Departamento de Matemática

Multivariate Analysis

Mater in Eng. and Data Science & Master in Mathematics and Applications

1^{st} Test
Duration: 1.5 hours

1^{st} Semester – 2020/2021
19/11/2020 – 20:00

Please justify conveniently your answers

Group I **10.0 points**

1. Let X_1, X_2, and X_3 be three independent, univariate Normal distributed random variables, with unitary mean and variance. Let $\boldsymbol{Y} = (Y_1, Y_2, Y_3)^t$, where $Y_1 = X_1 - 3X_2 + 2$, $Y_2 = 2X_1 - X_2 - 1$, and $Y_3 = X_3 - 1$. Determine the distribution of $\boldsymbol{Y}$. (2.5)

2. Suppose $\boldsymbol{X} = (X_1, X_2, X_3)^t$ has a multivariate normal distribution. Show that if: (2.5)
(i) X_1 and $X_2 + X_3$ are independent,
(ii) X_2 and $X_1 + X_3$ are independent, and
(iii) X_3 and $X_1 + X_2$ are independent,
then X_1, X_2, and X_3 are independent random variables.

3. Let $\boldsymbol{X} = (X_1, X_2, X_3, X_4)^t$ be a random vector with multivariate normal distribution with parameters: (2.5)

$$\boldsymbol{\mu} = \begin{pmatrix} 1 \\ 2 \\ 1 \\ 3 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{pmatrix} 2 & 1 & 2 & 3 \\ & 2 & 1 & 3 \\ & & 3 & 3 \\ & & & 7 \end{pmatrix}$$

Determine the distribution of $(X_3, X_4)^t | (X_1, X_2)^t = (x_1, x_2)^t$.

Suggestion: If $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then $E(\boldsymbol{X}_1 | \boldsymbol{X}_2 = \boldsymbol{x}_2) = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)$ and $Var(\boldsymbol{X}_1 | \boldsymbol{X}_2 = \boldsymbol{x}_2) = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$.

4. Let $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ be two independent random vectors, where $\boldsymbol{X}_i \sim \mathcal{N}_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, $i = 1, 2$. Consider two independent random samples, with sizes n_1 e n_2 from each population. Prove that (2.5)

$$\frac{n_1 n_2}{n_1 + n_2}(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2))^t \boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)) \sim \chi^2_{(p)}.$$

Group II **10.0 points**

The U.S. crime data set consists of the reported number of crimes in the 50 U.S. states in 1985. The crimes were classified according to 7 categories:

- X_1: murder,
- X_2: rape,
- X_3: robbery,
- X_4: assault
- X_5: burglary,
- X_6: larceny,
- X_7: auto theft.

```
xx<-princomp(USCrime, cor = TRUE, scores = TRUE)
summary(xx)
Importance of components:
                           Comp.1    Comp.2     Comp.3    Comp.4     Comp.5     Comp.6     Comp.7
Standard deviation     1.9775734          a 0.82235719 0.6095176 0.51509703 0.44367858 0.34052323

round(xx$loadings,3)
     Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
murd  0.272  0.653  0.023  0.245  0.310  0.009  0.586
rape  0.431  0.117  0.376 -0.061 -0.617 -0.523  0.015
robb  0.376 -0.051 -0.662 -0.613  0.025 -0.131  0.153
assa  0.397  0.455 -0.025 -0.009  0.095  0.279 -0.741
burg  0.425 -0.309  0.162  0.011 -0.282  0.737  0.274
larc  0.362 -0.370  0.470 -0.195  0.657 -0.200 -0.037
auto  0.360 -0.344 -0.414  0.723  0.046 -0.220 -0.091
```

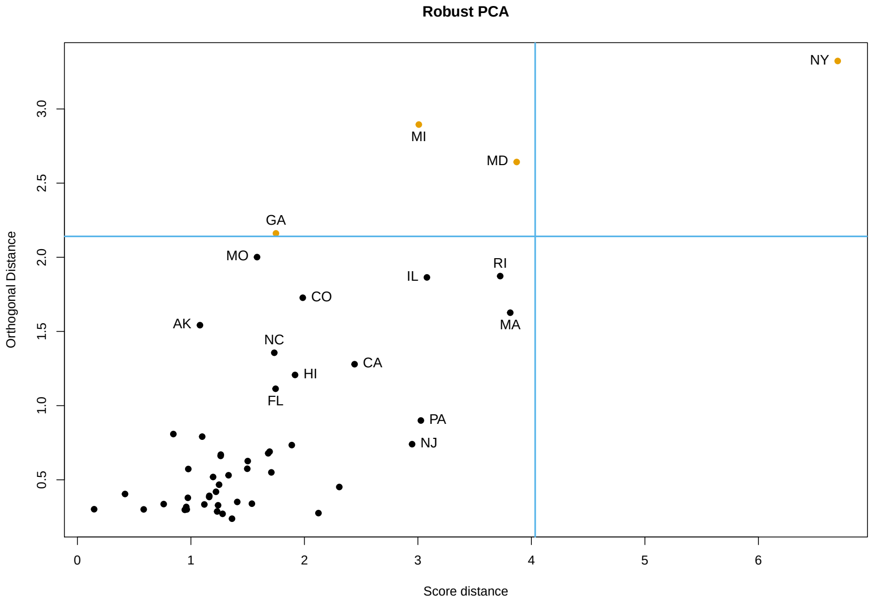
1. Obtain the the value a, missing in the previous R output. (2.0)

2. Decide how many principal components to retain. (2.0)

Suggestion: If you have not answered the previous question use a = 1.0.

3. Give an interpretation to the first two principal components. (3.0)

4. To identify atypical cities related with the crime, an outlier detection based on Robust PCA, using the first three robust principal components and a false alarm rate of 0.001, was estimated, leading to the following distance-distance plot.



Legend: AK - Alaska, CA - California, CO - Colorado, FL - Florida, GA - Georgia, HI - Hawaii, IL - Illinois, MA - Massachusetts, MD - Maryland, MI - Michigan, MO - Missouri, NC - North Carolina, NJ - New Jersey, NY - New York, PA - Pennsylvania, RI - Rhode Island.

(a) Obtain the critical value for the score distance. State any assumptions made. (2.0)

(b) Interpret this result. (1.0)