**25 Gennaio 2018 - Analisi Esplorativa**

Cognome: .....

Nome: .....

Matricola: .....

Tipologia d'esame: ☐ 12 CFU ☐ 15 CFU

---

**Prova scritta**

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

---

**Esercizio 1 (7 punti)**

Si consideri la seguente matrice di correlazione $\underset{3\times3}{R} = \begin{bmatrix} 1 & 1/2 & 1/2 \\ 1/2 & 1 & 2/3 \\ 1/2 & 2/3 & 1 \end{bmatrix}$.

a. Riportare l'indice di variabilità relativo, arrotondando al secondo decimale: .....

b. Sia $\underset{n\times3}{\tilde{X}}$ la matrice dei dati centrati. Determinare l'angolo (espresso in gradi) tra $\underset{n\times1}{\tilde{x}_1}$ e $\underset{n\times1}{\tilde{x}_3}$ : = .....

c. Sapendo che $s_{11} = 4$, $s_{22} = 9$ e $\text{tr}(S) = 14$, calcolare $\det(S) =$ ....., dove $\underset{3\times3}{S}$ rappresenta la matrice di varianze/covarianze.

d. Calcolare $S = \begin{bmatrix} ..... & ..... & ..... \\ ..... & ..... & ..... \\ ..... & ..... & ..... \end{bmatrix}$

e. Determinare gli autovalori di $S$, arrotondando al secondo decimale: $\lambda_1 =$ ....., $\lambda_2 =$ ....., $\lambda_3 =$ .....

f. Calcolare, arrotondando al secondo decimale, $S^{1/2} = \begin{bmatrix} ..... & ..... & ..... \\ ..... & ..... & ..... \\ ..... & ..... & ..... \end{bmatrix}$.

g. Calcolare la correlazione tra la terza colonna $\underset{n\times1}{\tilde{x}_3}$ di $\underset{n\times3}{\tilde{X}}$ e i punteggi $\underset{n\times1}{y_2}$ della seconda componente principale (calcolata a partire da $S$), arrotondando al secondo decimale: = .....

h. Si consideri la seguente trasformazione lineare: $\underset{n\times3}{W} = \underset{n\times3}{X}\,\underset{3\times3}{A'}$, dove $A = \text{diag}(1, 2, 3)$. Calcolare la matrice di varianze/covarianze di $W$, $S^W = \begin{bmatrix} ..... & ..... & ..... \\ ..... & ..... & ..... \\ ..... & ..... & ..... \end{bmatrix}$

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

```
## [1] 60
```

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

```
##      [,1] [,2] [,3]
## [1,]    4    3    1
## [2,]    3    9    2
## [3,]    1    2    1
```

```
## [1] 10.91  2.60  0.49
```

```
##      [,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] 0.04
```

```
##      [,1] [,2] [,3]
## [1,]    4    6    3
## [2,]    6   36   12
## [3,]    3   12    9
```

**Esercizio 2 (2 punti)**

Riportare le seguenti definizioni (in forma matriciale), specificando tutte le quantità coinvolte:

a. Vettore delle medie

b. Matrice di centramento

c. Matrice dei dati centrati

d. Matrice di varianze/covarianze

e. Matrice dei dati standardizzati

f. Matrice dei dati ortogonalizzati

**Esercizio 3 (4 punti)**

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

a. che la matrice di varianze/covarianze $S$ è semi-definita positiva, esplicitando tutti i passaggi.

b. che $\text{tr}(S) = \sum_{j=1}^{p} \lambda_j$, dove $\lambda_1, \ldots, \lambda_p$ sono gli autovalori di $S$.

d. Enunciare il Teorema di Eckart-Young.

**Esercizio 4 (2 punti)**

Un gruppo di $n = 112$ individui si è sottoposto a $p = 6$ prove di abilità e intelligenza. Caricare la matrice di varianza/covarianza `ability.cov` presente nella libreria `dataset` e si risponda alle seguenti domande:

a. Sulla base dalla matrice di correlazione $R$, si stimi il modello fattoriale con $k = 2$ fattori utilizzando il metodo della massima verosimiglianza ed effettuando la rotazione varimax. Arrontondando al secondo decimale, si riportino le stime delle comunalità

$$\hat{h}_1^2 = \ldots\ldots\ldots, \hat{h}_2^2 = \ldots\ldots\ldots, \hat{h}_3^2 = \ldots\ldots\ldots, \hat{h}_4^2 = \ldots\ldots\ldots, \hat{h}_5^2 = \ldots\ldots\ldots, \hat{h}_6^2 = \ldots\ldots\ldots$$

```
## [1] 0.54 0.41 0.78 0.23 0.95 0.67
```

**Esercizio 5 (2 punti)**

a. Siano date due unità statistiche $u_1' = (2, 3)$ e $u_2' = (1, 1)$. Riportare la distanza Euclidea $d_2(u_1, u_2) = \ldots\ldots\ldots$, di Manhattan $d_1(u_1, u_2) = \ldots\ldots\ldots$, di Lagrange $d_\infty(u_1, u_2) = \ldots\ldots\ldots$

b. Si consideri la seguente matrice di distanze relativa a tre unità statistiche $u_1$, $u_2$ e $u_3$:

| $d(u_i, u_l)$ | $u_1$ | $u_2$ | $u_3$ |
|---|---|---|---|
| $u_1$ | 0 | | |
| $u_2$ | 3 | 0 | |
| $u_3$ | 5 | 4 | 0 |

Completare la tabella sottostante calcolando la decomposizione della distanza totale $T = \frac{1}{2}\sum_{i=1}^{3}\sum_{l=1}^{3} d(u_i, u_l)$ in distanza entro i gruppi $W$ e tra i gruppi $B$ per le tre unità statistiche $u_1$, $u_2$ e $u_3$ raggruppate in due gruppi $G_1$ e $G_2$:

| $G_1, G_2$ | $W$ | $B$ | $T$ |
|---|---|---|---|
| $(u_1)$, $(u_2, u_3)$ | ..... | ..... | ..... |
| $(u_1, u_2)$, $(u_3)$ | ..... | ..... | ..... |
| $(u_1, u_3)$, $(u_2)$ | ..... | ..... | ..... |

**Esercizio 6 (2 punti)**

Riportare la statistica test con la correzione di Bartlett:

$T_{Bartlett} =$

**Esercizio 7 (3 punti)**

a. Si riporti il modello fattoriale con $k$ fattori in forma matriciale, specificando tutte le assunzioni.

b. Si dimostri che $S^Z = R^X$, ovvero che la matrice di varianze/covarianze calcolata per $Z$ risulta uguale alla matrice di correlazione calcolata per $X$.

**Esercizio 8 (4 punti)**

Si consideri il dataset `swiss` presente nella libreria `datasets`, che contiene $n = 47$ unità statistiche (provincie) relative alle seguenti 6 variabili:

- *Fertility* : common standardized fertility measure
- *Agriculture* : % of males involved in agriculture as occupation
- *Examination* : % draftees receiving highest mark on army examination
- *Education* : % education beyond primary school for draftees
- *Catholic* : % catholic (as opposed to protestant)
- *Infant.Mortality* : live births who live less than 1 year

a. Per ciascuna unità statistica, si calcoli la distanza di Mahalanobis dal baricentro e si riporti il nome delle provincie con distanza superiore a 3.6:

```
##    La Vallee V. De Geneve
##            19           45
```

b. Dopo aver standardizzato i dati, eseguire l'algoritmo delle $K$-medie (`algorithm = Hartigan-Wong`) per $K = 2, 4, 6$, inizializzando i centroidi con le osservazioni di riga $1, 2, \ldots, K$. Riportare per ciascun valore di $K$ il rispettivo valore dell'indice Calinski and Harabasz, arrotondando al secondo decimale.

```
##    [,1] [,2]  [,3]
## K  2.00  4.0  6.00
##   24.72 24.3 19.85
```

| $K$ | 2 | 4 | 6 |
|---|---|---|---|
| Indice CH | ..... | ..... | ..... |

c. Sulla base della matrice dei dati standardizzati, calcolare la matrice delle distanze utilizzando la metrica Euclidea. Riportare, arrotondando al secondo decimale, il valore medio della silhouette per i $K = 4$ gruppi determinati nel punto precedente.

| gruppo | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| silhouette (media) | ..... | ..... | ..... | ..... |

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

```
##    1    2    3    4
## 0.16 0.39 0.39 0.34
```