Cognome: .... Nome: .... Matricola: ....

Tipologia d'esame: ☐ 12 CFU ☐ 15 CFU

Prova scritta di ASM 12CFU e 15CFU - Modulo Analisi Esplorativa del 11.07.2017

La durata della prova è di 60 minuti.
Si svolgano gli esercizi 1, 2 e 3 riportando il risultato dove indicato.

Esercizio 1 (Punti: 8)

1.a) Siano date due unità statistiche $u_1' = (1,1)$ e $u_2' = (2,3)$. Calcolare

Distanza Euclidea $d_2(u_1, u_2) = \ldots\ldots$

Distanza di Manhattan $d_1(u_1, u_2) = \ldots\ldots$

Distanza di Lagrange $d_\infty(u_1, u_2) = \ldots\ldots$

Distanza di Canberra $d_C(u_1, u_2) = \ldots\ldots$

1.b) Si consideri la seguente matrice di distanze relativa a tre unità statistiche u_1, u_2 e u_3:

$\underset{3\times 3}{D} =$

	u_1	u_2	u_3
u_1	0		
u_2	7	0	
u_3	6	5	0

Se utilizziamo un algoritmo gerarchico agglomerativo, le unità u_1 e u_2 vengono messe assieme nel gruppo (u_1, u_2). Aggiornare la matrice delle distanze utilizzando il metodo del legame singolo:

	(u_1, u_2)	(u_3)
(u_1, u_2)	0	
(u_3)	...	0

Aggiornare la matrice delle distanze utilizzando il metodo del legame medio:

	(u_1, u_2)	(u_3)
(u_1, u_2)	0	
(u_3)	...	0

1.c) Si consideri la seguente decomposizione della distanza totale T in distanza entro i gruppi W e tra i gruppi B per 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)	2	...	12
(u_1, u_2), (u_3)	...	10.5	12
(u_1, u_3), (u_2)	...	8.5	...

Completare la tabella con le informazioni mancanti.

Qual è il raggruppamento ottimale? $G_1 = \ldots\ldots\ldots$, $G_2 = \ldots\ldots\ldots$

Esercizio 2 (Punti: 8)

Data la matrice di varianze/covarianze $\underset{3\times 3}{S} = \begin{bmatrix} 4 & 3 & 1 \\ 3 & 9 & 2 \\ 1 & 2 & 1 \end{bmatrix}$

2.a) Riportare

Varianza totale = ... e generalizzata = ...

L'indice di variabilità relativo (<u>arrotondare al secondo decimale</u>) = ...

2.b) Determinare la matrice di correlazione $\underset{3\times 3}{R} = \begin{bmatrix} \ldots\ldots & \ldots\ldots & \ldots\ldots \\ \ldots\ldots & \ldots\ldots & \ldots\ldots \\ \ldots\ldots & \ldots\ldots & \ldots\ldots \end{bmatrix}$.

2.c) Determinare, <u>arrotondando al secondo decimale</u>, $\underset{p\times p}{S^{1/2}} = \begin{bmatrix} \ldots\ldots & \ldots\ldots \\ \ldots\ldots & \ldots\ldots \end{bmatrix}$

2.d) Calcolare la correlazione tra la seconda colonna $\underset{n\times 1}{\tilde{x}_2}$ di $\underset{n\times p}{\tilde{X}}$ e i punteggi $\underset{n\times 1}{y_2}$ della seconda componente principale, <u>arrotondando al secondo decimale</u>.

= ....

```
## [1] 14
```

```
## [1] 14
```

```
## [1] 0.39
```

```
##      [,1]      [,2]      [,3]
## [1,]  1.0 0.5000000 0.5000000
## [2,]  0.5 1.0000000 0.6666667
## [3,]  0.5 0.6666667 1.0000000
```

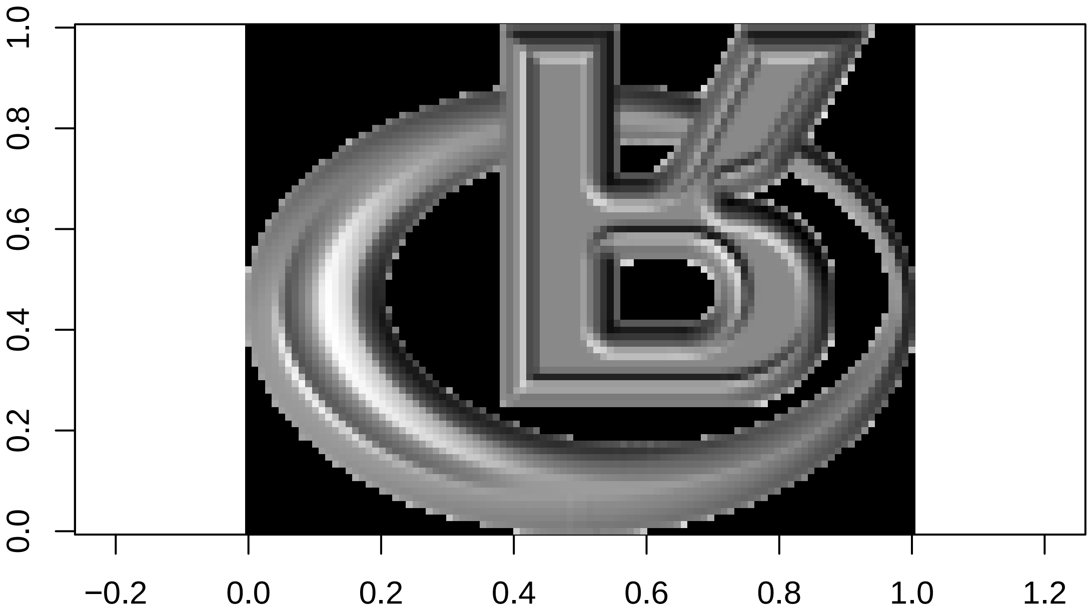
```
##      [,1] [,2] [,3]
## [1,] 1.89  0.6 0.26
## [2,] 0.60  2.9 0.50
## [3,] 0.26  0.5 0.83
```

```
## [1] -0.23
```

Esercizio 3 (Punti: 9)

Importare il logo di R nella matrice $\underset{n\times p}{\tilde{X}}$ con $n = 100$ e $p = 76$ tramite i comandi

```
rm(list=ls())
library(png)
logo <- readPNG(system.file("img", "Rlogo.png", package="png"))
X <- t(logo[,,1])
n = nrow(X)
p = ncol(X)
# visualizza immagine
image(X, col=gray(0:255/255), asp=p/n)
```



3.a) Effettuare l'analisi delle componenti principali dei dati centrati $\underset{n\times p}{\tilde{X}}$ (utilizzare il comando `princomp`), riportando il numero delle componenti necessarie per spiegare almeno 85% della variabilità:

NUMERO DI COMPONENTI = ... PROPORZIONE DI VARIABILITA' SPIEGATA = ...

```
## Comp.8
##      8
```

```
##    Comp.8
## 0.8636509
```

3.b) Dopo aver calcolato la matrice dei punteggi $\underset{n\times p}{Y} = \underset{n\times p}{\tilde{X}}\underset{p\times p}{V}$ e la matrice dei pesi $\underset{p\times p}{V}$, ottenere la migliore approssimazione per $\underset{n\times p}{\tilde{X}}$ di rango q, $\underset{n\times p}{A_q} = \underset{n\times q}{Y_q}\underset{q\times p}{V_q'}$, con q determinato al punto precedente (q = numero delle componenti necessarie per spiegare almeno 85% della variabilità). Costruire l'immagine compressa $\underset{n\times p}{C} = \underset{n\times p}{A_q} + \underset{n\times 1}{1}\underset{1\times p}{\bar{x}}'$, assicurandosi che tutti gli elementi di $\underset{n\times p}{C}$ siano compresi tra 0 e 1. Riportare infine i 5 numeri di sintesi per l'immagine compressa:

MEDIA = ...

MEDIANA = ...

1 QUARTILE = ...

3 QUARTILE = ...

Min = ...

Max = ...

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.03987 0.31959 0.30156 0.50744 0.98670
```

3.c) Confrontare l'immagine compressa con l'immagine originale calcolando il fattore di riduzione in termini di pixels e bytes (utilizzando il comando `object.size`)

FATTORE DI RIDUZIONE (PIXELS) = ...

FATTORE DI RIDUZIONE (BYTES) = ...

```
## [1] 5.12
## 4.4 bytes
```