17 Aprile 2019 - Analisi Esplorativa

Cognome: ....

Nome: ....

Matricola: ....

Tipologia d'esame: ☐ 12 CFU ☐ 15 CFU

Prova scritta - fila A

Si svolgano gli esercizi riportando il risultato dove indicato. Durata: 70 minuti

---

Esercizio 1 (10 punti)

Il dataset `Boston` presente nella libreria `MASS` è composto da 506 osservazioni e 14 variabili: *crim*, *zn*, *indus*, *chas*, *nox*, *rm*, *age*, *dis*, *rad*, *tax*, *ptratio*, *black*, *lstat*, *medv*.

Le variabili di interesse per l'analisi sono trasformazioni delle variabili originali come elencato nel seguito:

- $x_1 = log(crim)$
- $x_2 = zn/10$
- $x_3 = log(indus)$
- $x_4 = log(nox)$
- $x_5 = log(rm)$
- $x_6 = (age)^{2.5}/10000$
- $x_7 = log(dis)$
- $x_8 = log(rad)$
- $x_9 = log(tax)$
- $x_{10} = exp(0.4 \cdot ptratio)/1000$
- $x_{11} = black/100$
- $x_{12} = \sqrt{lstat}$
- $x_{13} = log(medv)$

dove log indica il logaritmo naturale.

Si costruisca la matrice dei dati $\underset{506\times 13}{X}$ che contiene le variabili $x_1, x_2, \ldots, x_{13}$.

a. Si calcoli $d^2_M(x_i, \bar{x})$, il quadrato della distanza di Mahalanobis di ciascuna osservazione (ciascuna riga della matrice X) dal baricentro. Si riportino i valori di $d^2_M(x_i, \bar{x})$ solo se superano il valore 38, specificando anche l'indice della riga di X a cui si fa riferimento.

```
library(MASS)
Y = Boston
X = cbind(
  log(Y[,"crim"]),
  Y[,"zn"]/10,
  log(Y[,c("indus","nox","rm")]),
  (Y[,"age"]^(2.5))/10000,
```

```r
  log(Y[,c("dis","rad","tax")]),
  exp(0.4*Y[,"ptratio"])/1000,
  Y[,"black"]/100,
  sqrt(Y[,"lstat"]),
  log(Y[,"medv"])
)
n = nrow(X)
p = ncol(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) )
round(dM2[dM2 >  38],2)
```

```
  343   356   366   368   369   413
44.19 38.70 72.85 58.54 49.32 41.73
```

b. Sulla base della matrice dei dati standardizzati $\underset{506\times 13}{Z}$, applicare l'algoritmo delle K medie (`algorithm = "Hartigan-Wong"`) inizializzando i K centrodi utilizzando le prime K osservazioni (righe $1,\ldots,K$ della matrice Z). Arrotondando il risultato alla seconda cifra decimale, riportare per $K=5,6,\ldots,10$

- il valore dell'indice $\mathrm{CH}(K)=\frac{B/(K-1)}{W/(n-K)}$ di Calinski and Harabasz
- il valore medio della *silhouette* considerando come matrice delle distanze quella ottenuta con la metrica Euclidea basata su Z

K	5	6	7	8	9	10
$CH(K)$						
silhouette(K)						

```r
Z = scale(X, center=T, scale=diag(S)^(1/2))
D = dist(Z, method = "euclidean")
K = 5:10
CH <- vector()
sil <- vector()
library(cluster)
for (k in 1:length(K)){
km = kmeans(Z, centers=Z[1:K[k],], algorithm ="Hartigan-Wong")
CH[k] = (km$betweenss/(K[k]-1))/(km$tot.withinss/(n-K[k]))
sil[k] = summary(silhouette(x=km$cluster,dist=D))$avg.width
}
round(rbind(K,CH,sil),2)
```

```
      [,1]   [,2]   [,3]   [,4]   [,5]   [,6]
K     5.00   6.00   7.00   8.00   9.00  10.00
CH  202.35 174.22 148.52 143.76 158.47 139.62
sil   0.24   0.22   0.20   0.23   0.25   0.21
```

c. Si consideri la stima di massima verosimiglianza per il modello fattoriale con k fattori basato sui dati standardizzati Z. Riportare il p-value del primo test non significativo al livello 5% (e il corrispondente valore di k) per la sequenza di ipotesi nulle $H_0(k=1),H_0(k=2),H_0(k=3),\ldots$ dove $H_0(k)$="il modello fattoriale con k fattori è corretto".


```r
k = which.max(sapply(1:8, function(k) factanal(Z,factors=k)$PVAL > 0.05 ))
k
```

```
objective
        7
```

```r
round(factanal(Z,factors=k)$PVAL,4)
```

```
objective
   0.0806
```

d. Stimare il modello fattoriale con $k = 5$ fattori con il metodo della massima verosimiglianza utilizzando i dati standardizzati Z e senza effettuare alcuna rotazione. Riportare il valore della statistica test rapporto di verosimiglianza $T = n \log\left(\dfrac{det(\hat{\Lambda}\hat{\Lambda}' + \hat{\Psi})}{det(R)}\right)$ (arrotondando al terzo decimale)

```r
R = cor(X)
af5 = factanal(Z,factors=5, rotation="none", method="mle")
Lambda = af5$loadings[,]
Psi = diag(af5$uniqueness)
fit = Lambda %*% t(Lambda) + Psi
lrt = n*log(det(fit)/det(R))
round(lrt,3)
```

```
[1] 111.317
```

e. Stimare il modello fattoriale con $k = 7$ fattori con il metodo della massima verosimiglianza utilizzando i dati standardizzati Z e senza effettuare alcuna rotazione. Riportare i punteggi fattoriali $\hat{f}_i$ con il metodo di Thompson (arrotondando alla seconda cifra decimale) per l'unità statistica corrispondente alla riga $i = 100$ della matrice Z.

```r
af7 = factanal(Z, factors=7, rotation="none", method="mle", scores = "regression")
round(af7$scores[100,],2)
```

```
Factor1 Factor2 Factor3 Factor4 Factor5 Factor6 Factor7
  -1.26    0.85   -0.08   -0.18   -0.37   -0.65    0.42
```

Esercizio 2 (6 punti)

Sia

$$R = \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}$$

con $0 < r < 1$.

a. Determinare gli autovalori di R:

$$\lambda_1 = \dots\dots\dots\dots \qquad \lambda_2 = \dots\dots\dots\dots$$

b. Determinare gli autovettori normalizzati di R:

$$v_1 = \begin{bmatrix} \cdots\cdots\cdots \\ \\ \cdots\cdots\cdots \end{bmatrix}, \qquad v_2 = \begin{bmatrix} \cdots\cdots\cdots \\ \\ \cdots\cdots\cdots \end{bmatrix}$$

c. Determinare i punteggi delle componenti principali

$$y_{i1} = \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots, \qquad y_{i2} = \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots, \qquad i = 1, \ldots, n$$

Esercizio 3 (4 punti)

Dimostrare, esplicitando tutti i passaggi e le quantità coinvolte, che il vettore $\underset{p\times 1}{u}$ di lunghezza unitaria che risolve il problema di massimo

$$\max_{u:\|u\|=1} u'Su$$

è l'autovettore $\underset{p\times 1}{v_1}$ (con segno positivo o negativo) della matrice di varianze/covarianze $\underset{p\times p}{S}$.

Esercizio 4 (3 punti)

Si consideri la seguente matrice di varianze/covarianze $\underset{3\times 3}{S} = \begin{bmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{bmatrix}$

dove $a > 0$, $b > 0$ e $c > 0$ sono costanti non note.

a. Calcolare la varianza totale

$$= \ldots\ldots\ldots\ldots$$

b. Calcolare la varianza generalizzata

$$= \ldots\ldots\ldots\ldots$$

c. Calcolare l'indice di variabilità relativo

$$= \ldots\ldots\ldots\ldots$$

d. Determinare l'inversa della matrice di correlazione

$$\underset{3\times 3}{R}^{-1} = \begin{bmatrix} \ldots\ldots & \ldots\ldots & \ldots\ldots \\ \ldots\ldots & \ldots\ldots & \ldots\ldots \\ \ldots\ldots & \ldots\ldots & \ldots\ldots \end{bmatrix}.$$

Esercizio 5 (3 punti)

Enunciare il teorema di Eckart-Young.