

17 Aprile 2019 - Analisi Esplorativa

Cognome:

Nome:

Matricola:

Tipologia d'esame: ☐ 12 CFU ☐ 15 CFU

Prova scritta - fila A

Si svolgono gli esercizi riportando il risultato dove indicato. Durata: 70 minuti

Esercizio 1 (10 punti)

Il dataset **Boston** presente nella libreria **MASS** è composto da 506 osservazioni e 14 variabili: *crim, zn, indus, chas, nox, rm, age, dis, rad, tax, ptratio, black, lstat, medv*.

Le variabili di interesse per l'analisi sono trasformazioni delle variabili originali come elencato nel seguito:

- $x_1 = \log(\text{crim})$
- $x_2 = \text{zn}/10$
- $x_3 = \log(\text{indus})$
- $x_4 = \log(\text{nox})$
- $x_5 = \log(\text{rm})$
- $x_6 = (\text{age})^{2.5}/10000$
- $x_7 = \log(\text{dis})$
- $x_8 = \log(\text{rad})$
- $x_9 = \log(\text{tax})$
- $x_{10} = \exp(0.4 \cdot \text{ptratio})/1000$
- $x_{11} = \text{black}/100$
- $x_{12} = \sqrt{\text{lstat}}$
- $x_{13} = \log(\text{medv})$

dove \log indica il logaritmo naturale.

Si costruisca la matrice dei dati $X_{506 \times 13}$ che contiene le variabili x_1, x_2, \dots, x_{13} .

- a. Si calcoli $d_M^2(x_i, \bar{x})$, il quadrato della distanza di Mahalanobis di ciascuna osservazione (ciascuna riga della matrice X) dal baricentro. Si riportino i valori di $d_M^2(x_i, \bar{x})$ solo se superano il valore 38, specificando anche l'indice della riga di X a cui si fa riferimento.

- b. Sulla base della matrice dei dati standardizzati $Z_{506 \times 13}$, applicare l'algoritmo delle K medie (**algorithm = "Hartigan-Wong"**) inizializzando i K centri utilizzando le prime K osservazioni (righe $1, \dots, K$ della matrice Z). Arrotondando il risultato alla seconda cifra decimale, riportare per $K = 5, 6, \dots, 10$
- 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	5	6	7	8	9	10
$CH(K)$						
$\text{silhouette}(K)$						

- c. 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 $H_0(k = 1), H_0(k = 2), H_0(k = 3), \dots$ dove $H_0(k)$ ="il modello fattoriale con k fattori è corretto".

- d. Stimare il modello fattoriale con $k = 5$ 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)

- e. Stimare il modello fattoriale con $k = 7$ fattori con il metodo della massima verosimiglianza utilizzando i dati standardizzati Z e senza effettuare alcuna rotazione. Riportare i punteggi fattoriali \hat{f}_i con il metodo di Thompson (arrotondando alla seconda cifra decimale) per l'unità statistica corrispondente alla riga $i = 100$ della matrice Z .

Esercizio 2 (6 punti)

Sia

$$R = \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}$$

con $0 < r < 1$.

- a. Determinare gli autovalori di R :

$$\lambda_1 = \dots\dots\dots \quad \lambda_2 = \dots\dots\dots$$

- b. Determinare gli autovettori normalizzati di R :

$$v_1 = \begin{bmatrix} \dots\dots\dots \\ \dots\dots\dots \end{bmatrix}, \quad v_2 = \begin{bmatrix} \dots\dots\dots \\ \dots\dots\dots \end{bmatrix}$$

- c. Determinare i punteggi delle componenti principali

$$y_{i1} = \dots\dots\dots, \quad y_{i2} = \dots\dots\dots, \quad i = 1, \dots, n$$

Esercizio 3 (4 punti)

Dimostrare, esplicitando tutti i passaggi e le quantità coinvolte, che il vettore $\underset{p \times 1}{u}$ di lunghezza unitaria che risolve il problema di massimo

$$\max_{u: \|u\|=1} u' S u$$

è l'autovettore $\underset{p \times 1}{v_1}$ (con segno positivo o negativo) della matrice di varianze/covarianze $\underset{p \times p}{S}$.

Esercizio 4 (3 punti)

Si consideri la seguente matrice di varianze/covarianze $S_{3 \times 3} = \begin{bmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{bmatrix}$

dove $a > 0$, $b > 0$ e $c > 0$ sono costanti non note.

a. Calcolare la varianza totale

=

b. Calcolare la varianza generalizzata

=

c. Calcolare l'indice di variabilità relativo

=

d. Determinare l'inversa della matrice di correlazione

$$R_{3 \times 3}^{-1} = \begin{bmatrix} \dots & \dots & \dots \\ \dots & \dots & \dots \\ \dots & \dots & \dots \end{bmatrix}.$$

Esercizio 5 (3 punti)

Enunciare il teorema di Eckart-Young.