

Cognome: Nome: Matricola:

Prova scritta di ASM - Modulo Analisi Esplorativa del 23.02.2017

La durata della prova è di 60 minuti.

Si svolgano gli esercizi 1 e 2 riportando il risultato dove indicato.

Esercizio 1 (Punti: 7)

Decomposizione Spettrale e Analisi delle Componenti Principali

Data la matrice di varianze/covarianze $S_{3 \times 3} = \begin{bmatrix} 4 & 3 & 1 \\ 3 & 9 & 2 \\ 1 & 2 & 1 \end{bmatrix}$

a. Riportare

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

b. Determinare la matrice di correlazione $R_{3 \times 3} = \begin{bmatrix} \dots & \dots & \dots \\ \dots & \dots & \dots \\ \dots & \dots & \dots \end{bmatrix}$.

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.

=

```
## [1] 14
```

```
## [1] 14
```

```
## [1] 0.39
```

```
##      [,1]      [,2]      [,3]
## [1,]  1.0 0.5000000 0.5000000
## [2,]  0.5 1.0000000 0.6666667
## [3,]  0.5 0.6666667 1.0000000
```

```
##      [,1] [,2] [,3]
## [1,] 1.89  0.6 0.26
## [2,] 0.60  2.9 0.50
## [3,] 0.26  0.5 0.83
```

```
## [1] -0.23
```

Distanze e Cluster Analysis

- e. Per una generica matrice di dati $X_{n \times p}$, si riporti la definizione di distanza di Lagrange $d_\infty(u_i, u_l)$ tra due unità statistiche u'_i e u'_l .

$$d_\infty(u_i, u_l) =$$

- f. 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 .

$$d(G_1, G_2) =$$

Esercizio 2 (Punti: 6.5)

Si consideri il dataset `crabs` presente nella libreria `MASS` (per caricare la libreria si usi il comando `library(MASS)`), che contiene $n = 200$ unità statistiche (granchi *Leptograpsus*) relative alle seguenti 8 variabili:

- *sp* (specie: B = blu o O = arancione)
- *sex* (sesso: F = femmina o M = maschio)
- *index* (indice)
- *FL* (misura del lobo frontale in mm)
- *RW* (larghezza della parte posteriore in mm)
- *CL* (lunghezza del carapace in mm)
- *CW* (larghezza del carapace in mm)
- *BD* (profondità del corpo in mm)

- a. Si consideri la matrice $X_{200 \times 5}$ che contiene le seguenti variabili: *FL*, *RW*, *CL*, *CW* e *BD*. Per ciascuna unità statistica u'_i , si calcoli la distanza di Mahalanobis dal baricentro \bar{x}' e si riporti il numero di unità statistiche con distanza di Mahalanobis superiore a 3.33:

$$\sum_{i=1}^n I\{d_M(u_i, \bar{x}) > 3.33\} = \dots$$

```
## [1] 12
```

- b. Calcolare la matrice dei dati standardizzati $Z_{200 \times 5}$. Sulla base di $Z_{200 \times 5}$, calcolare la matrice delle distanze $D_{200 \times 200}$ utilizzando la metrica di Lagrange, ed effettuare l'analisi dei cluster gerarchica utilizzando il legame completo, ricavandone 3 gruppi. Riportare i valori della tabella a doppia entrata che incrocia la classificazione ottenuta e la variabile `sp`.

```
##
```

```
## gruppi  B  O
##      1 21  9
##      2 62 43
##      3 17 48
```

	Specie	
	B	O
Gruppo 1		
Gruppo 2		
Gruppo 3		

- c. Calcolare, arrotondando al secondo decimale, il valore medio della silhouette per i tre gruppi individuati al punto b. (utilizzando il comando `silhouette` presente nella libreria `cluster`).

```
## Loading required package: cluster
```

```
##      1      2      3
## 0.62 0.31 0.40
```

	Valore medio della Silhouette
Gruppo 1	
Gruppo 2	
Gruppo 3	