

Cognome: Nome: Matricola:

Prova scritta di ASM - Modulo Analisi Esplorativa del 14.02.2017

La durata della prova è di 90 minuti.

Si svolgano gli esercizi A e B riportando il risultato dove indicato.

Esercizio A (Punti: 14)

1. Decomposizione Spettrale e Analisi delle Componenti Principali

Alla matrice $X_{n \times p}$ sono associati i seguenti autovalori e autovettori normalizzati: $\lambda_1 = 9$, $\lambda_2 = 6$, $v_1 =$

$$\begin{bmatrix} 1/\sqrt{5} \\ 2/\sqrt{5} \end{bmatrix} \text{ e } v_2 = \begin{bmatrix} 2/\sqrt{5} \\ -1/\sqrt{5} \end{bmatrix}.$$

a. Determinare la matrice di varianze/covarianze $S_{p \times p} = \begin{bmatrix} \dots & \dots \\ \dots & \dots \end{bmatrix}$

b. Riportare

- varianza totale = e generalizzata =
- l'indice di variabilità relativo (arrotondare al secondo decimale) =

c. Determinare, arrotondando al secondo decimale, $S_{p \times p}^{1/2} = \begin{bmatrix} \dots & \dots \\ \dots & \dots \end{bmatrix}$

d. Calcolare la correlazione tra la seconda colonna \tilde{x}_2 di $\tilde{X}_{n \times p}$ e i punteggi y_2 della seconda componente principale, arrotondando al secondo decimale.

=

2. Distanze e Cluster Analysis

a. Per una generica matrice di dati $X_{n \times p}$, si riporti la definizione di distanza di Minkowski $d_m(u_i, u_l)$ di ordine $m \geq 1$ tra due unità statistiche u'_i e u'_l .

$1 \times p \quad 1 \times p$

$$d_m(u_i, u_l) =$$

b. Per una generica matrice di distanze $D_{n \times n}$ con elemento di posizione (i, l) pari a $d(u_i, u_l)$, si riporti la definizione di legame medio tra due gruppi G_1 e G_2 .

$$d(G_1, G_2) =$$

c. Si calcoli il valore dell'indice di similarità di Jaccard per il seguente esempio:

$$\begin{bmatrix} 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} u'_1 \\ u'_2 \end{bmatrix}$$

$$s_J(u_1, u_2) =$$

3. Analisi Fattoriale

a. Si riportino le assunzioni del modello fattoriale con k fattori $x = \Lambda f + u$.

$$- \mathbb{E} \begin{pmatrix} x \\ p \times 1 \end{pmatrix} =$$

$$- \mathbb{E} \begin{pmatrix} f \\ k \times 1 \end{pmatrix} = \quad , \text{Cov} \begin{pmatrix} f \\ k \times 1 \end{pmatrix} =$$

$$- \mathbb{E} \begin{pmatrix} u \\ p \times 1 \end{pmatrix} = \quad , \text{Cov} \begin{pmatrix} u \\ p \times 1 \end{pmatrix} =$$

$$- \text{Cov} \begin{pmatrix} u \\ p \times 1 \end{pmatrix}, \begin{pmatrix} f \\ k \times 1 \end{pmatrix} =$$

b. La seguente tabella riporta la stima della matrice di pesi fattoriali $\hat{\Lambda}_{5 \times 2}$ di un modello fattoriale a due fattori ottenuta a partire dalla matrice di correlazione $R_{5 \times 5}$.

$$\hat{\Lambda}_{5 \times 2} = \begin{bmatrix} .56 & ? \\ .78 & -.53 \\ .65 & .75 \\ .94 & -.10 \\ ? & -.54 \end{bmatrix} \begin{matrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{matrix}$$

a. Sapendo che le varianze specifiche di x_1 e x_5 sono pari a $\hat{\psi}_1 = 0.02$ e $\hat{\psi}_5 = 0.07$, determinare, arrotondando al secondo decimale,

$$\hat{\lambda}_{12} = \dots\dots\dots, \quad \hat{\lambda}_{51} = \dots\dots\dots$$

4. Dimostrazione

Dimostrare che la matrice di centrimento $H_{n \times n}$ è idempotente, giustificano tutti i passaggi.

Esercizio B (Punti: 13)

Si consideri il dataset `iris` presente nella libreria `datasets` che contiene $n = 150$ unità statistiche (fiori di genere *Iris*) relative alle 4 variabili

- *Sepal.Length* (lunghezza dei sepali)
- *Sepal.Width* (larghezza dei sepali)
- *Petal.Length* (lunghezza dei petali)
- *Petal.Width* (larghezza dei petali)

più l'ultima colonna *Species* che specifica la specie (con modalità *setosa*, *versicolor* e *virginica*).

1. Si consideri la matrice $X_{150 \times 4}$ che contiene le seguenti variabili: *Sepal.Length*, *Sepal.Width*, *Petal.Length* e *Petal.Width*. Si calcoli il quadrato della distanza di Mahalanobis di ciascuna unità statistica u_i dal baricentro \bar{x} e si riporti il valore medio e il valore massimo, arrotondando i calcoli al secondo decimale.

$$\frac{1}{150} \sum_{i=1}^{150} d_M^2(u_i, \bar{x}) = \dots\dots\dots \quad \max_{i=1, \dots, 150} \{d_M^2(u_i, \bar{x})\} = \dots\dots\dots$$

2. Per la matrice di dati $X_{150 \times 4}$, utilizzare l'algoritmo delle K-medie (specificando `algorithm = "Lloyd"`) per formare $K = 3$ gruppi, inizializzando i centroidi con le osservazioni di riga 30, 80 e 110, ed eseguendo l'algoritmo una sola volta. Riportare
 - a. la numerosità dei 3 gruppi ottenuti;
 - b. i valori della tabella a doppia entrata che incrocia la classificazione ottenuta e la variabile *Species*;
 - c. il valore medio della silhouette (arrotondando al secondo decimale) per i tre gruppi (utilizzando il comando `silhouette` presente nella libreria `cluster`) considerando come matrice delle distanze quella ottenuta con la metrica Euclidea.

a. Numerosità gruppo 1 = , gruppo 2 = , gruppo 3 =

	setosa	versicolor	virginica
b.			
gruppo 1
gruppo 2
gruppo 3

c. Valore medio silhouette per il gruppo 1 = , gruppo 2 = , gruppo 3 =

3. Partendo dalla matrice $X_{150 \times 4}$ determinata al punto a., si calcoli la matrice dei dati standardizzati $Z_{150 \times 4}$. Si conduca l'analisi delle componenti principali basata su $Z_{150 \times 4}$ utilizzando il comando `prcomp()`, riportando (arrotondando alla seconda cifra decimale)
- la proporzione di varianza spiegata dalle prime due componenti principali
 - l'equazione del punteggio (*score*) y_{i2} della seconda componente principale per l' i -sima unità statistica
 - la correlazione tra i punteggi della prima componente principale e la prima colonna di Z

a. Proporzione di varianza spiegata dalle prime due componenti principali:

b. Punteggio y_{i2} della prima componente principale per l' i -sima unità statistica:

$$y_{i1} = \dots \cdot z_{i1} + \dots \cdot z_{i2} + \dots \cdot z_{i3} + \dots \cdot z_{i4}$$

c. Correlazione tra i punteggi della prima componente principale e la prima colonna di Z :