Analisi Esplorativa (Analisi Statistica Multivariata)

Prova d'esame

24 Giugno 2022

Tempo a disposizione: 120 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. Data la seguente matrice dei dati

$$X = \begin{bmatrix} 9 & 1 \\ 5 & 3 \\ 1 & 2 \end{bmatrix} \tag{1}$$

   a. Si calcolino il vettore delle medie $\bar{x}$ e i vettori scarto dalla media $\tilde{x}_1$ e $\tilde{x}_2$.

   b. Si calcolino la lunghezza dei vettori $\tilde{x}_1$ e $\tilde{x}_2$ e il coseno dell'angolo compreso.

   c. Si ricavino la matrice di varianze e covarianze S e la matrice di correlazione R facendo riferimento ai risultati ottenuti nel punto precedente.

2. Mostrare che il determinante di una matrice simmetrica a valori reali A di dimensioni $p \times p$ può essere espresso come prodotto degli autovalori di A.

3. Siano S e $\tilde{S}$ due matrici di varianze e covarianze definite come segue:

$$S = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \tilde{S} = \begin{pmatrix} 1 & -\frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & 1 & -\frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{2} & 1 \end{pmatrix}$$

   Si calcolino per S e $\tilde{S}$ la varianza totale e la varianza generalizzata.

PARTE B: esercizi di analisi dei dati

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

Problema 2

Si consideri il dataset `USArrest` presente nella libreria `datasets`. Per ciascuno dei $n = 50$ stati negli Stati Uniti, il data set contiene il numero degli arresti per 100.000 residenti per ciascuno dei seguenti tre reati: `Assault` (aggressione); `Murder` (omicidio); `Rape` (stupro). E' inoltre presente la variabile `UrbanPop`, che indica la percentuale della popolazione in ogni stato che vive nelle aree urbane. Sia X la matrice 49×4 corrispondente al dataset `USArrest` ma rimuovendo la riga 1 **(nel vostro esercizio su esamionline il numero di riga potrebbe essere diverso)**.

1. Calcolare $S^{-1/2}$, dove S è la matrice di varianze/covarianze di X. Riportare il valore massimo presente in $S^{-1/2}$, arrotondando al **terzo decimale** (si ricordi l'uso della **virgola** per i decimali).

```
# 1
X <- USArrests[-1,]
n <- nrow(X)
p <- ncol(X)
S = var(X) * ((n-1)/n)
sqrtinvS = eigen(S)$vector %*% diag(sqrt(1/eigen(S)$values)) %*% t(eigen(S)$vector)
a = max(sqrtinvS)
round(a,3)
```

```
## [1] 0.404
```

2. Svolgere l'analisi delle componenti principali, decidendo opportunamente se basarla sui dati originali o sui dati standardizzati. Calcolare la correlazione **in valore assoluto** tra il vettore dei punteggi di ciascuna componente principale (prima, seconda, terza e quarta) e ciascuna variabile (`Assault`, `Murder`, `Rape` e `UrbanPop`, standardizzata oppure no a seconda della scelta effettuata), quindi in totale 16 correlazioni in valore assoluto (quattro componenti principali per quattro variabili). Riportare il valore massimo di queste 16 correlazioni in valore assoluto, arrotondando al **terzo decimale** (si ricordi l'uso della **virgola** per i decimali).

```
# 2
R = cor(X)
V = eigen(R)$vector
lambdas = eigen(R)$values
abs_cor = sapply(1:p, function(j) abs( V[,j] * sqrt(lambdas[j]) ))
b = max(abs_cor)
round(b,3)
```

```
## [1] 0.917
```

```
# Soluzione alternativa
# Z = scale(X,center=TRUE, scale = sqrt(diag(var(X)*((n-1)/n)))  )
# scores = princomp(X, cor=TRUE)$scores
# sapply(1:p, function(j) abs(cor(Z[,j],scores)))
```

3. Determinare il numero q delle componenti principali in modo tale da spiegare almeno il 80% della variabilità. Sia A la migliore approssimazione di rango q della matrice W, dove W indicata la matrice dei dati centrati o standardizzati (a seconda della scelta effettuata al punto 2.). 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).

```
# 3
q = which.max(cumsum(lambdas)/p > .8)
residuals = n * sum(lambdas[(q+1):p])
c=residuals
round(c,3)
```

```
## [1] 26.014
```

Problema 3

Si consideri il dataset `USArrest` presente nella libreria `datasets` (per la descrizione si veda il Problema 2). Sia X la matrice 49×3 corrispondente al dataset `USArrest`, escludendo la riga 1 **(nel vostro esercizio su esamionline il numero di riga potrebbe essere diverso)** e la variabile `UrbanPop`.

1. Si consideri la decomposizione in valori singolari (SVD) di X. Calcolare i valori singolari, e riportare il valore minimo arrotondando al **terzo decimale** (si ricordi l'uso della **virgola** per i decimali).

```
# 2
rm(list=ls())
X <- USArrests[-1,-3]
n <- nrow(X)
p <- ncol(X)
a = min(svd(X)$d)
round(a,3)
```

```
## [1] 17.951
```

2. Sulla base dalla matrice di correlazione, si stimi il modello fattoriale con $k = 1$ fattore utilizzando il metodo della massima verosimiglianza senza effettuare alcuna rotazione. Calcolare i valori delle comunalità, e riportare il valore minimo arrotondando al **terzo decimale** (si ricordi l'uso della **virgola** per i decimali).

```
# 2
af = factanal(X, rotation="none", factors = 1)
Lambda = af$loadings[,]
h2 = Lambda^2
b = min(h2)
round(b,3)
```

```
## [1] 0.48
```

3. Calcolare il valore della statistica test $t = n \log\left(\frac{\det(\hat{\Lambda}\hat{\Lambda}' + \hat{\Psi})}{\det(R)}\right)$. Riportare il risultato arrotondando al **terzo decimale** (si ricordi l'uso della **virgola** per i decimali).

```
# 3
Psi = diag(af$uniqueness)
fit = Lambda %*% t(Lambda) + Psi
R = cor(X)
stat = n*log(det(fit)/det(R))
round(stat,3)
```

```
## [1] 0
```