Analisi Esplorativa (Analisi Statistica Multivariata)

Prova d'esame

15 Luglio 2022

Tempo a disposizione: 110 minuti

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

PARTE A: esercizi di teoria

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

Problema 1

1. Si supponga che la matrice dei dati X consista di due colonne $\underset{n\times 1}{x_1}$ e $\underset{n\times 1}{x_2}$ tali che $x_2 = -x_1$.
   a. Calcolare la matrice di correlazione R.
   b. Determinare gli autovalori di R.
   c. Calcolare l'indice relativo di variabilità.
2. Sia H la matrice di centramento di dimensione $n \times n$, e $\underset{n\times n}{J} = \frac{1}{n}11'$, dove $\underset{n\times 1}{1}$ è il vettore unitario di lunghezza n.
   a. Calcolare $H1$.
   b. Calcolare $J1$.
   c. Calcolare UH, dove $U = nJ$.
3. Si consideri il modello fattoriale con le seguenti matrici di pesi fattoriali e varianze specifiche:

$$\Lambda = \begin{bmatrix} 4 & 1 \\ 7 & 2 \\ -1 & 6 \\ 1 & 8 \end{bmatrix}, \qquad \Psi = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix}$$

   a. Calcolare la covarianza tra la prima variabile e il primo fattore comune, i.e. $\mathbb{C}\text{ov}(x_1, f_1)$
   b. Calcolare la varianza delle prima variabile, i.e. $\mathbb{V}\text{ar}(x_1)$
   c. Calcolare la covarianza tra le prime due variabili, i.e. $\mathbb{C}\text{ov}(x_1, x_2)$

PARTE B: esercizi di analisi dei dati

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

Problema 2

La seguente tabella riporta le stime di un modello fattoriale ottenute, previa standardizzazione dei dati, da misurazioni su $p = 6$ variabili **(nel vostro esercizio su esamionline la tabella potrebbe essere diversa)**.

	Factor1	Factor2
VAR1	NA	0.543
VAR2	0.156	0.622
VAR3	0.206	0.860
VAR4	0.109	0.468
VAR5	0.956	0.182
VAR6	0.785	NA

1. Sapendo che le stime delle varianze specifiche per le variabili VAR1 e VAR6 sono 0.455 e 0.334, rispettivamente, calcolare il valore mancante per VAR1. Riportare il risultato arrotondando al **terzo decimale** (si ricordi l'uso della **virgola** per i decimali).

```
# 1
x[1]
```

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

```
NA1 <- sqrt(1 - Lambda[x[1],2]^2 - Psi[x[1]])
a = round(NA1, 3)
a
```

```
## [1] 0.5
```

- Si calcoli la stima della comunalità per la variabile VAR6. Riportare il risultato arrotondando al **terzo decimale** (si ricordi l'uso della **virgola** per i decimali).

```
# b.
x[2]
```

```
## [1] 6
```

```
NA2 <- sqrt(1 - Lambda[x[2],1]^2 - Psi[x[2]] )
b = round(Lambda[x[2],1]^2 + NA2^2 ,3)
b
```

```
## [1] 0.666
```

- Si calcoli la proporzione di varianza spiegata dal primo fattore comune. Riportare il risultato arrotondando al **terzo decimale**. Si ricordi l'uso della **virgola** per i decimali.

```
# c.
PROP1 = (sum(Lambda[-x[1],1]^2) + NA1^2 )/6
c = round(PROP1,3)
c
```

```
## [1] 0.31
```

Problema 3

Si consideri il dataset `Animals` presente nella libreria `MASS`, che contiene $n = 28$ osservazioni misurate su $p = 2$ variabili: `body` (body weight in kg) e `brain` (brain weight in g). Sia X la matrice 27×2 contenente le

variabili `log(body)` e `log(brain)` (ovvero le variabili body e brain weight trasformate al logaritmo) e che esclude la riga 1 del dataset `Animals`.

- Calcolare il numero di osservazioni anomale presenti in X verificando se la distanza di Mahalanobis al quadrato di ciascuna osservazione dal baricento è superiore alla soglia s, dove s corrisponde al quantile 0.95 di una variabile casuale χ^2_p (dove p è il numero di colonne di X).

```
library(MASS)
X <- log(Animals[-1,])
n <- nrow(X)
p <- ncol(X)

# a.
xbar = matrix(colMeans(X), nrow=p, ncol=1)
S = var(X) * ((n-1)/n)
InvS = solve(S)
dM2 = apply(X,MARGIN=1, function(u) t(u-xbar) %*% InvS %*% (u - xbar) )
s = qchisq(0.95, df=p)
a = sum(dM2 > s)
a
```

```
## [1] 2
```

- Sia W la matrice dei dati X escludendo le osservazioni anomale individuate al punto precedente. Si consideri la matrice A, che rappresenta la migliore approssimazione di rango 1 della matrice W. Riportare $\|W - A\|_F^2 = \sum_i \sum_j (w_{ij} - a_{ij})^2$, arrotondando al **terzo decimale** (si ricordi l'uso della **virgola** per i decimali).

```
# b.
out = which(dM2 > s)
W = X[-out, ]
n_W = nrow(W)
S_W = var(W) * ((n_W-1)/n_W)
lambdas = eigen(S_W)$values
b = round(n_W * lambdas[2],3)
b
```

```
## [1] 21.406
```

- Si consideri l'algoritmo delle K-medie sui dati W, considerando come attribuzione iniziale dei centroidi le unità statistiche Potar monkey e Rat. Si consideri il primo passo dell'algoritmo, che individua per ciascuna unità statistica il centroide più vicino. Quante unità statistiche sono più vicine al centroide identificato da Potar monkey ?

```
# c.
u1
```

```
## [1] 8
```

```
u2
```

```
## [1] 23
```

```
D2 = as.matrix(dist(W, diag = T, upper = T))^2
gruppi = sapply(1:n_W, function(i) which.min(c(D2[i,u1],D2[i,u2])) )
c = sum(gruppi==1)
c
```

```
## [1] 20
```