

Cognome:
 Nome:
 Matricola:
 Tipologia d'esame: ☐ 12 CFU ☐ 15 CFU

Si svolgano gli esercizi riportando il risultato dove indicato. Durata: 80 minuti

- **Population** in migliaia
- **Income** in dollari pro capita
- **Illiterarcy** Percentuale della popolazione
- **Life Exp** Anni di aspettativa di vita alla nascita
- **Murder** Numero di omicidi e omicidi colposo per 100000 persone
- **HS Grad** Percentuale di adulti diplomati
- **Frost** Numero medio di giorni freddi all'anno con temperature sotto lo zero
- **Area** in miglia quadrate

Stati Uniti d'America



- a. Sia X la matrice 50×8 corrispondente al dataset `state.x77`. Sulla base della corrispondente matrice di varianze/covarianze S , svolgere l'analisi delle componenti principali e riportare la percentuale di varianza spiegata dalla prima componente principale.

- b. Riportare le varianze delle variabili presenti in X , arrotondando al primo decimale.

Alla luce dei risultati sopra ottenuti, indicare quali sono le problematiche per l'analisi delle componenti principali svolta al punto a.

- c. Svolgere l'analisi delle componenti principali sui dati standardizzati Z , riportando
- la media c delle percentuali di varianza spiegata da ciascuna componente principale
 - il numero di componenti principali con varianza spiegata superiore a c

- d. Si consideri la matrice dei dati Y di dimensioni 50×11 dove le prime 8 colonne sono uguali a quelle di X mentre le restanti 3 colonne sono le nuove variabili

- **Longitude**, ricavabile dal dataset `state.center`
- **Latitude**, ricavabile dal dataset `state.center`
- **Density** = `Population` / `Area` ricavabile dal dataset `state.x77`

Sia Q la matrice dei dati standardizzati ottenuta a partire da Y . Si svolga l'analisi delle componenti principali basata su Q (ovvero sulla base della matrice di correlazione R^Y), riportando i punteggi (*scores*) dello stato dell'Alaska relativamente alle prime tre componenti principali (arrotondati alla terza cifra decimale)

- e. La variabile **Density** costruita al punto precedente è funzione delle variabili **Population** e **Area**. Questo comporta che la matrice Y ha colonne linearmente dipendenti? Giustificare la risposta.

- f. Si consideri la stima di massima verosimiglianza per il modello fattoriale con k fattori basato sui dati standardizzati Q . 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".

- g. Stimare il modello fattoriale con $k = 5$ fattori con il metodo della massima verosimiglianza utilizzando i dati standardizzati Q e senza effettuare alcuna rotazione. Riportare le "stime" dei punteggi fattoriali con il metodo di Thomson per lo stato dell'Alaska (arrotondando al secondo decimale)

- h. Applicare l'algoritmo delle K medie (**algorithm = "Hartigan-Wong"**) per i dati standardizzati Q iniziando i K centri utilizzando le prime K osservazioni (righe $1, \dots, K$ della matrice dei dati Q). Arrotondando il risultato alla seconda cifra decimale, riportare per $K = 2, \dots, 8$

- il valore dell'indice $CH(K) = \frac{B/(K-1)}{W/(n-K)}$ di Calinski and Harabasz
- il valore medio della *silhouette* considerando come matrice delle distanze quella ottenuta con la metrica Euclidea basata su Q

K	2	3	4	5	6	7	8
$CH(K)$							
$silhouette(K)$							

Esercizio 2 (Punti 3)

Si consideri il modello fattoriale con 1 fattore:

$$z_1 = \lambda_1 f + u_1$$

$$z_2 = \lambda_2 f + u_2$$

$$z_3 = \lambda_3 f + u_3$$

$$\text{dove } \widehat{Cov}(z) = R_{3 \times 3} = \begin{bmatrix} 1 & 0.5 & 0.6 \\ & 1 & 0.7 \\ & & 1 \end{bmatrix}.$$

Arrotondando il risultato al secondo decimale, riportare le stime $\hat{\Lambda}$ e $\hat{\Psi}$ utilizzando il metodo di stima *naive*.

Esercizio 3 (Punti 3)

Alla matrice di varianze/covarianze $S_{p \times p}$ sono associati i seguenti autovalori $\lambda_1 = 6, \lambda_2 = 4$ e autovettori

$$\text{normalizzati } v_1 = \begin{bmatrix} 1/\sqrt{5} \\ 2/\sqrt{5} \end{bmatrix}, v_2 = \begin{bmatrix} 2/\sqrt{5} \\ -1/\sqrt{5} \end{bmatrix}.$$

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

- b. Calcolare la correlazione tra la prima colonna \tilde{x}_1 di \tilde{X} e i punteggi y_1 della prima componente principale, arrotondando il risultato al secondo decimale:

Esercizio 4 (punti 5)

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

- a. $d_m(y_i, y_l) = d_m(x_i, x_l)$ dove d_m è la distanza di Minkowski di ordine $m \geq 1$, $y'_i = x'_i + b$ e $y'_l = x'_l + b$

- b. se $\det(S) = 0$, allora le colonne di \tilde{X} sono linearmente dipendenti;