Departamento de Matemática

Multivariate Analysis

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

1^{st} Test — 1^{st} Semester – 2020/2021

Duration: 1.5 hours — 19/11/2020 – 20:00

Group I

1. Let X_1, X_2, and X_3 be three independent, univariate Normal distributed random variables, with unitary mean and variance ($\mathcal{N}(1,1)$). 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 $\boldsymbol{Y}$. (2.0)

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix} = \begin{bmatrix} X_1 - 3X_2 + 2 \\ 2X_1 - X_2 - 1 \\ X_3 - 1 \end{bmatrix} = \begin{bmatrix} 1 & -3 & 0 \\ 2 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} + \begin{bmatrix} 2 \\ -1 \\ -1 \end{bmatrix}$$

Thus, $E(\underset{\sim}{Y}) = A\underset{\sim}{\mu} + \begin{bmatrix} 2 \\ -1 \\ -1 \end{bmatrix} = \begin{bmatrix} 1 & -3 & 0 \\ 2 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} 2 \\ -1 \\ -1 \end{bmatrix}$

$$= \begin{bmatrix} -2 + 2 \\ 1 - 1 \\ 1 - 1 \end{bmatrix} = \underset{\sim}{0}$$

$$Var(Y) = \underset{\sim}{A}\, Var(\underset{\sim}{X})\, \underset{\sim}{A}^t = \begin{bmatrix} 1 & -3 & 0 \\ 2 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 0 \\ -3 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix} =$$

$$= \begin{bmatrix} 10 & 5 & 0 \\ 5 & 5 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \underset{\sim}{\Sigma}_Y$$

Since $\underset{\sim}{Y} = \underset{\sim}{A}\underset{\sim}{X} + \underset{\sim}{k}$ and $\underset{\sim}{X} \sim \mathcal{N}\left(\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \underset{\sim}{I}\right)$ then

$\underset{\sim}{Y} \sim \mathcal{N}_3(\underset{\sim}{0}, \underset{\sim}{\Sigma}_Y)$

2. Suppose $\boldsymbol{X} = (X_1, X_2, X_3)^t$ has a multivariate normal distribution. Show that if:
(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.

$$\operatorname{diag}\left(\operatorname{Cov}\left(\underset{\sim}{X}, \underbrace{\begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}}_{=A} \underset{\sim}{X}\right)\right) = \underset{\sim}{0}$$

$$(\Leftarrow) \quad \operatorname{diag}(\Sigma A) = \underset{\sim}{0} \quad (\Leftarrow)$$

$$(\Leftarrow) \quad \operatorname{diag}\left(\begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} \end{bmatrix}\begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}\right) = \underset{\sim}{0}$$

$(\Leftarrow)$

$$\begin{cases} \sigma_{12} + \sigma_{13} = 0 \\ \sigma_{12} + \sigma_{23} = 0 \\ \sigma_{13} + \sigma_{23} = 0 \end{cases} \quad \begin{cases} -\sigma_{23} - \sigma_{23} = 0 \\ \sigma_{12} = -\sigma_{23} \\ \sigma_{13} = -\sigma_{23} \end{cases} \quad \begin{cases} \sigma_{23} = 0 \\ \sigma_{12} = 0 \\ \sigma_{13} = 0 \end{cases}$$

thus, $\underset{\sim}{\Sigma} = \operatorname{diag}(\sigma_{11}, \sigma_{22}, \sigma_{33})$

and given that $\underset{\sim}{X} \sim N_p(\mu, \Sigma)$ then $X = (X_1, X_2, X_3)^t$ are indep. r.v

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 $Var(\boldsymbol{X}_1 | \boldsymbol{X}_2 = \boldsymbol{x}_2) = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$.

Let $Y = (X_1, X_2)^t$, $Z = (X_3, X_4)^t$

Then $Y | Z = z \sim N_2(\mu_{Y|z}, \Sigma_{Y|z})$

$$E(Y | Z = z) = E(Y) + \Sigma_{12}\Sigma_{22}^{-1}(z - E(Z))$$

$$= \begin{bmatrix} 1 \\ 3 \end{bmatrix} + \begin{bmatrix} 2 & 1 \\ 3 & 3 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}^{-1} \left(z - \begin{bmatrix} 1 \\ 2 \end{bmatrix} \right)$$

$$Var(Y | Z = z) = \begin{bmatrix} 3 & 3 \\ 3 & 7 \end{bmatrix} - \begin{bmatrix} 2 & 1 \\ 3 & 3 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}^{-1} \begin{bmatrix} 2 & 3 \\ 1 & 3 \end{bmatrix} =$$

$$\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}^{-1} = \frac{1}{4-1} \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} = \frac{1}{3} \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$$

$$Var(Y | Z = z) = \begin{bmatrix} 3 & 3 \\ 3 & 7 \end{bmatrix} - \frac{1}{3} \begin{bmatrix} 2 & 1 \\ 3 & 3 \end{bmatrix} \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} 2 & 3 \\ 1 & 3 \end{bmatrix}$$

$$= \begin{bmatrix} 3 & 3 \\ 3 & 7 \end{bmatrix} - \frac{1}{3} \begin{bmatrix} 2 & 1 \\ 3 & 3 \end{bmatrix} \begin{bmatrix} 4-1 & 6-3 \\ -2+2 & -3+6 \end{bmatrix}$$

$$= \begin{bmatrix} 3 & 3 \\ 3 & 7 \end{bmatrix} - \frac{1}{3} \begin{bmatrix} 2 & 1 \\ 3 & 3 \end{bmatrix} \begin{bmatrix} 3 & 3 \\ 0 & 3 \end{bmatrix}$$

$$= \begin{bmatrix} 3 & 3 \\ 3 & 7 \end{bmatrix} - \frac{1}{3} \begin{bmatrix} 6 & 6+3 \\ 9 & 9+9 \end{bmatrix} = \begin{bmatrix} 3 & 3 \\ 3 & 7 \end{bmatrix} - \frac{1}{3} \begin{bmatrix} 6 & 9 \\ 9 & 18 \end{bmatrix}$$

$$= \begin{bmatrix} 3-2 & 3-3 \\ 3-3 & 7-6 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$E(\underset{\sim}{Y} \mid \underset{\sim}{Z} = \underset{\sim}{z}) = E(\underset{\sim}{Y}) + \Sigma_{12}\,\Sigma_{22}^{-1}\,(\underset{\sim}{z} - E(\underset{\sim}{Z}))$$

$$= \begin{bmatrix} 1 \\ 3 \end{bmatrix} + \begin{bmatrix} 2 & 1 \\ 3 & 3 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}^{-1} \left(\underset{\sim}{x} - \begin{bmatrix} 1 \\ 2 \end{bmatrix} \right)$$

$$= \begin{bmatrix} 1 \\ 3 \end{bmatrix} + \begin{bmatrix} 2 & 1 \\ 3 & 3 \end{bmatrix} \frac{1}{3} \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \left(\underset{\sim}{x} - \begin{bmatrix} 1 \\ 2 \end{bmatrix} \right)$$

$$= \begin{bmatrix} 1 \\ 3 \end{bmatrix} + \frac{1}{3} \begin{bmatrix} 4-1 & -2+2 \\ 6-3 & -3+6 \end{bmatrix} \left(\underset{\sim}{x} - \begin{bmatrix} 1 \\ 2 \end{bmatrix} \right)$$

$$= \begin{bmatrix} 1 \\ 3 \end{bmatrix} + \frac{1}{3} \begin{bmatrix} 3 & 0 \\ 3 & 3 \end{bmatrix} \left(\underset{\sim}{x} - \begin{bmatrix} 1 \\ 2 \end{bmatrix} \right)$$

$$= \begin{bmatrix} 1 \\ 3 \end{bmatrix} + \begin{bmatrix} x_1 - 1 \\ (x_1 - 1) + (x_2 - 2) \end{bmatrix}$$

$$= \begin{bmatrix} x_1 \\ x_1 + x_2 \end{bmatrix}$$

Thus, $\begin{bmatrix} x_3 \\ x_4 \end{bmatrix} \mid \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim N_2\left(\begin{bmatrix} x_1 \\ x_1 + x_2 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right)$

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

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

Given that $\underset{\sim}{X}_i \sim N_p(\underset{\sim}{\mu}_i, \underset{\sim}{\Sigma})$ then

$\underset{\sim}{\bar{X}}_i$ is also $N_p(E(\underset{\sim}{\bar{X}}_i), Cov(\underset{\sim}{\bar{X}}_i))$

where $E(\underset{\sim}{\bar{X}}_i) = \sum_{h=1}^{n_i} \frac{E(\underset{\sim}{X}_{hi})}{n_i} = \frac{n_i E(\underset{\sim}{\mu}_i)}{n_i} = \underset{\sim}{\mu}_i$

$Cov(\underset{\sim}{\bar{X}}_i) = Cov\left(\sum_{h=1}^{n_i} \frac{\underset{\sim}{X}_{hi}}{n_i}\right) = \frac{1}{n_i^2} \sum_{h=1}^{n_i} Cov(\underset{\sim}{X}_{hi}) = \frac{n_i}{n_i^2} \underset{\sim}{\Sigma} = \frac{\underset{\sim}{\Sigma}}{n_i}$

since $\underset{\sim}{X}_{hi} \perp\!\!\!\perp \underset{\sim}{X}_{h'i}$, $\forall h \neq h'$

ie. $\underset{\sim}{\bar{X}}_i \sim N_p(\underset{\sim}{\mu}_i, \underset{\sim}{\Sigma})$.

Given that the two random samples are indep, so

$\underset{\sim}{\bar{X}}_1 \perp\!\!\!\perp \underset{\sim}{\bar{X}}_2$ and $\underset{\sim}{\bar{X}}_i - \underset{\sim}{\bar{X}}_2 \sim N_p(E(\underset{\sim}{\bar{X}}_1 - \underset{\sim}{\bar{X}}_2), Cov(\underset{\sim}{\bar{X}}_1 - \underset{\sim}{\bar{X}}_2))$

thus,

$E(\underset{\sim}{\bar{X}}_1 - \underset{\sim}{\bar{X}}_2) = E(\underset{\sim}{\bar{X}}_1) - E(\underset{\sim}{\bar{X}}_2) = \underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2$

$Cov(\underset{\sim}{\bar{X}}_1 - \underset{\sim}{\bar{X}}_2) \underset{\underset{\sim}{\bar{X}}_1 \perp\!\!\!\perp \underset{\sim}{\bar{X}}_2}{=} Cov(\underset{\sim}{\bar{X}}_1) + Cov(\underset{\sim}{\bar{X}}_2) = \frac{1}{n_1} \underset{\sim}{\Sigma} + \frac{1}{n_2} \underset{\sim}{\Sigma} =$

$= \frac{(n_1 + n_2)}{n_1 n_2} \underset{\sim}{\Sigma}$

Being so $\underset{\sim}{Z} = \underset{\sim}{\bar{X}}_1 - \underset{\sim}{\bar{X}}_2 \sim N_p\left(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2, \frac{n_1 + n_2}{n_1 n_2} \underset{\sim}{\Sigma}\right)$

Thus $(\underset{\sim}{z} - E(\underset{\sim}{z}))^t \underset{\sim}{\Sigma}_{zz}^{-1} (\underset{\sim}{z} - E(\underset{\sim}{z})) \sim \chi^2_{(p)}$

$(=)$

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

QED

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.

1. Obtain the the standard deviation of the second sample principal component, `a`, missing in the previous R output.

We know that $\hat{Var}(\hat{Y}_i) = \hat{\lambda}_i$

and $\sum_{i=1}^{p} \hat{\lambda}_i = p$ since the loadings are estimated based on the sample correlation matrix. Being so, $\sum_{i=1}^{p} \lambda_i = 7$

and $\hat{\lambda}_2 = 7 - \sum_{\substack{i=1 \\ i \neq 2}}^{7} \lambda_i$

$= 7 - 1.9776^2 - 0.8224^2 - \cdots - 0.3405^2$

$= 7 - 5.536711 = 1.463289$

and $a = \sqrt{\hat{\lambda}_2} = \sqrt{1.463289} = 1.209665$

2. Decide how many principal components to retain.

Suggestion: If you have not answer the previous question use $\mathtt{a} = 1.2$.

$\hat{\lambda}_i$:	3.9107	1.4633	0.6767	0.3715
$\hat{\lambda}_i / 7$	0.5587	0.2090	0.0966	0.0531
$\sum_{i=1}^{k} \hat{\lambda}_i / 7$	0.5587	0.7677	0.8643	0.9174

Thus, if the criteria is:

(i) $\hat{\lambda}_i > \bar{\lambda}$ (or) $\hat{\lambda}_i > 1 \Rightarrow k = 2$

(ii) if % variability explained is $> 0.80 \Rightarrow k = 3$

For simplicity reasons we choose $k = 2$.

3. Interpret the first two principal components.

The 1st PC is a weighted mean of all the murders types, so it is a global measure of crime in US.

High (low) values on the 1st PC represents states with high (low) crime levels.

	Comp.1	Comp.2	Comp.
murd	0.272	0.653	
rape	0.431	0.117	
robb	0.376	-0.051	
assa	0.397	0.455	
burg	0.425	-0.309	
larc	0.362	-0.370	
auto	0.360	-0.344	

The 2nd PC is a contrast between murder and assault against burglary, larceny, and auto theft. A state with high (low) value on the 2nd PC is a state with high (low) levels of crimes injured to people (murder and assault) and low (high) levels of crimes injured to property (burglary

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, where GA - Georgia, MD - Maryland, MI - Michigan, and NY- New York.

(a) Obtained the critical value for the score distance.

(b) Interprete the results.

(a) $k = 3$, $\alpha = 0.001$, Assuming $\underset{\sim}{X} \sim N_p(\underset{\sim}{\mu}, \underset{\sim}{\Sigma})$

then $\hat{\underset{\sim}{Y}} = (\hat{Y}_1, \hat{Y}_2, \hat{Y}_3)^t \sim N_3(\underset{\sim}{\mu}_Y, \underset{\sim}{\Sigma}_Y)$

where $\hat{\underset{\sim}{\mu}}_Y = E(\hat{\underset{\sim}{Y}}) = E(\underset{(A\times 3)}{\Gamma^t} \underset{\sim}{Z}) = \underset{(2\times 3)}{\Gamma^t} E(\underset{\sim}{Z}) = \underset{\sim}{0}$

$\underset{\sim}{\Sigma}_Y = Var(\underset{\sim}{Y}) = diag(\hat{\lambda}_1, \hat{\lambda}_2, \hat{\lambda}_3)$

thus, $SD(\underset{\sim}{X}) = \hat{\underset{\sim}{y}}^t \underset{\sim}{\Sigma}_y^{-1} \hat{\underset{\sim}{y}}^t = RD^2(\underset{\sim}{Y}, \underset{\sim}{0})$

$$= \sum_{i=1}^{3} \frac{\hat{y}_i^2}{\hat{\lambda}_i} \sim \chi^2_{(3)}$$

Being so,

c_{OD} : $P(RD^2(\underset{\sim}{X}) > c^2_{OD}) = \alpha = 0.001$

$c^2_{OD} = F^{-1}_{\chi^2_{(3)}}(1 - 0.001) = 16.26624$

and $c_{OD} = \sqrt{16.26624} = 4.03162$

(b) According with the classification rule there are 4 STATES classified as atypical:

New York - with high SD and OD score and

Georgia, Maryland, and Michigan all target by having a high OD.