Analisi Esplorativa (Analisi Statistica Multivariata)

Prova d'esame

8 Febbrario 2022

Tempo a disposizione: 150 minuti

Modalità di consegna: svolgere gli esercizi di teoria (parte A) riportando le soluzioni sul foglio protocollo, e consegnare il foglio protocollo assieme al testo della prova d'esame. 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.

Compilare con nome, cognome e numero di matricola. E' obbligatorio consegnare il testo della prova d'esame all'interno del foglio protocollo contente le soluzioni degli esercizi di teoria.

NOME:

COGNOME:

MATRICOLA:

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 $\underset{n\times n}{J} = \frac{1}{n}11'$, dove $\underset{n\times 1}{1}$ è il vettore unitario di lunghezza n.
   a. Calcolare $J1$.
   b. Si dimostri che $\underset{n\times n}{J}$ è una matrice idempotente.
   c. Calcolare UH, dove $U = nJ$ e $\underset{n\times n}{H}$ è la matrice di centramento.
   d. La matrice di centramento $\underset{n\times n}{H}$ è una matrice singolare? Giustificare la risposta.
3. Si supponga che la matrice dei dati X consista di due colonne, contenente i dati rilevati su un campione di n individui. La prima variabile (prima colonna) rappresenta il numero di anni compiuti dalla nascita (numero intero), la seconda l'altezza. Siano u_1' e u_2' le prime due righe della matrice X, a cui corrispondono a due individui entrambi con età pari a 35 anni. La distanza di Lagrange fra u_1' e u_2' è $d_\infty(u_1, u_2) = 10$.

a. Calcolare la distanza di Manhattan tra u_1' e u_2', i.e. $d_1(u_1, u_2)$.

b. Sia u_3' la terza riga della matrice X, a cui corrisponde un individuo di 29 anni, alto come u_2. Calcolare la distanza Euclidea tra u_1' e u_3', i.e. $d_2(u_1, u_3)$.

c. Il gruppo G_1 contiene solo l'unità statistica u_1', mentre il gruppo G_2 le unità statistiche u_2' e u_3'. Calcolare la distanza tra i due gruppi, utilizzando la distanza Euclidea con il metodo del legame singolo.

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 il numero di riga potrebbe essere diverso).

1. Calcolare il numero di osservazioni anomale 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).

```
# 1
X <- USArrests[-1,]
n <- nrow(X)
p <- ncol(X)
states <- row.names(X)
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)
sum(dM2 > s)
```

```
## [1] 4
```

2. Rimuovere da X le osservazioni anomale individuate al punto precedente e 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 y_1 della prima componente principale e ciascuna variabile (`Assault`, `Murder`, `Rape` e `UrbanPop`, standardizzata oppure no a seconda della scelta effettuata). Riportare il valore massimo delle correlazioni in valore assoluto, arrotondando al **terzo decimale** (si ricordi l'uso della **virgola** per i decimali).

```
# 2
out = which(dM2 > s)
X_no_out = X[-out, ]
R_no_out = cor(X_no_out)
V_no_out = eigen(R_no_out)$vector
lambdas_no_out = eigen(R_no_out)$values
abs_correlation = sapply(1:p, function(i) abs( sqrt(lambdas_no_out[1]) * V_no_out[i,1] ) )
round(max(abs_correlation),3)
```

```
## [1] 0.924
```

3. Determinare il numero q delle componenti principali in modo tale da spiegare almeno il 90% 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.) escludendo le osservazioni anomale individuate al punto 1. 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_no_out)/p > .9)
n_no_out = nrow(X_no_out)
residuals = n_no_out * sum(lambdas_no_out[(q+1):p])
round(residuals,3)
```

```
## [1] 17.779
```

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 il numero di riga potrebbe essere diverso) e la variabile `UrbanPop` .

1. 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).

```
# 1
rm(list=ls())
X <- USArrests[-1,-3]
n <- nrow(X)
p <- ncol(X)
af = factanal(X, rotation="none", factors = 1, scores = "regression")
Lambda = af$loadings[,]
h2 = Lambda^2
round(min(h2),3)
```

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

2. 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).

```
# 2
Psi = diag(af$uniqueness)
fit = Lambda %*% t(Lambda) + Psi
R = cor(X)
stat = round( n*log(det(fit)/det(R)) , 3)
round(stat,3)
```

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

3. Calcolare i punteggi fattoriali con il metodo di Thompson. Riportare il punteggio massimo (in valore assoluto) arrotondando al **terzo decimale** (si ricordi l'uso della **virgola** per i decimali).

```
round(max(abs(af$scores)),3)
```

```
## [1] 1.927
```

```r
Z <- scale(X, center=TRUE, scale=sqrt(diag(var(X))))
punteggi <- apply(Z,1, function(z) t(Lambda) %*% solve(R) %*% z)
round(max(abs(punteggi)),3)
```

```
## [1] 1.927
```