22 Febbraio 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: 80 minuti

---

Esercizio 1 (10 punti)

Il dataset `Cars93` presente nella libreria `MASS` descrive 93 modelli di automobile rispetto a 27 variabili. Le variabili di interesse per l'analisi sono $p = 12$ come elencato nel seguito:

- `Weight` peso (in libbre)
- `Width` larghezza (in pollici)
- `Length` lunghezza (in pollici)
- `EngineSize` cilindrata (in litri)
- `Horsepower` potenza
- `RPM` giri al minuto
- `Fuel.tank.capacity` capacità di carburante
- `Passengers` numero di passeggeri
- `MPG.highway` Miglia per gallone in autostrada
- `MPG.city` Miglia per gallone in città
- `Price` prezzo (in migliaia di dollari)
- `Wheelbase` interasse (in pollici)

Sia X la matrice 93×12 corrispondente al dataset `Cars93` ridotto selezionando le variabili sopra elencate.

a. Si calcoli $d^2_M(x_i, \bar{x})$, il quadrato della distanza di Mahalanobis di ciascuna auto (ciascuna riga della matrice X) dal baricentro. Si riportino i valori di $d^2_M(x_i, \bar{x})$ solo se superano il valore 21, specificando anche il nome dell'automobile a cui si fa riferimento (informazione reperibile dalla variabile `Model` del dataset `Cars93`).

b. Svolgere l'analisi delle componenti principali sui dati standardizzati Z che si ottengono a partire da X. Si riportino i punteggi (*scores*) per l'automobile `Firebird` relativamente alle componenti principali 2, 8 e 10 (arrotondati alla terza cifra decimale).

c. Applicare l'algoritmo delle K medie (`algorithm = "Hartigan-Wong"`) per i dati standardizzati Z 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 $K = 2, 3, 4, 5$

- 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 di Lagrange basata su Z

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

d. Si consideri la stima di massima verosimiglianza per il modello fattoriale con k fattori basato sui dati standardizzati Z. Riportare il p-value del primo test non significativo al livello 5% (e il corrispondente valore di k) per la sequenza di ipotesi nulle $H_0(k = 1), H_0(k = 2), H_0(k = 3), \ldots$ dove $H_0(k)$="il modello fattoriale con k fattori è corretto".

e. Stimare il modello fattoriale con $k = 2$ 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 $T = n \log\left(\frac{|\hat{\Lambda}\hat{\Lambda}' + \hat{\Psi}|}{|R|}\right)$ (arrotondando al terzo decimale)

Esercizio 2 (punti 2)

Dimostrare, esplicitando tutti i passaggi, e specificando tutte le quantità coinvolte.

Se esiste un $\underset{p\times 1}{c} \neq 0$ tale che $\underset{p\times p}{S}\underset{p\times 1}{c} = \underset{p\times 1}{0}$, allora le colonne di $\underset{n\times p}{\tilde{X}}$ sono linearmente dipendenti.

Esercizio 3 (3 punti)

Si consideri la seguente matrice di correlazione $\underset{3\times 3}{R} = \begin{bmatrix} 1 & 1/2 & 1/2 \\ 1/2 & 1 & 2/3 \\ 1/2 & 2/3 & 1 \end{bmatrix}$.

a. Sia $\underset{n\times 3}{\tilde{X}}$ la matrice dei dati centrati. Determinare l'angolo (espresso in gradi) tra $\underset{n\times 1}{\tilde{x}_1}$ e $\underset{n\times 1}{\tilde{x}_3}$: = ...

b. Sapendo che $s_{11} = 4$, $s_{22} = 9$ e $\text{tr}(S) = 14$, calcolare la matrice di varianze/covarianze S e $S^{1/2}$ (arrotondando al secondo decimale)

$$S = \begin{bmatrix} \dots\dots & \dots\dots & \dots\dots \\ \dots\dots & \dots\dots & \dots\dots \\ \dots\dots & \dots\dots & \dots\dots \end{bmatrix} \qquad S^{1/2} = \begin{bmatrix} \dots\dots & \dots\dots & \dots\dots \\ \dots\dots & \dots\dots & \dots\dots \\ \dots\dots & \dots\dots & \dots\dots \end{bmatrix}.$$

Esercizio 4 (2 punti)

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

	a	b	c	d
a	1	0.8	0.1	0.6
b	0.8	1	0.3	0.7
c	0.1	0.3	1	0.4
d	0.6	0.7	0.4	1

Sulla base dalla matrice di correlazione, si stimi il modello fattoriale con $k = 1$ fattori utilizzando il metodo della massima verosimiglianza senza effettuare alcuna rotazione. Si determini il punteggio fattoriale con il metodo di Thompson (arrotondando alla quarta cifra decimale) per una certa unità statistica sapendo che i suoi valori standardizzati nelle quattro variabili a, b, c e d sono

	a	b	c	d
unità	1.2426	0.7828	-0.5209	-0.0034

$\hat{f} =$

Esercizio 5 (3 punti)

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

a. 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 al secondo decimale

$$= \dots\dots\dots$$

b. Determinare (arrotondando al secondo decimale) la matrice di correlazione

$$\underset{p\times p}{R} = \Bigg[$$

Esercizio 6 (6 punti)

Dimostrare, esplicitando tutti i passaggi, e specificando tutte le quantità coinvolte.

a. Nel modello fattoriale a k fattori, $\mathbb{E}(\underset{p\times 1}{x}\underset{1\times k}{f'}) = \underset{p\times k}{\Lambda}$.

b. Una generica matrice di varianze/covarianze S è semidefinita positiva.

c. $\det(S^Y) = \det(S)$ dove $Y = \tilde{X}V$ e le colonne di V sono gli autovettori normalizzati di S