

Analisi Esplorativa (Analisi Statistica Multivariata)

Prova d'esame (simulazione)

25 Gennaio 2022

Tempo a disposizione: 150 minuti

Modalità di consegna: svolgere gli esercizi di teoria (parte A) riportando le soluzioni sul foglio protocollo, e consegnare il foglio protocollo assieme al testo della prova d'esame. Successivamente, accedere alla piattaforma examonline tramite computer e svolgere gli esercizi di analisi dei dati (parte B). In questo caso la consegna si svolge tramite piattaforma examonline. Il tempo da dedicare alla parte A e alla parte B è a discrezione dello studente.

Compilare con nome, cognome e numero di matricola. E' obbligatorio consegnare il testo della prova d'esame all'interno del foglio protocollo contenente le soluzioni degli esercizi di teoria.

NOME:

COGNOME:

MATRICOLA:

PARTE A: esercizi di teoria

Esercizi da svolgere sul foglio protocollo senza l'ausilio di R/Rstudio.

Problema 1

1. Trovare la matrice di varianza e covarianza per le variabili standardizzate z_1, z_2, z_3 corrispondente al seguente modello fattoriale

$$z_1 = 0.9f + u_1$$

$$z_2 = 0.7f + u_2$$

$$z_3 = 0.5f + u_3$$

dove $\text{Var}(f) = 1, \text{Cov}(u, f) = 0$ e

$$\Psi = \text{Cov}(u) = \begin{bmatrix} 0.19 & 0 & 0 \\ 0 & 0.51 & 0 \\ 0 & 0 & 0.75 \end{bmatrix}$$

Calcolare le comunalità e $\text{Corr}(z_i, f)$, $i = 1, 2, 3$

2. Si dimostri che una matrice quadrata $A_{n \times n}$ idempotente ha autovalori $\lambda_i \in \{0, 1\}$ per $i = 1, \dots, n$.
3. Supponiamo che alla matrice dei dati $X_{n \times p}$ sia associata la varianza/covarianza $S = \text{diag}(s_{11}, \dots, s_{pp})$ con $s_{11} \geq \dots \geq s_{pp} > 0$. Per questa particolare matrice di varianza/covarianze, ha senso effettuare l'analisi delle componenti principali basata sulla corrispondente matrice di correlazione R ? E' l'analisi fattoriale basata sulle variabili standardizzate? Giustificare le risposte.

PARTE B: esercizi di analisi dei dati

Esercizi da svolgere con il computer sulla piattaforma examonline con l'ausilio di R/Rstudio.

Problema 2

Si consideri il dataset `state.x77` presente nella libreria `datasets`. Questo dataset descrive 50 stati degli Stati Uniti d'America rispetto alle seguenti 8 variabili:

- **Population** Popolazione (in migliaia)
 - **Income** Reddito (in dollari pro capite)
 - **Illiterarcy** Analfabetismo (Percentuale della popolazione)
 - **Life Exp** Aspettativa di vita alla nascita (in anni)
 - **Murder** Numero di omicidi per 100,000 abitanti
 - **HS Grad** Percentuale di adulti diplomati
 - **Frost** Numero medio di giorni freddi all'anno con temperature sotto lo zero
 - **Area** Superficie (in miglia quadrate)
1. Sia X la matrice 50×8 corrispondente al dataset `state.x77`. Calcolare, riportando tutti i risultati arrotondando al **secondo decimale** (si ricordi l'uso della **virgola** per i decimali)
 - a. la lunghezza del vettore scarto dalla media \tilde{x}_2 (variabile **Income**);
 - b. il coseno dell'angolo (espresso in radianti) compreso tra \tilde{x}_1 e \tilde{x}_2 (variabili **Population** e **Income**);
 - c. Calcolare l'elemento di posizione $[2, 2]$ (riga 2, colonna 2) della matrice $R^{-1/2}$, dove R indica la matrice di correlazione associata a X .
 2. Calcolare il numero di osservazioni anomale verificando se la distanza di Mahalanobis al quadrato di ciascuna osservazione dal baricentro è superiore alla soglia s , dove s corrisponde al quantile 0.95 di una variabile casuale χ_p^2 (dove p è il numero di colonne di X).
 3. Rimuovere da X le osservazioni anomale individuate al punto precedente e svolgere l'analisi delle componenti principali, decidendo opportunamente se basarla sui dati originali o sui dati standardizzati. Calcolare la correlazione in valore assoluto tra il vettore dei punteggi y_1 della prima componente principale e la variabile **Area** (standardizzata oppure no a seconda della scelta effettuata). Riportare il risultato arrotondando al **terzo decimale** (si ricordi l'uso della **virgola** per i decimali).
 4. Si supponga di aggiungere al dataset X considerato al punto 1. una nuova variabile: **Density** = **Population** / **Area**. Riportare il numero di colonne linearmente indipendenti per il nuovo dataset.

Problema 3

Si consideri il dataset `state.center` (anch'esso presente nella libreria `datasets`), che riporta la longitudine (con segno negativo, variabile **x**) e la latitudine (variabile **y**) del centro geografico di ogni stato considerato nel Problema 2 (tranne che per l'Alaska e le Hawaii, che sono messe artificialmente da qualche parte a ovest della costa).

1. Calcolare la matrice D delle distanze Euclidee tra gli Stati, e riportare il valore della distanza (non nulla) più piccola, arrotondando al **terzo decimale** (si ricordi l'uso della **virgola** per i decimali).
2. Utilizzare l'algoritmo agglomerativo gerarchico con il legame medio. Decidere il numero di gruppi K ottimale secondo il criterio della *silhouette*, per $K = 2, \dots, 10$. Riportare
 - a. il numero di gruppi K ottimale;
 - b. il valore medio della *silhouette* corrispondente alla scelta effettuata, arrotondando al **terzo decimale** (si ricordi l'uso della **virgola** per i decimali).

3. Riportare i comandi R per visualizzare il diagramma di dispersione delle variabili latitudine e longitudine, colorandogli Stati secondo l'attribuzione in gruppi ottenuta al punto precedente. Il codice R deve essere riproducibile, ovvero se eseguito da cima a fondo non deve produrre errori. Questo implica che bisogna iniziare dall'import dei dati, caricare le opportune librerie, etc.