Multivariate Analysis - Homework 2

Please upload your homework on SPOC before 8:00pm, March 23, including all details needed. For R exercises, R markdown is highly encouraged; for other parts, try to use LaTex.

1. Find the maximum likelihood estimates of the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ based on the random sample

$$\mathbf{Y} = \begin{pmatrix} 3 & 6 \\ 4 & 4 \\ 5 & 7 \\ 4 & 7 \end{pmatrix}$$

from a bivariate normal population.

2. Complete the following.

   (a) Evaluate T^2, for testing H_0: $\boldsymbol{\mu} = (7, 11)'$, using data

$$\begin{pmatrix} 2 & 12 \\ 8 & 9 \\ 6 & 9 \\ 8 & 10 \end{pmatrix}$$

   (b) Specify the distribution of T^2 for the situation in (a).

   (c) Using (a) and (b), test H_0 at the significance level 0.05. What conclusion do you reach?

3. Suppose $\mathbf{y} = (\mathbf{y}_1', \mathbf{y}_2')' \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $|\boldsymbol{\Sigma}| \neq 0$; $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are partitioned accordingly.

   (a) Check that

$$\begin{aligned}(\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu}) \quad = \quad & \{\mathbf{y}_1 - \boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{y}_2 - \boldsymbol{\mu}_2)\}' \\ & \times(\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})^{-1}\{\mathbf{y}_1 - \boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{y} - \boldsymbol{\mu}_2)\} \\ & +(\mathbf{y}_2 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{y}_2 - \boldsymbol{\mu}_2)\end{aligned}$$

   (b) Derive the conditional density $f(\mathbf{y}_1|\mathbf{y}_2)$, and verify that it is still normal.

4. Let $Y_1 \sim N(0,1)$, and let

$$Y_2 = \begin{cases} -Y_1 & \text{if } -1 \le Y_1 \le 1 \\ Y_1 & \text{otherwise.} \end{cases}$$

Show each of the following.

(a) Y_2 also has an $N(0,1)$ distribution. (Hint: Compute the CDF of Y_2.)

(b) Y_1 and Y_2 do not have a bivariate normal distribution.

5. Let $\mathbf{y} = (Y_1, Y_2, Y_3) \sim N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} = (2, -3, 1)'$ and

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 3 & 2 \\ 1 & 2 & 2 \end{bmatrix}$$

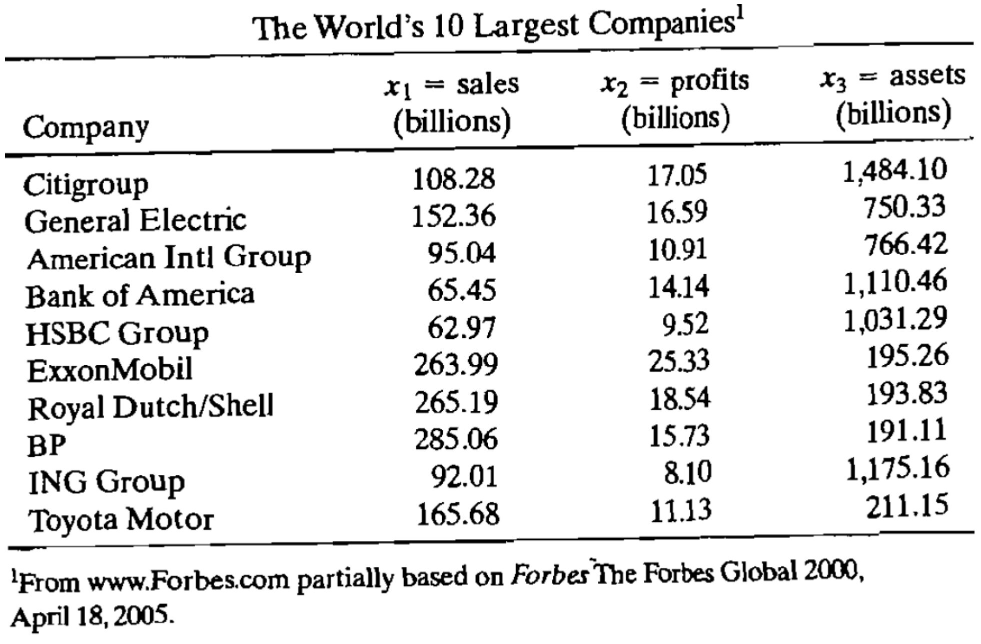
(a) Find the distribution of $3Y_1 - 2Y_2 + Y_3$.

(b) Find the conditional distribution of $Y_1 | (Y_2, Y_3)$.

(c) Find a 2×1 vector $\mathbf{a}$ such that Y_2 and $Y_2 - \mathbf{a}'(Y_1, Y_3)'$ are independent.

6. For one sample case, prove that the likelihood ratio test leads to Hotelling's T^2 test for multivariate normal samples.

7. (R exercise.) The world's 10 largest companies (2005 database) yield the following data (also attached as .txt file):



The World's 10 Largest Companies[1]

Company	x_1 = sales (billions)	x_2 = profits (billions)	x_3 = assets (billions)
Citigroup	108.28	17.05	1,484.10
General Electric	152.36	16.59	750.33
American Intl Group	95.04	10.91	766.42
Bank of America	65.45	14.14	1,110.46
HSBC Group	62.97	9.52	1,031.29
ExxonMobil	263.99	25.33	195.26
Royal Dutch/Shell	265.19	18.54	193.83
BP	285.06	15.73	191.11
ING Group	92.01	8.10	1,175.16
Toyota Motor	165.68	11.13	211.15

[1] From www.Forbes.com partially based on *Forbes* The Forbes Global 2000, April 18, 2005.

For all the three variables:

(a) Construct individual QQ plots to investigate univariate normality. Interpret the output.

(b) Conduct formal statistical tests for the individual normality. Explain the results.

(c) Check the multivariate normality of $(X_1, X_2, X_3)'$ using the pairwise scatter plot matrix and the χ^2 QQ plot.

8. (R exercise) Recall the relationship between the hypothesis testing and the confidence interval, i.e. the conclusion of a test can be directly obtained from the related confidence interval. For the multivariate case, the confidence interval becomes the "confidence region".

   (a) Analogous to the definition of confidence interval, define the $1-\alpha$ confidence region for the population mean vector $\boldsymbol{\mu}$. And derive the mathematical formula of this confidence region when the covariance matrix $\boldsymbol{\Sigma}$ is unknown. Suppose the sample is $\{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$, each $\mathbf{y}_i$ is p-variate, the sample mean vector is $\bar{\mathbf{y}}$ and the sample covariance matrix is $\mathbf{S}$.

   (b) For the sweat data (Page 20 in the slides, and data attached as sweat.dat), suppose we only have the information of the first two variables with mean μ_1 and μ_2. Find the 95% confidence region for $\boldsymbol{\mu} = (\mu_1, \mu_2)'$.

   (c) Describe the confidence region geometrically using the eigenvalue and eigenvectors of the sample covariance matrix $\mathbf{S}$.

   (d) Consturct the 95% univariate confidence interval for each variable.

   (e) Condsider the test H_0: $\boldsymbol{\mu} = \boldsymbol{\mu}_0$. Give an example of $\boldsymbol{\mu}_0$ such that the multivariate test rejects H_0 but both univariate tests fails to do so. You should answer this question based on the confidence region and confidence intervals obtained in (b) and (d).