25 Gennaio 2018 - Analisi Esplorativa

Cognome: ....

Nome: ....

Matricola: ....

Tipologia d'esame: ☐ 12 CFU ☐ 15 CFU

Prova scritta - fila B

Si svolgano gli esercizi riportando il risultato dove indicato. Durata: 40 minuti

---

Esercizio 1 (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

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

```r
# a
rm(list=ls())
n = 112
S = ability.cov$cov
D = diag(diag(S)^(-1/2))
R = D %*% S %*%D
af <- factanal(covmat=R, factors=2, n.obs = n, rotation = "varimax")
round(apply(af$loadings[,]^2,1,sum),2)
```

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

Esercizio 2 (2 punti)

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

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 3 (2 punti)

Riportare la statistica test con la correzione di Bartlett:

$T_{Bartlett} =$

Esercizio 4 (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 5 (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:

```r
# a
rm(list=ls())
X = as.matrix(swiss)
n = nrow(X)
xbar = matrix(colMeans(X), ncol=1)
S = var(X)*((n-1)/n)
InvS = solve(S)
dM2 = apply(X,1, function(u) t(u-xbar) %*% InvS %*% (u - xbar) )
which(sqrt(dM2) > 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.

```r
# b
Z = scale(X, center=T, scale= diag(S)^(1/2))
K = c(2,4,6)
CH <- vector()
for (i in 1:length(K)){
  k = K[i]
  km = kmeans(Z, centers = Z[1:k,])
  W = km$tot.withinss
  B = km$betweenss
  CH[i] = (B/(k-1)) / (W/(n-k))
}
rbind(K, round(CH,2))
```

```
##    [,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)	......	......	......	......

```r
# c
require(cluster)
```

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

```r
D = dist(Z)
gruppi = kmeans(Z, centers = Z[1:4,])$cluster
sil = silhouette(gruppi, dist=D)
round( summary(sil)$ clus.avg.widths, 2)
```

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