6 Febbraio 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: 80 minuti

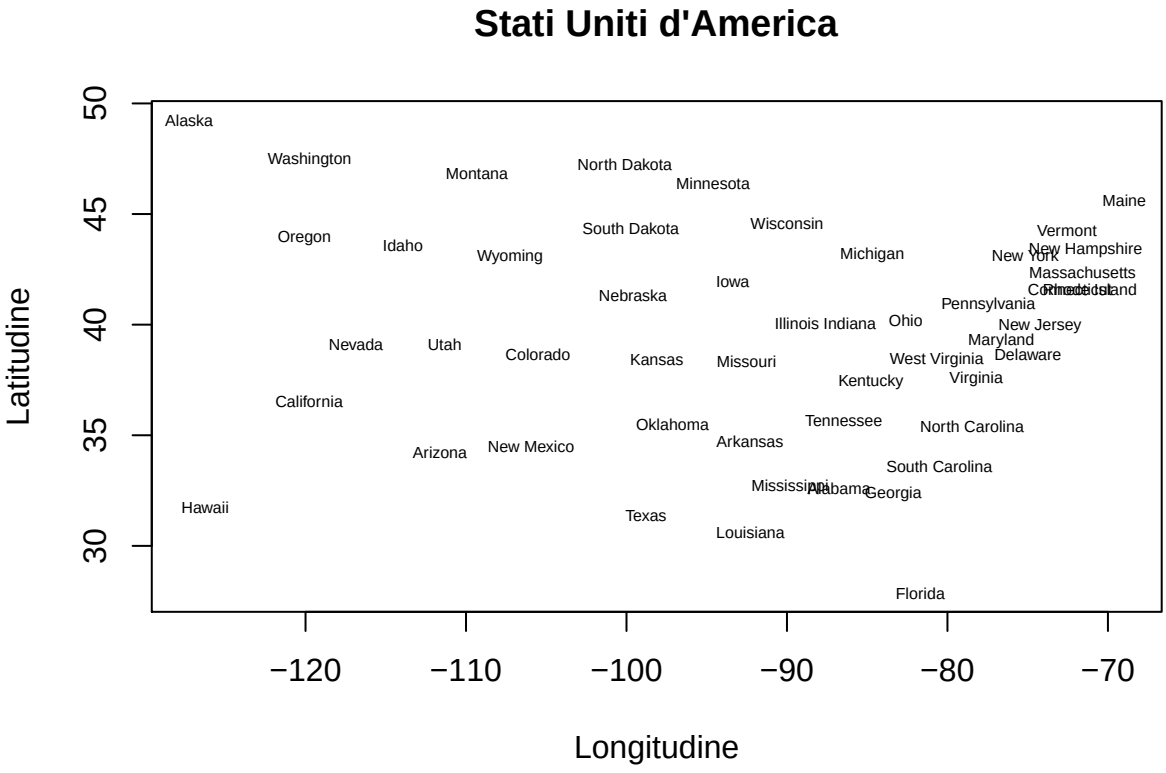
---

Esercizio 1 (Punti 15)

Il dataset `state.x77` presente nella libreria `datasets` descrive 50 stati degli Stati Uniti d'America rispetto alle seguenti 8 variabili:

- `Population` in migliaia
- `Income` in dollari pro capita
- `Illiterarcy` Percentuale della popolazione
- `Life Exp` Anni di aspettativa di vita alla nascita
- `Murder` Numero di omicidi e omicidi colposo per 100000 persone
- `HS Grad` Percentuale di adulti diplomati
- `Frost` Numero medio di giorni freddi all'anno con temperature sotto lo zero
- `Area` in miglia quadrate

Inoltre il dataset `state.center` (anch'esso presente nella libreria `datasets`) riporta la longitudine (con segno negativo) e la latitudine del centro geografico di ogni stato (tranne che per l'Alaska e le Hawaii, che sono messe artificialmente da qualche parte a ovest della costa), come illustrato nella seguente Figura:



a. Sia X la matrice 50×8 corrispondente al dataset `state.x77`. Sulla base della corrispondente matrice di varianze/covarianze S, svolgere l'analisi delle componenti principali e riportare la percentuale di varianza spiegata dalla prima componente principale.

```
rm(list=ls())
X = state.x77
n = nrow(X)
p = ncol(X)
summary(prcomp(X))$importance[3,1]
```

```
[1] 0.99723
```

b. Riportare le varianze delle variabili presenti in X, arrotondando al primo decimale.

```
round(apply(X,2,var)*((n-1)/n),1)
```

```
  Population       Income   Illiteracy     Life Exp       Murder
  19533050.1     370021.8          0.4          1.8         13.4
     HS Grad        Frost         Area
        63.9       2648.0 7135133099.6
```

Alla luce dei risultati sopra ottenuti, indicare quali sono le problematiche per l'analisi delle componenti principali svolta al punto a.

c. Svolgere l'analisi delle componenti principali sui dati standardizzati Z, riportando
- la media c delle percentuali di varianza spiegata da ciascuna componente principale
- il numero di componenti principali con varianza spiegata superiore a c

```
R = cor(X)
c = mean(eigen(R)$values/p)
c
```

```
[1] 0.125
```

```
sum(eigen(R)$values/p > c)
```

```
[1] 3
```

d. Si consideri la matrice dei dati Y di dimensioni 50×11 dove le prime 8 colonne sono uguali a quelle di X mentre le restanti 3 colonne sono le nuove variabili

- `Longitude`, ricavabile dal dataset `state.center`
- `Latitude`, ricavabile dal dataset `state.center`
- `Density` = `Population` / `Area` ricavabile dal dataset `state.x77`

Sia Q la matrice dei dati standardizzati ottenuta a partire da Y. Si svolga l'analisi delle componenti principali basata su Q (ovvero sulla base della matrice di correlazione R^Y), riportando i punteggi (*scores*) dello stato dell'Alaska relativamente alle prime tre componenti principali (arrotondati alla terza cifra decimale)

```
Y = cbind(X,state.center$x,state.center$y,state.x77[,"Population"]/state.x77[,"Area"])
colnames(Y) = c(colnames(state.x77),"Longitude",
"Latitude", "Density")
round(princomp(Y, cor=T)$scores["Alaska",1:3],3)
```

```
Comp.1 Comp.2 Comp.3
-2.370 -5.715  1.519
```

e. La variabile `Density` costruita al punto precedente è funzione delle variabili `Population` e `Area`. Questo comporta che la matrice Y ha colonne linearmente dipendenti? Giustificare la risposta.

```
qr(Y)$rank
```

```
[1] 11
```

f. Si consideri la stima di massima verosimiglianza per il modello fattoriale con k fattori basato sui dati standardizzati Q. 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".

```
k = which.max(sapply(1:6, function(k) factanal(Y,factors=k)$PVAL > 0.05 ))
k
```

```
objective
        5
```

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

```
objective
   0.1016
```

g. Stimare il modello fattoriale con $k = 5$ fattori con il metodo della massima verosimiglianza utilizzando i dati standardizzati Q e senza effettuare alcuna rotazione. Riportare le "stime" dei punteggi fattoriali con il metodo di Thomson per lo stato dell'Alaska (arrotondando al secondo decimale)

```
round(factanal(Y,factors=5, scores="regression", rotation="none", method="mle")$scores["Alaska",],2)
```

```
Factor1 Factor2 Factor3 Factor4 Factor5
  -0.80    3.42    2.16    3.65    1.04
```

h. Applicare l'algoritmo delle K medie (`algorithm = "Hartigan-Wong"`) per i dati standardizzati Q inizializzando i K centrodi utilizzando le prime K osservazioni (righe $1, \ldots, K$ della matrice dei dati Q). Arrotondando il risultato alla seconda cifra decimale, riportare per $K = 2, \ldots, 8$
   - il valore dell'indice $\text{CH}(K) = \frac{B/(K-1)}{W/(n-K)}$ di Calinski and Harabasz
   - il valore medio della *silhouette* considerando come matrice delle distanze quella ottenuta con la metrica Euclidea basata su Q

K	2	3	4	5	6	7	8
$CH(K)$							
silhouette(K)							

```
varY = apply(Y,2,var)*((n-1)/n)
W = scale(Y,center=TRUE, scale=sqrt(varY))
D = dist(W, method="euclidean")
K = 2:8
CH <- vector()
sil <- vector()
library(cluster)
for (k in 1:length(K)){
km = kmeans(W, centers=W[1:K[k],])
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]  [,7]
K   2.00  3.00  4.00  5.00  6.00  7.00  8.00
CH  6.07 16.14 17.11 18.26 16.41 16.67 16.02
sil 0.49  0.30  0.26  0.30  0.25  0.25  0.25
```

Esercizio 2 (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 $\widehat{\mathbb{C}ov}(z) = \underset{3\times 3}{R} = \begin{bmatrix} 1 & 0.5 & 0.6 \\ & 1 & 0.7 \\ & & 1 \end{bmatrix}$.

Arrontondando il risultato al secondo decimale, riportare le stime $\hat{\Lambda}$ e $\hat{\Psi}$ utilizzando il metodo di stima *naive*.

```
rm(list=ls())
lambda1 = sqrt(0.5*0.6/0.7)
lambda2 = sqrt(0.5*0.7/0.6)
lambda3 = sqrt(0.6*0.7/0.5)
Lambda = matrix(c(lambda1,lambda2,lambda3), ncol=1)
Psi = diag(1-c(lambda1,lambda2,lambda3)^2)
round(Lambda,2)
```

```
     [,1]
[1,] 0.65
[2,] 0.76
[3,] 0.92
```

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

```
     [,1] [,2] [,3]
[1,] 0.57 0.00 0.00
[2,] 0.00 0.42 0.00
[3,] 0.00 0.00 0.16
```

Esercizio 3 (Punti 3)

Alla matrice di varianze/covarianze $\underset{p\times p}{S}$ sono associati i seguenti autovalori $\lambda_1 = 6, \lambda_2 = 4$ e autovettori normalizzati $\underset{2\times 1}{v_1} = \begin{bmatrix} 1/\sqrt{5} \\ 2/\sqrt{5} \end{bmatrix}$, $\underset{2\times 1}{v_2} = \begin{bmatrix} 2/\sqrt{5} \\ -1/\sqrt{5} \end{bmatrix}$.

a. Riportare la matrice di correlazione $\underset{p\times p}{R}$ e la matrice $\underset{p\times p}{S^{2/3}}$ (arrotondando i risultati al secondo decimale):

b. Calcolare la correlazione tra la prima colonna $\underset{n\times 1}{\tilde{x}_1}$ di $\underset{n\times p}{\tilde{X}}$ e i punteggi $\underset{n\times 1}{y_1}$ della prima componente principale, arrotondando il risultato al secondo decimale:

```
rm(list=ls())
Lambda = diag(c(6,4))
V = matrix(c(1/sqrt(5),2/sqrt(5),2/sqrt(5),-1/sqrt(5)), byrow=F, ncol=2)
S = V %*% Lambda %*% t(V)
# a.
R = diag(diag(S)^(-1/2)) %*% S %*% diag(diag(S)^(-1/2))
round(R,2)
```

```
     [,1] [,2]
```

```
[1,] 1.00 0.16
[2,] 0.16 1.00
```

```
round(V %*% Lambda^(2/3) %*% t(V),2)
```

```
     [,1] [,2]
[1,] 2.68 0.31
[2,] 0.31 3.15
```

```
# b
round( V[1,1]*sqrt(Lambda[1,1])/sqrt(S[1,1]) , 2)
```

```
[1] 0.52
```

Esercizio 4 (punti 5)

Dimostrare, esplicitando tutti i passaggi, e specificando tutte le quantità coinvolte,

a. $d_m(y_i, y_l) = d_m(x_i, x_l)$ dove d_m è la distanza di Minkowski di ordine $m \geq 1$, $\underset{1\times p}{y_i'} = \underset{1\times p}{x_i'} + \underset{1\times p}{b}$ e $\underset{1\times p}{y_l'} = \underset{1\times p}{x_l'} + \underset{1\times p}{b}$

b. se $\det(S) = 0$, allora le colonne di $\underset{n\times p}{\tilde{X}}$ sono linearmente dipendenti;