21 Febbraio 2020 - Analisi Esplorativa (Analisi Statistica Multivariata)

Cognome: ....

Nome: ....

Matricola: ....

Prova scritta

Si svolgano gli esercizi riportando il risultato dove indicato. Durata: 90 minuti

---

Esercizio 1 (Punti 5)

Sia R la matrice di varianze e covarianze dei dati standardizzati $\underset{n\times p}{Z}$: $R = \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}$ con $r \in (-1, 0)$.

a. (Punti 1) Determinare gli autovalori λ_1 e λ_2 di R (rispettando $\lambda_1 > \lambda_2$)

b. (Punti 1) Determinare gli autovettori normalizzati $\underset{p\times 1}{v_1}$ e $\underset{p\times 1}{v_2}$ di R corrispondenti a λ_1 e λ_2

c. (Punti 2) Determinare il vettore dei punteggi $\underset{n\times 1}{y_1}$ della prima componente principale di Z

d. (Punti 1) Determinare la proporzione di varianza spiegata dalla prima componente principale.

Esercizio 2 (Punti 3)

Sia S la matrice di varianze e covarianze dei dati centrati $\underset{n\times p}{\tilde{X}}$: $S = \begin{bmatrix} 5 & 2 \\ 2 & 2 \end{bmatrix}$.

a. (Punti 1) Determinare la proporzione di varianza spiegata dalla prima componente principale di $\underset{n\times p}{\tilde{X}}$, arrotondando alla sesta cifra decimale.

```
rm(list=ls())
S = matrix(c(5,2,2,2), byrow=T, ncol=2)
round(eigen(S)$value[1]/sum(eigen(S)$value),6)
```

```
[1] 0.857143
```

b. (Punti 1) Determinare la proporzione di varianza spiegata dalla prima componente principale di $\underset{n\times p}{Z}$ (dati standardizzati), arrotondando alla sesta cifra decimale.

```
R = diag(diag(S)^(-1/2)) %*% S %*% diag(diag(S)^(-1/2))
round(eigen(R)$value[1]/sum(eigen(R)$value),6)
```

```
[1] 0.816228
```

c. (Punti 1) Si calcoli la correlazione tra $\underset{n\times 1}{z_k}$ (j-sima colonna dei dati standardizzati $\underset{n\times p}{Z}$) e i punteggi della seconda componente principale di $\underset{n\times p}{Z}$ per $j = 1, \ldots, p$, arrotondando alla sesta cifra decimale.

```
v2 = eigen(R)$vector[,2]
lambda2 = eigen(R)$value[2]
round(v2*sqrt(lambda2),6)
```

```
[1] -0.428687  0.428687
```

Esercizio 3 (Punti 3)

Si consideri la seguente matrice dei dati:

Presidente	Luogo di Nascita	Eletto	Partito	Esperienze pregresse al congresso	Vicepresidente
Nixon	ovest	si	rep.	si	si
Kennedy	est	si	dem.	si	no
Johnson	sud	no	dem.	si	si

Si definiscano le seguenti variabili binarie:

- $X_1 = 1$ se Luogo di Nascita = sud, 0 altrimenti
- $X_2 = 1$ se Eletto = si, 0 altrimenti
- $X_3 = 1$ se Partito = rep., 0 altrimenti
- $X_4 = 1$ se Esperienze pregresse al congresso = si, 0 altrimenti
- $X_5 = 1$ se Vicepresidente = si, 0 altrimenti

a. (Punti 1) Quali tra le variabili $X_1, \ldots, X_5$ sono variabili binarie asimmetriche?

a. (Punti 2) Calcolare l'indice di corrispondenza semplice s_c e quello di Jaccard s_J per i presidenti (i) Nixon e Johnson (ii) Nixon e Kennedy

```r
X = matrix(c(0,1,1,1,1,
             0,1,0,1,0,
             1,0,0,1,1),byrow=T,ncol=5)
# Nixon e Johnson
tab = table(X[1,],X[3,])
d = tab[1,1]
a = tab[2,2]
p = sum(tab)
( s_c = (a + d)/p )
```

```
[1] 0.4
```

```r
( s_J = (a)/(p-d) )
```

```
[1] 0.4
```

```r
# Nixon e Kennedy
tab = table(X[1,],X[2,])
d = tab[1,1]
a = tab[2,2]
p = sum(tab)
( s_c = (a + d)/p )
```

```
[1] 0.6
```

```r
( s_J = (a)/(p-d) )
```

```
[1] 0.5
```

Esercizio 4 (Punti 5)

Sia S la matrice di varianze e covarianze dei dati centrati $\underset{n\times p}{\tilde{X}}$: $S = \begin{bmatrix} a & 0 & 0 & 0 \\ 0 & b & 0 & 0 \\ 0 & 0 & c & 0 \\ 0 & 0 & 0 & d \end{bmatrix}$ dove $a > b > c > d > 0$.

a. (Punti 1) Determinare gli autovalori $\lambda_1, \ldots, \lambda_p$ di S.

b. (Punti 2) Determinare la matrice $\underset{p\times p}{V}$ degli autovalori normalizzati di S.

c. (Punti 1) Determinare la matrice dei punteggi $\underset{n\times p}{Y} = \tilde{X}V$ delle componenti principali.

d. (Punti 1) Determinare la percentuale di varianza spiegata dalla prima componente principale.

Esercizio 5 (Punti 4)

Si supponga che la matrice dei dati $\underset{n\times 3}{X}$ sia tale che la seconda colonna sia pari a 2 volte la prima, i.e. $\underset{n\times 1}{x_2} = 2\underset{n\times 1}{x_1}$, e la terza colonna sia pari a 2 volte la seconda, $\underset{n\times 1}{x_3} = 2\underset{n\times 1}{x_2}$.

a. (Punti 1) Determinare la matrice di correlazione R di X

```
R = matrix(c(1,1,1,1,1,1,1,1,1), byrow=T, ncol=3)
```

b. (Punti 2) Determinare gli autovalori $\lambda_1, \ldots, \lambda_p$ di R

c. (Punti 1) Qual è la percentuale di varianza spiegata dalla prima componente principale?

Esercizio 6 (Punti 5)

Si consideri il dataset `quakes` presente nella libreria `datasets`, che contiene $n = 1000$ osservazioni (eventi sismici) su cui sono state misurate le seguenti 5 variabili:

- *lat* latitudine dell'evento sismico
- *long* longitudine dell'evento sismico
- *depth* profondità (in km) dell'evento sismico
- *mag* magnitudo (scala Richter)
- *stations* Numero di stazioni che hanno riportato l'evento sismico

a. (Punti 1) Si consideri la matrice $\underset{1000\times 5}{X}$ che contiene le seguenti variabili: *lat*, *long*, *depth*, *mag* e *stations*. Si costruisca il diagramma a scatola con baffi (*boxplot*) per ciascuna delle variabili presenti in $\underset{1000\times 5}{X}$ e si riporti il numero di valori anomali (*outliers*).

	lat	long	depth	mag	stations
numero di valori anomali					

```r
rm(list=ls())
X = quakes
apply(X,2,function(x) length(boxplot.stats(x)$out))
```

```
     lat     long    depth      mag stations 
      32      204        0        7       54 
```

b. (Punti 2) Per la matrice $\underset{1000\times 5}{X}$ calcolata al punto a., si calcoli il quadrato della distanza di Mahalanobis di ciascuna unità statistica u_i' dal baricentro $\bar{x}'$ e si riporti il valore minimo e il valore massimo, arrotondando i calcoli al <u>secondo decimale</u>.

```r
n = nrow(X)
# vettore medie
xbar = matrix(colMeans(X), ncol=1)
S = var(X)*((n-1)/n)
# matrice inversa
InvS = solve(S)
# quadrato della distanza di Mahalanobis per le n osservazioni
dM2 = apply(X,1, function(u) t(u-xbar) %*% InvS %*% (u - xbar) )
# valore minimo e massimo delle distanze di Mahalanobis al quadrato
round( min(dM2) , 2)
```

```
[1] 0.57
```

```r
round( max(dM2) , 2)
```

```
[1] 25.9
```

$\min_{i=1,\ldots,1000}\{d^2_M(u_i,\bar{x})\} = \ldots\ldots\ldots\ldots$ $\max_{i=1,\ldots,1000}\{d^2_M(u_i,\bar{x})\} = \ldots\ldots\ldots\ldots$

c. (Punti 1) Utilizzare l'algoritmo delle K-medie (specificando `algorithm = "Lloyd"`) per formare $K = 4$ gruppi sulla base della matrice dei dati standardizzati $\underset{1000\times 5}{Z}$ ottenuta a partire da $\underset{1000\times 5}{X}$, inizializzando i centroidi con le osservazioni di riga 200, 400, 600 e 800, ed eseguendo l'algoritmo una sola volta. Riportare i valori dei centroidi dei 4 gruppi ottenuti, arrotondando alla seconda cifra decimale.

```r
# kmeans
Z <- scale(X, center=T, scale = sqrt(diag(S)))
km = kmeans(Z, centers=Z[c(200,400,600,800),], algorithm = "Lloyd")
round( km$centers, 2)
```

```
    lat  long depth   mag stations
1 -0.15  0.23  0.02  1.61     1.82
2  0.95 -1.89 -0.76  0.22    -0.10
3 -0.51  0.70 -0.79 -0.26    -0.34
4  0.01  0.25  1.03 -0.59    -0.46
```

	Centroidi				
	lat	long	depth	mag	stations
Gruppo 1					
Gruppo 2					
Gruppo 3					
Gruppo 4					

d. (Punti 1) Calcolare, arrotondando al secondo decimale, l'indice di Calinski and Harabasz per i quattro gruppi individuati al punto c.

```
# indice di Calinski and Harabasz
W = km$tot.withinss
B = km$betweenss
K = 4
( CH = (B/(K-1)) / (W/(n-K)) )
```

```
[1] 474.8148
```

Indice di Calinski and Harabasz =