

Corso di laurea: ☐ SGI ☐ SSE Anno: ☐ II ☐ III o più

La durata della prova è di 60 minuti.
Si svolgano gli esercizi 1 e 2 riportando il risultato dove indicato.

Alla matrice $S_{p \times p}$ sono associati i seguenti autovalori $\lambda_1 = 6, \lambda_2 = 4$ e autovettori 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. Quante sono le colonne di X ? $p = \dots$

- b. Determinare la matrice di varianze/covarianze $S_{p \times p}$

- c. Riportare

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

- d. Bestimmen Sie $S^2_{p \times p}$

- e. Calcolare la proporzione di varianza spiegata dalla prima componente principale

=====

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

[illegible]

- g. Determinare (arrotondando al secondo decimale) la matrice di correlazione $R_{p \times p}$ =

- h. Riportare (arrotondando al secondo decimale) gli autovalori e gli autovettori normalizzati di R calcolata al punto precedente.

```

##      [,1] [,2]
## [1,]  4.4  0.8
## [2,]  0.8  5.6

## [1] 24
## [1] 10
## [1] 0.97

##      [,1] [,2]
## [1,]   20   8
## [2,]    8  32

## [1] 0.6
## [1] 0.52

##      [,1] [,2]
## [1,] 1.00 0.16
## [2,] 0.16 1.00

## [1] 1.16 0.84

##      [,1] [,2]
## [1,] 0.71 -0.71
## [2,] 0.71  0.71

```

Esercizio 2. Punteggio: 6.5

Si consideri il dataset **quakes** presente nella libreria **datasets**, che contiene $n = 1000$ osservazioni (eventi sismici) su cui sono state misurate le seguenti 5 variabili:

- *lat* latitudine dell'evento sismico
- *long* longitudine dell'evento sismico
- *depth* profondità (in km) dell'evento sismico
- *mag* magnitudo (scala Richter)
- *stations* Numero di stazioni che hanno riportato l'evento sismico

- a. Si consideri la matrice $X_{1000 \times 5}$ che contiene le seguenti variabili: *lat*, *long*, *depth*, *mag* e *stations*. Si costruisca il diagramma a scatola con baffi (*boxplot*) per ciascuna delle variabili presenti in $X_{1000 \times 5}$ e si riporti il numero di valori anomali (*outliers*).

	lat	long	depth	mag	stations
numero di valori anomali					

```
##      lat      long      depth      mag stations
##      32      204         0         7         54
```

- b. Per la matrice $X_{1000 \times 5}$ calcolata al punto a., si calcoli il quadrato della distanza di Mahalanobis di ciascuna unità statistica u'_i dal baricentro \bar{x} e si riporti il valore minimo e il valore massimo, arrotondando i calcoli al secondo decimale.

```
## [1] 0.57
```

```
## [1] 25.9
```

$\min_{i=1, \dots, 1000} \{d_M^2(u_i, \bar{x})\} = \dots\dots\dots$	$\max_{i=1, \dots, 1000} \{d_M^2(u_i, \bar{x})\} = \dots\dots\dots$
---	---

- c. Utilizzare l'algoritmo delle K-medie (specificando **algorithm = "Lloyd"**) per formare $K = 4$ gruppi sulla base della matrice dei dati standardizzati $Z_{1000 \times 5}$ ottenuta a partire da $X_{1000 \times 5}$, inizializzando i centroidi con le osservazioni di riga 200, 400, 600 e 800, ed eseguendo l'algoritmo una sola volta. Riportare i valori dei centroidi dei 4 gruppi ottenuti, arrotondando alla seconda cifra decimale.

```
##      lat      long      depth      mag      stations
## 1 -0.15      0.23      0.02      1.61         1.82
## 2  0.95     -1.89     -0.76      0.22         -0.10
## 3 -0.51      0.70     -0.79     -0.26         -0.34
## 4  0.01      0.25      1.03     -0.59         -0.46
```

	Centroidi				
	lat	long	depth	mag	stations
Gruppo 1					
Gruppo 2					
Gruppo 3					
Gruppo 4					

- d. Calcolare, arrotondando al secondo decimale, l'indice di Calinski and Harabasz per i quattro gruppi individuati al punto c.

[1] 474.8148

Indice di Calinski and Harabasz =