3 Luglio 2019 - Analisi Esplorativa

Cognome: ....

Nome: ....

Matricola: ....

Tipologia d'esame: ☐ 12 CFU ☐ 15 CFU

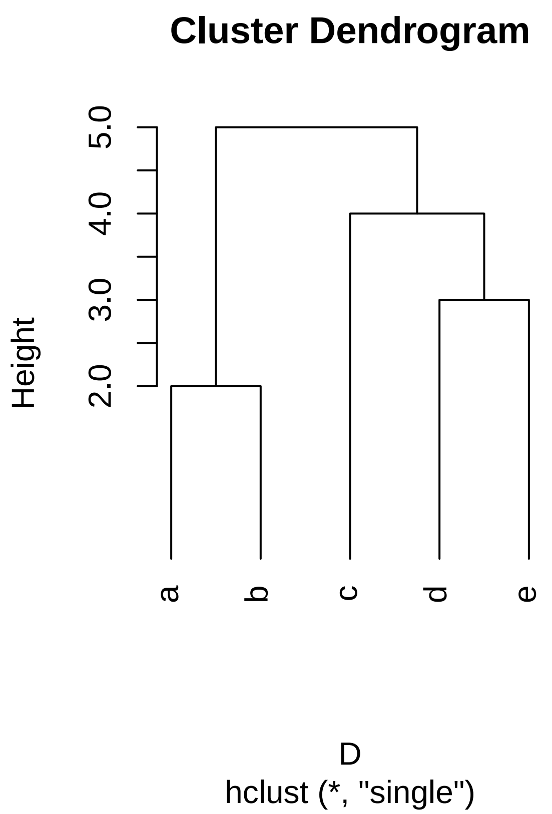
Prova scritta - fila A

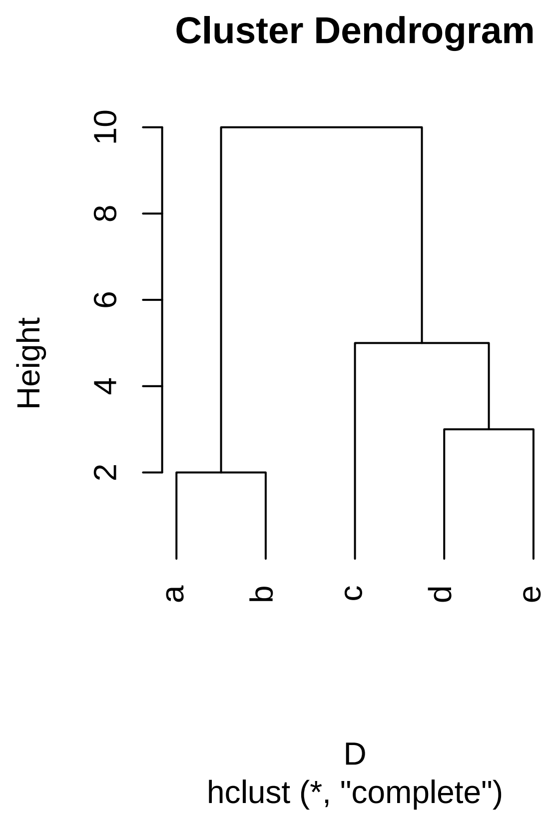
Si svolgano gli esercizi riportando il risultato dove indicato. Durata: 60 minuti

---

Esercizio 1 (Punti 3)

```
rm(list=ls())
M = matrix(
c(0,2,6,10,9,
  2,0,5,9,8,
  6,5,0,4,5,
  10,9,4,0,3,
  9,8,5,3,0),byrow=T, ncol=5
)
colnames(M)<-c("a","b","c","d","e")
D = as.dist(M)
op <- par(mfrow = c(1, 2))
plot(hclust(D,"single"), hang=-1)
plot(hclust(D,"complete"), hang=-1)
```






```r
par(op)
```

Sulla base dei due dendogrammi sopra riportati, completare la seguente matrice di distanze che gli ha generati.

	a	b	c	d	e
a	0				
b	...	0			
c	6	5	0		
d	...	9	4	0	
e	9	8	5	...	0

Esercizio 2 (Punti 3)

Quali dei seguenti vettori sono ortogonali?

$$x = \begin{pmatrix} 1 \\ -2 \\ 3 \\ -4 \end{pmatrix} \quad y = \begin{pmatrix} 6 \\ 7 \\ 1 \\ -2 \end{pmatrix} \quad z = \begin{pmatrix} 5 \\ -4 \\ 5 \\ 7 \end{pmatrix}$$

```r
rm(list=ls())
x = matrix(c(1,-2,3,-4), ncol=1)
y = matrix(c(6,7,1,-2), ncol=1)
z = matrix(c(5,-4,5,7), ncol=1)
round(c(x)/sqrt( t(x)%*%x ),2)
```

```
[1]  0.18 -0.37  0.55 -0.73
```

```r
round(c(z)/sqrt( t(z)%*%z ),2)
```

```
[1]  0.47 -0.37  0.47  0.65
```

Riportare la versione normalizzata dei due vettori ortogonali (arrotondare al secondo decimale).

Esercizio 3 (Punti 9)

Si consideri il dataset *USArrests* presente nella libreria *datasets*. Per ciascuno dei 50 stati degli USA, l'insieme di dati contiene il numero di arresti per 100000 residenti per ognuno dei tre reati: Rapina (*Assault*), Omicidio (*Murder*) e Stupro (*Rape*). La variabile *UrbanPop* indica la percentuale di popolazione nelle aree urbane. Sia $\underset{50\times 4}{X}$ la matrice dei dati corrispondente al dataset *USArrests*.

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 8, specificando anche il nome della riga di X (lo stato) a cui si fa riferimento.


```r
rm(list=ls())
X <- USArrests
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 >  8],2)
```

```
        Alaska        Georgia         Nevada North Carolina   Rhode Island 
         15.48           9.75           8.31          12.87           9.98 
```

b. Sulla base della matrice dei dati standardizzati $\underset{50\times 4}{Z}$, 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 $K = 2, 3, \ldots, 7$
   - 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	2	3	4	5	6	7
$CH(K)$						
silhouette(K)						

```
Z = scale(X, center=T, scale=diag(S)^(1/2))
D = dist(Z, method = "euclidean")
K = 2:7
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    2.00  3.00  4.00  5.00  6.00  7.00
CH  43.46 30.62 37.95 32.22 28.12 29.50
sil  0.41  0.28  0.34  0.34  0.27  0.29
```

c. Determinare l'appartenenza di ciascuna osservazione a $K = 2$ gruppi utilizzando

- il metodo gerarchico agglomerativo con funzione di legame completo considerando come matrice delle distanze quella ottenuta con la metrica di Manhattan basata su Z;
- il metodo delle delle K medie (`algorithm = "Hartigan-Wong"` inizializzando i K centrodi utilizzando le prime K osservazioni) applicato alla matrice dei punteggi (*scores*) $\underset{50\times 2}{Y}$ ottenuta dalle prime due componenti principali di Z.

```
Dlag = dist(Z, method = "manhattan")
hc = hclust(Dlag, method="complete")
clusterhc = cutree(hc, k=2)
Y = princomp(Z)$scores[,1:2]
km = kmeans(Y, centers=Y[1:2,], algorithm ="Hartigan-Wong")
clusterkm = km$cluster
table(clusterhc)
```

```
clusterhc
 1  2
19 31
```

```
table(clusterkm)
```

```
clusterkm
 1  2
30 20
```

```
table(clusterhc,clusterkm)
```

```
         clusterkm
clusterhc  1  2
        1  0 19
        2 30  1
```

```
rownames(X)[(clusterhc ==2 & clusterkm ==2)]
```

```
[1] "Missouri"
```

Riportare il numero delle osservazioni classificate nei cluster 1 e 2 secondo i due approcci (gerarchico e K-medie)

Approccio	n.ro osservazioni cluster 1	n.ro osservazioni cluster 2
Gerarchico	...	...
K-medie	...	...

Riportare i valori della tabella a doppia entrata che incrocia la classificazione ottenuta con l'approccio gerarchico e quello delle K-medie

Gerarchico / K-medie	cluster 1	cluster 2
cluster 1	...	...
cluster 2	...	...

Riportare il nome dell'unico stato classificato nel cluster 2 da entrambi gli approcci.

d. Stimare il modello fattoriale con $k = 1$ 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 $T = n \log\left(\frac{det(\hat{\Lambda}\hat{\Lambda}' + \hat{\Psi})}{det(R)}\right)$ (arrotondando al terzo decimale)

```
R = cor(X)
af = factanal(Z,factors=1, rotation="none", method="mle")
Lambda = af$loadings[,]
Psi = diag(af$uniqueness)
fit = Lambda %*% t(Lambda) + Psi
lrt = n*log(det(fit)/det(R))
round(lrt,3)
```

```
[1] 9.908
```

Esercizio 4 (Punti 3)

Si consideri il modello fattoriale con 1 fattore:

$$
\begin{aligned}
z_1 &= \lambda_1 f + u_1 \\
z_2 &= \lambda_2 f + u_2 \\
z_3 &= \lambda_3 f + u_3
\end{aligned}
$$

dove $\widehat{\mathbb{C}ov}(z) = \underset{3\times 3}{R} = \begin{bmatrix} 1 & 0.25 & 0.25 \\ & 1 & 0.25 \\ & & 1 \end{bmatrix}$.

Riportare le stime $\hat{\Lambda}$ e $\hat{\Psi}$ utilizzando il metodo di stima *naive*.

```
rm(list=ls())
lambda1 = sqrt(0.25*0.25/0.25)
lambda2 = sqrt(0.25*0.25/0.25)
lambda3 = sqrt(0.25*0.25/0.25)
Lambda = matrix(c(lambda1,lambda2,lambda3), ncol=1)
Psi = diag(1-c(lambda1,lambda2,lambda3)^2)
round(Lambda,2)
```

```
     [,1]
[1,]  0.5
[2,]  0.5
[3,]  0.5
```

```
round(Psi,2)
```

```
     [,1] [,2] [,3]
[1,] 0.75 0.00 0.00
[2,] 0.00 0.75 0.00
[3,] 0.00 0.00 0.75
```

Esercizio 5 (Punti 3)

Sulla base la matrice dei dati centrati $\underset{n\times 3}{\tilde{X}} = [\underset{n\times 1}{\tilde{x}_1}\ \underset{n\times 1}{\tilde{x}_2}\ \underset{n\times 1}{\tilde{x}_3}]$ è stata calcolata la seguente matrice di varianze/covarianze:

$$\underset{3\times 3}{S} = \begin{bmatrix} 2 & 0 & 0 \\ & 3 & 0 \\ & & 4 \end{bmatrix}$$

Si determinino i punteggi delle 3 componente principali:

$\underset{n\times 1}{y_1} =$

$\underset{n\times 1}{y_2} =$

$\underset{n\times 1}{y_3} =$

```
rm(list=ls())
S = diag(c(2,3,4))
princomp(covmat=S)$loadings
```

```

Loadings:
     Comp.1 Comp.2 Comp.3
[1,]               1
[2,]        1
[3,] 1

               Comp.1 Comp.2 Comp.3
```

```
SS loadings     1.000  1.000  1.000
Proportion Var  0.333  0.333  0.333
Cumulative Var  0.333  0.667  1.000
```

Esercizio 6 (Punti 5)

Dimostrare, esplicitando tutti i passaggi e le quantità coinvolte, che

a. $\det(S^Y) = \det(S)$ dove $Y = \tilde{X}V$ e le colonne di V sono gli autovettori normalizzati di S

b. nel modello fattoriale a k fattori, $\mathbb{E}(\underset{p\times 1}{x}\underset{1\times k}{f'}) = \underset{p\times k}{\Lambda}$