$ $\hat{\lambda}_2 = \ldots\ldots\ldots\ldots\ldots$ $\hat{\lambda}_3 = \ldots\ldots\ldots\ldots\ldots$ $\hat{\lambda}_4 = \ldots\ldots\ldots\ldots\ldots$

b. (Punti 2) Si determini il punteggio fattoriale con il metodo di Thompson (arrotondando alla quarta cifra decimale) per l'unità statistica "Arizona" sapendo che i suoi valori nelle quattro variabili standardizzate sono

	Murder	Assault	UrbanPop	Rape
Arizona	1.2426	0.7828	-0.5209	-0.0034

c. (Punti 2) Si riporti la stima delle comunalità per le quattro variabili, arrotondando al secondo decimale. Indicare quale variabile è spiegata meglio dal modello, motivando la risposta.

	Murder	Assault	UrbanPop	Rape
Comunalità				

Variabile spiegata meglio:

d. (Punti 1) Si valuti l'opportunità di stimare un modello a 2 fattori, motivando la risposta.

e. (Punti 3) Si calcolino le correlazioni tra la j-sima variabile standardizzata $\underset{n\times 1}{z_j}$ e la k-sima componente principale $\underset{n\times 1}{y_k}$ per $j = 1, \ldots, 4$ e $k = 1, \ldots, 4$, riportandole nella seguente tabella (arrotondando al secondo decimale).

	y_1	y_2	y_3	y_4
z_1				
z_2				
z_3				
z_4				

Potrebbe tornare utile sapere che le covarianze tra variabili standardizzate e componenti principali sono date dalla matrice

$$\frac{1}{n}Z'Y$$

dove $\underset{n\times p}{Y} = ZV$ è la matrice dei punteggi delle componenti principali e $R = V\Lambda V'$ è la matrice di correlazione.

Esercizio 6 (Punti 4)

Si consideri il dataset `swiss` presente nella libreria `datasets`, che contiene $n = 47$ unità statistiche (provincie) relative alle seguenti 6 variabili:

- *Fertility* : common standardized fertility measure
- *Agriculture* : % of males involved in agriculture as occupation
- *Examination* : % draftees receiving highest mark on army examination
- *Education* : % education beyond primary school for draftees
- *Catholic* : % catholic (as opposed to protestant)
- *Infant.Mortality* : live births who live less than 1 year

a. (Punti 2) Per ciascuna unità statistica, si calcoli la distanza di Mahalanobis dal baricentro, e si riportino i valori delle distanze superiori a 3.35 (arrotondando al terzo decimale) con i rispettivi nomi delle unità (provincie)

b. (Punti 2) Standardizzare i dati `swiss`, e utilizzare l'algoritmo delle K-medie (`algorithm = Hartigan-Wong`) per $K = 2, 4, 6$, inizializzando i centroidi con le osservazioni di riga $1, 2, \ldots, K$. Riportare per ciascun valore di K il rispettivo valore dell'indice Calinski and Harabasz (arrotondando al terzo decimale).

K	2	4	6
Indice CH	....	....	....