Multivariate Analysis Homework 1

A49109720 Yi-Chen Zhang

March 16, 2018

4.2. Consider a bivariate normal population with $\mu_1 = 0$, $\mu_2 = 2$, $\sigma_{11} = 2$, $\sigma_{22} = 1$, and $\rho_{12} = 0.5$.

(a) Write out the bivariate normal density.

(b) Write out the squared generalized distance expression $(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})$ as a function of x_1 and x_2.

(c) Determine (and sketch) the constant-density contour that contains 50% of the probability.

Sol. (a) The multivariate normal density is defined by the following equation.

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right\}.$$

In this question, we have $p = 2$, $\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$, $\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$, $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}$, and $\sigma_{12} = \rho_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}$. Note that $\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 2 \end{pmatrix}$, $\boldsymbol{\Sigma} = \begin{pmatrix} 2 & \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & 1 \end{pmatrix}$, $|\boldsymbol{\Sigma}| = 2 \times 1 - \left(\frac{\sqrt{2}}{2}\right)^2 = \frac{3}{2}$, $|\boldsymbol{\Sigma}|^{1/2} = \sqrt{\frac{3}{2}}$, and $\boldsymbol{\Sigma}^{-1} = \frac{2}{3}\begin{pmatrix} 1 & -\frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} & 2 \end{pmatrix}$. So the bivariate normal density is

$$\begin{aligned} f(\boldsymbol{x}) &= \frac{1}{(2\pi)^{2/2}\sqrt{\frac{3}{2}}} \exp\left\{-\frac{1}{2}\begin{pmatrix} x_1 & x_2 - 2 \end{pmatrix} \frac{2}{3}\begin{pmatrix} 1 & -\frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} & 2 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 - 2 \end{pmatrix}\right\} \\ &= \frac{1}{\sqrt{6}\pi} \exp\left\{-\frac{1}{3}\left(x_1^2 - \sqrt{2}x_1(x_2 - 2) + 2(x_2 - 2)^2\right)\right\} \end{aligned}$$

(b)

$$\begin{aligned} (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) &= \begin{pmatrix} x_1 & x_2 - 2 \end{pmatrix} \frac{2}{3}\begin{pmatrix} 1 & -\frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} & 2 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 - 2 \end{pmatrix} \\ &= \frac{2}{3}\left(x_1^2 - \sqrt{2}x_1(x_2 - 2) + 2(x_2 - 2)^2\right). \end{aligned}$$

(c) For $\alpha = 0.5$, the solid ellipsoid of (x_1, x_2) satisfy $(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \le \chi^2_{p,\alpha} = c^2$ will have probability 50%. From the quantile function in R we have $\chi^2_{2,0.5} =$ `qchisq(0.5,df=2)` $= 1.3863$, therefore, $c = 1.1774$. The eigenvalues of $\boldsymbol{\Sigma}$ are $(\lambda_1, \lambda_2) = (2.3660, 0.6340)$ with eigenvectors $\begin{pmatrix} \boldsymbol{e}_1 & \boldsymbol{e}_2 \end{pmatrix} = \begin{pmatrix} -0.8881 & 0.4597 \\ -0.4597 & -0.8881 \end{pmatrix}$. Therefore, we have the axes as: $c\sqrt{\lambda_1} = 1.8111$ and $c\sqrt{\lambda_2} = 0.9375$. The contour is plotted in Figure 1.

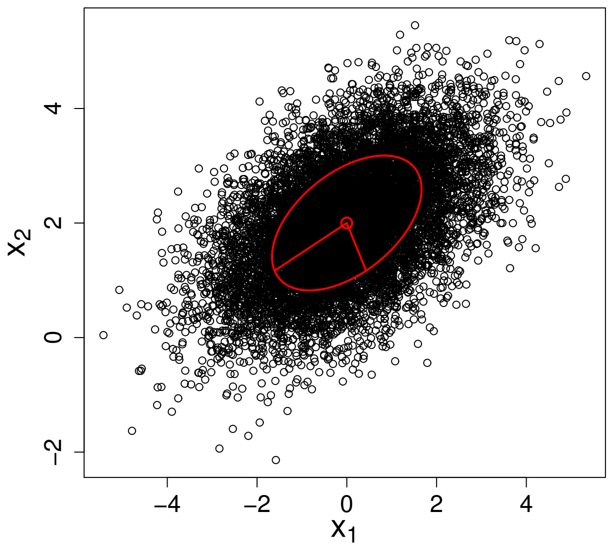

Figure 1: Contour that contains 50% of the probability

4.4. Let $\boldsymbol{X}$ be $N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu}^T = (2, -3, 1)$ and $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 3 & 2 \\ 1 & 2 & 2 \end{pmatrix}$

(a) Find the distribution of $3X_1 - 2X_2 + X_3$.

(b) Relabel the variables if necessary, and find a 2×1 vector $\boldsymbol{a}$ such that X_2 and $X_2 - \boldsymbol{a}^T \begin{pmatrix} X_1 \\ X_3 \end{pmatrix}$ are independent.

Sol. (a) Let $\boldsymbol{a} = (3, -2, 1)^T$, then $\boldsymbol{a}^T\boldsymbol{X} = 3X_1 - 2X_2 + X_3$. Therefore,

$$\boldsymbol{a}^T\boldsymbol{X} \sim N(\boldsymbol{a}^T\boldsymbol{\mu}, \boldsymbol{a}^T\boldsymbol{\Sigma}\boldsymbol{a}),$$

where

$$\boldsymbol{a}^T\boldsymbol{\mu} = \begin{pmatrix} 3 & -2 & 1 \end{pmatrix} \begin{pmatrix} 2 \\ -3 \\ 1 \end{pmatrix} = 13$$

and

$$\boldsymbol{a}^T\boldsymbol{\Sigma}\boldsymbol{a} = \begin{pmatrix} 3 & -2 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \\ 1 & 3 & 2 \\ 1 & 2 & 2 \end{pmatrix} \begin{pmatrix} 3 \\ -2 \\ 1 \end{pmatrix} = 9$$

The distribution of $3X_1 - 2X_2 + X_3$ is $N_3(13, 9)$.

(b) Let $\boldsymbol{a} = \begin{pmatrix} a_1 & a_2 \end{pmatrix}^T$, then $Y = X_2 - \boldsymbol{a}^T \begin{pmatrix} X_1 \\ X_3 \end{pmatrix} = -a_1X_1 + X_2 - a_2X_3$.

Now, let $\boldsymbol{A} = \begin{pmatrix} 0 & 1 & 0 \\ -a_1 & 1 & -a_2 \end{pmatrix}$, then $\boldsymbol{A}\boldsymbol{X} = \begin{pmatrix} X_2 \\ Y \end{pmatrix} \sim N(\boldsymbol{A}\boldsymbol{\mu}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$, where

$$\begin{aligned} \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T &= \begin{pmatrix} 0 & 1 & 0 \\ -a_1 & 1 & -a_2 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \\ 1 & 3 & 2 \\ 1 & 2 & 2 \end{pmatrix} \begin{pmatrix} 0 & -a_1 \\ 1 & 1 \\ 0 & -a_2 \end{pmatrix} \\ &= \begin{pmatrix} 3 & -a_1 - 2a_2 + 3 \\ -a_1 - 2a_2 + 3 & a_1^2 - 2a_1 - 4a_2 + 2a_1a_2 + 2a_2^2 + 3 \end{pmatrix} \end{aligned}$$

Since we want to have X_2 and Y independent, this implies that $-a_1 - 2a_2 + 3 = 0$. So we have vector

$$\boldsymbol{a} = \begin{pmatrix} 3 \\ 0 \end{pmatrix} + c\begin{pmatrix} -2 \\ 1 \end{pmatrix}, \text{ for } c \in \mathbb{R}$$

4.6. Let $\boldsymbol{X}$ be distributed as $N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}^T = (1, -1, 2)$ and $\boldsymbol{\Sigma} = \begin{pmatrix} 4 & 0 & -1 \\ 0 & 5 & 0 \\ -1 & 0 & 2 \end{pmatrix}$. Which of the following random variables are independent? Explain.

(a) X_1 and X_2

(b) X_1 and X_3

(c) X_2 and X_3

(d) (X_1, X_3) and X_2

(e) X_1 and $X_1 + 3X_2 - 2X_3$

Sol. (a) $\sigma_{12} = \sigma_{21} = 0$, X_1 and X_2 are independent.

(b) $\sigma_{13} = \sigma_{31} = -1$, X_1 and X_3 are not independent.

(c) $\sigma_{23} = \sigma_{32} = 0$, X_2 and X_3 are independent.

(d) We rearrange the covariance matrix and partition it. The new covariance matrix is as following:

$$\boldsymbol{\Sigma}^* = \left(\begin{array}{cc|c} 4 & -1 & 0 \\ -1 & 2 & 0 \\ \hline 0 & 0 & 5 \end{array}\right)$$

It is clear that (X_1, X_3) and X_2 are independent.

(e) Let $\boldsymbol{A} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 3 & -2 \end{pmatrix}$, then $\boldsymbol{AX} = \begin{pmatrix} X_1 \\ X_1 + 3X_2 - 2X_3 \end{pmatrix}$ and $\boldsymbol{AX} \sim N(\boldsymbol{A\mu}, \boldsymbol{A\Sigma A}^T)$, where

$$\begin{aligned} \boldsymbol{A\Sigma A}^T &= \begin{pmatrix} 1 & 0 & 0 \\ 1 & 3 & -2 \end{pmatrix} \begin{pmatrix} 4 & 0 & -1 \\ 0 & 5 & 0 \\ -1 & 0 & 2 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 3 \\ 0 & -2 \end{pmatrix} \\ &= \begin{pmatrix} 4 & 6 \\ 6 & 61 \end{pmatrix} \end{aligned}$$

It is clear that X_1 and $X_1 + 3X_2 - 2X_3$ are not independent.

4.7. Refer to Exercise 4.6 and specify each of the following.

(a) The conditional distribution of X_1, given that $X_3 = x_3$.

(b) The conditional distribution of X_1, given that $X_2 = x_2$ and $X_3 = x_3$.

Sol. We use the result 4.6 from textbook. Let $\boldsymbol{X} = \left(\frac{\boldsymbol{X}_1}{\boldsymbol{X}_2}\right) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} = \left(\frac{\boldsymbol{\mu}_1}{\boldsymbol{\mu}_2}\right)$ and $\boldsymbol{\Sigma} = \left(\begin{array}{c|c} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \hline \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{array}\right)$ and $|\boldsymbol{\Sigma}_{22}| > 0$. Then

$$\boldsymbol{X}_1 \big| \boldsymbol{X}_2 = \boldsymbol{x}_2 \sim N\left(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right)$$

(a)

$$X_1 \big| X_3 = x_3 \sim N\left(1 + (-1)(2)^{-1}(x_3 - 2), 4 - (-1)(2)^{-1}(-1)\right)$$
$$\Rightarrow \quad X_1 \big| X_3 = x_3 \sim N\left(-\frac{1}{2}x_3 + 2, \right)$$

(b)

$$X_1 \big| X_2 = x_2, X_3 = x_3$$
$$\sim N\left(1 + \begin{pmatrix}0 & -1\end{pmatrix}\begin{pmatrix}5 & 0\\ 0 & 2\end{pmatrix}^{-1}\begin{pmatrix}x_2 - (-1)\\ x_3 - 2\end{pmatrix}, 4 - \begin{pmatrix}0 & -1\end{pmatrix}\begin{pmatrix}5 & 0\\ 0 & 2\end{pmatrix}^{-1}\begin{pmatrix}0\\ -1\end{pmatrix}\right)$$
$$\Rightarrow \quad X_1 \big| X_2 = x_2, X_3 = x_3 \sim N\left(-\frac{1}{2}x_3 + 2, \right)$$

4.16. Let $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, $\boldsymbol{X}_3$, and $\boldsymbol{X}_4$ be independent $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ random vectors.

(a) Find the marginal distributions for each of the random vectors

$$\boldsymbol{V}_1 = \frac{1}{4}\boldsymbol{X}_1 - \frac{1}{4}\boldsymbol{X}_2 + \frac{1}{4}\boldsymbol{X}_3 - \frac{1}{4}\boldsymbol{X}_4$$

and

$$\boldsymbol{V}_2 = \frac{1}{4}\boldsymbol{X}_1 + \frac{1}{4}\boldsymbol{X}_2 - \frac{1}{4}\boldsymbol{X}_3 - \frac{1}{4}\boldsymbol{X}_4$$

(b) Find the joint density of the random vectors $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$ defined in (a).

Sol. (a) By result 4.8 in the textbook, $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$ have the following distribution

$$N_p\left(\sum_{i=1}^{n} c_i \boldsymbol{\mu}, \left(\sum_{i=1}^{n} c_i^2\right)\boldsymbol{\Sigma}\right)$$

Then we have $\boldsymbol{V}_1 \sim N_p(\boldsymbol{0}, \frac{1}{4}\boldsymbol{\Sigma})$ and $\boldsymbol{V}_2 \sim N_p(\boldsymbol{0}, \frac{1}{4}\boldsymbol{\Sigma})$.

(b) Also by result 4.8, $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$ are jointly multivariate normal with covariance matrix

$$\begin{pmatrix} \left(\sum_{i=1}^{n} c_i^2\right)\boldsymbol{\Sigma} & (\boldsymbol{b}^T\boldsymbol{c})\boldsymbol{\Sigma} \\ (\boldsymbol{b}^T\boldsymbol{c})\boldsymbol{\Sigma} & \left(\sum_{j=1}^{n} b_j^2\right)\boldsymbol{\Sigma} \end{pmatrix},$$

with $\boldsymbol{c} = (\frac{1}{4}, -\frac{1}{4}, \frac{1}{4}, -\frac{1}{4})^T$ and $\boldsymbol{b} = (\frac{1}{4}, \frac{1}{4}, -\frac{1}{4}, -\frac{1}{4})^T$. So that we have the joint distribution of $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$ as following:

$$\begin{pmatrix}\boldsymbol{V}_1\\ \boldsymbol{V}_2\end{pmatrix} \sim N_{2p}\left(\begin{pmatrix}\boldsymbol{0}\\ \boldsymbol{0}\end{pmatrix}, \begin{pmatrix}\frac{1}{4}\boldsymbol{\Sigma} & \boldsymbol{0}\\ \boldsymbol{0} & \frac{1}{4}\boldsymbol{\Sigma}\end{pmatrix}\right)$$

4.18. Find the maximum likelihood estimates of the 2×1 mean vector $\boldsymbol{\mu}$ and the 2×2 covariance matrix $\boldsymbol{\Sigma}$ based on the random sample

$$\boldsymbol{X} = \begin{pmatrix}3 & 6\\ 4 & 4\\ 5 & 7\\ 4 & 7\end{pmatrix}$$

from a bivariate normal population.

Sol. Since the random samples $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, $\boldsymbol{X}_3$, and $\boldsymbol{X}_4$ are from normal population, the maximum likelihood estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are $\bar{\boldsymbol{X}}$ and $\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^T$. Therefore,

$$\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{X}} = \begin{pmatrix} 4 \\ 6 \end{pmatrix} \text{ and } \widehat{\boldsymbol{\Sigma}} = \frac{1}{4}\sum_{i=1}^{4}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^T = \begin{pmatrix} 1/2 & 1/4 \\ 1/4 & 3/2 \end{pmatrix}$$

4.19. Let $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, $\ldots$, $\boldsymbol{X}_{20}$ be a random sample of size $n = 20$ from an $N_6(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ population. Specify each of the following completely.

(a) The distribution of $(\boldsymbol{X}_1 - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_1 - \boldsymbol{\mu})$

(b) The distributions of $\bar{\boldsymbol{X}}$ and $\sqrt{n}(\bar{\boldsymbol{X}} - \boldsymbol{\mu})$

(c) The distribution of $(n-1)\boldsymbol{S}$

Sol. (a) $(\boldsymbol{X}_1 - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_1 - \boldsymbol{\mu})$ is distributed as χ^2_6

(b) $\bar{\boldsymbol{X}}$ is distributed as $N_6\left(\boldsymbol{\mu}, \frac{1}{20}\boldsymbol{\Sigma}\right)$ and $\sqrt{n}\left(\bar{\boldsymbol{X}} - \boldsymbol{\mu}\right)$ is distributed as $N_6\left(\boldsymbol{0}, \boldsymbol{\Sigma}\right)$

(c) $(n-1)\boldsymbol{S}$ is distributed as Wishart distribution $\sum_{i=1}^{20-1}\boldsymbol{Z}_i\boldsymbol{Z}_i^T$, where $\boldsymbol{Z}_i \sim N_6(\boldsymbol{0}, \boldsymbol{\Sigma})$.

We write this as $W_6(19, \boldsymbol{\Sigma})$, i.e., Wishart distribution with dimensionality 6, degrees of freedom 19, and covariance matrix $\boldsymbol{\Sigma}$.

4.21. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{60}$ be a random sample of size 60 from a four-variate normal distribution having mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. Specify each of the following completely.

(a) The distribution of $\bar{\boldsymbol{X}}$

(b) The distribution of $(\boldsymbol{X}_1 - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_1 - \boldsymbol{\mu})$

(c) The distribution of $n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu})$

(d) The approximate distribution of $n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})^T\boldsymbol{S}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu})$

Sol. (a) $\bar{\boldsymbol{X}}$ is distributed as $N_4\left(\boldsymbol{\mu}, \frac{1}{60}\boldsymbol{\Sigma}\right)$.

(b) $(\boldsymbol{X}_1 - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_1 - \boldsymbol{\mu})$ is distributed as χ^2_4.

(c) $n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu})$ is distributed as χ^2_4.

(d) Since $60 \gg 4$, $n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})^T\boldsymbol{S}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu})$ can be approximated as χ^2_4.

4.23. Consider the annual rates of return (including dividends) on the Dow-Jones industrial average for the years 1996-2005. These data, multiplied by 100, are

$$-0.6 \quad 3.1 \quad 25.3 \quad -16.8 \quad -7.1 \quad -6.2 \quad 25.2 \quad 22.6 \quad 26.0$$

Use these 10 observations to complete the following.

(a) Construct a Q-Q plot. Do the data seem to be normally distributed? Explain.

(b) Carry out a test of normality based on the correlation coefficient r_Q. Let the significance level be $\alpha = 0.1$.

Sol. (a) The Q-Q plot of this data is plotted in Figure 2. It seems that all the sample quantiles are close the theoretical quantiles. However, the Q-Q plots are not particularly informative unless the sample size is moderate to large, for instance, $n \geq 20$. There can be quite a bit of variability in the straightness of the Q-Q plot for small samples, even when the observations are known to come from a normal population.

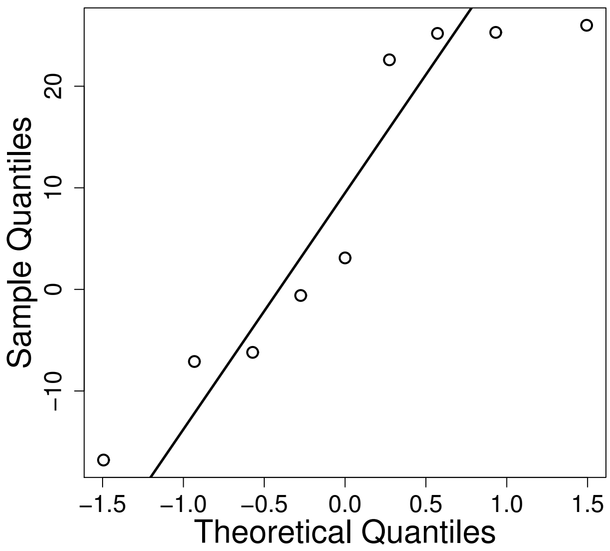

Figure 2: Normal Q-Q plot

(b) From (4-31) in the textbook, the q_Q is defined by

$$r_Q = \frac{\sum_{j=1}^{n}(x_{(j)} - \bar{x})(q_{(j)} - \bar{q})}{\sqrt{\sum_{j=1}^{n}(x_{(j)} - \bar{x})^2}\sqrt{\sum_{j=1}^{n}(q_{(j)} - \bar{q})^2}}$$

Using the information from the data, we have $r_Q = 0.9351$. The R code of this calculation is compiled in Appendix. From Table 4.2 in the textbook we know that the critical point to test of normality at the 10% level of significance corresponding to $n = 9$ and $\alpha = 0.1$ is between 0.9032 and 0.9351. Since $r_Q = 0.9351 >$ the critical point, we do not reject the hypothesis of normality.

4.26. Exercise 1.2 gives the age x_1, measured in years, as well as the selling price x_2, measured in thousands of dollars, for $n = 10$ used cars. These data are reproduced as follows:

x_1	1	2	3	3	4	5	6	8	9	11
x_2	18.95	19.00	17.95	15.54	14.00	12.95	8.94	7.49	6.00	3.99

(a) Use the results of Exercise 1.2 to calculate the squared statistical distances $(\boldsymbol{x}_j - \bar{\boldsymbol{x}})^T\boldsymbol{S}^{-1}(\boldsymbol{x}_j - \bar{\boldsymbol{x}})$, $j = 1, 2, \ldots, 10$, where $\boldsymbol{x}_j^T = (x_{j1}, x_{j2})$.

(b) Using the distances in Part (a), determine the proportion of the observations falling within the estimated 50% probability contour of a bivariate normal distribution.

(c) Order the distances in Part (a) and construct a chi-square plot.

(d) Given the results in Parts (b) and (c), are these data approximately bivariate normal? Explain.

Sol. (a) From Exercise 1.2 we have $\bar{\boldsymbol{x}} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \end{pmatrix} = \begin{pmatrix} 5.2 \\ 12.481 \end{pmatrix}$ and $\boldsymbol{S} = \begin{pmatrix} 10.6222 & -17.7102 \\ -17.7102 & 30.8544 \end{pmatrix}$.
The squared statistical distances $d_j^2 = (\boldsymbol{x}_j - \bar{\boldsymbol{x}})^T\boldsymbol{S}^{-1}(\boldsymbol{x}_j - \bar{\boldsymbol{x}})$, $j = 1, \ldots, 10$ are calculated and listed below

d_1^2	d_2^2	d_3^2	d_4^2	d_5^2	d_6^2	d_7^2	d_8^2	d_9^2	d_{10}^2
1.8753	2.0203	2.9009	0.7352	0.3105	0.0176	3.7329	0.8165	1.3753	4.2152

(b) We plot the data points and 50% probability contour (the blue ellipse) in Figure 3. It is clear that subject 4, 5, 6, 8, and 9 are falling within the estimated 50% probability contour. The proportion of that is 0.5.

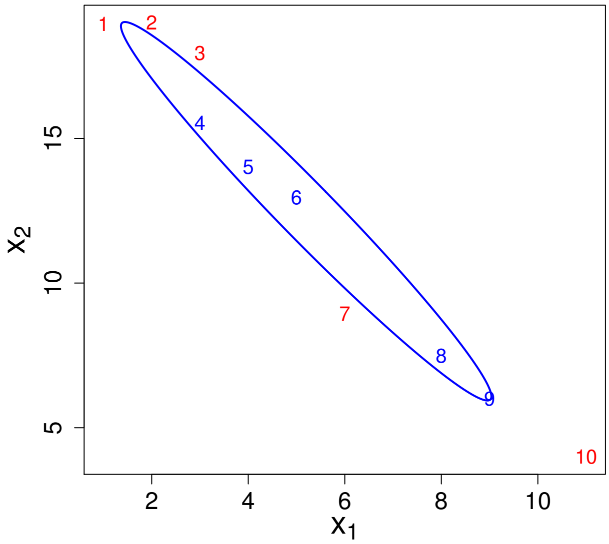

Figure 3: Contour of a bivariate normal

(c) The squared distances in Part (a) are ordered as below. The chi-square plot is shown in Figure 4.

d_6^2	d_5^2	d_4^2	d_8^2	d_9^2	d_1^2	d_2^2	d_3^2	d_7^2	d_{10}^2
0.0176	0.3105	0.7353	0.8165	1.3753	1.8753	2.0203	2.9009	3.7329	4.2153

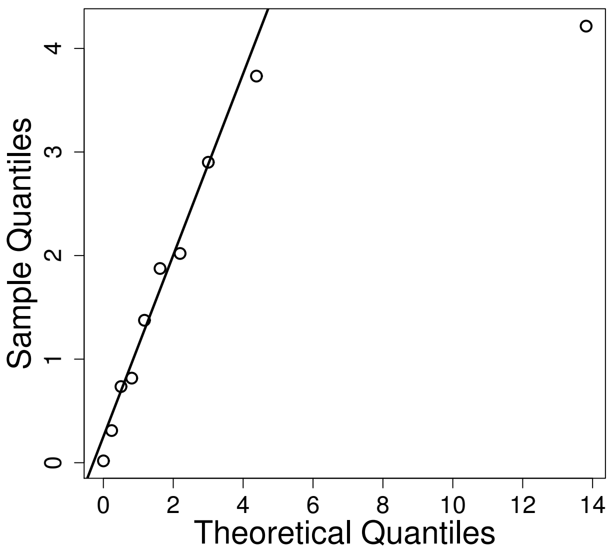

Figure 4: Chi-square plot

(d) Given the results in Parts (b) and (c), we conclude these data are approximately bivariate normal. Most of the data are around the theoretical line.

Appendix

R code for Problem 4.2 (c).

```
> library(ellipse)
> library(MASS)
> library(mvtnorm)
> set.seed(123)
>
> mu <- c(0,2)
> Sigma <- matrix(c(2,sqrt(2)/2,sqrt(2)/2,1), nrow=2, ncol=2)
> X <- mvrnorm(n=10000,mu=mu, Sigma=Sigma)
> lambda <- eigen(Sigma)$values
> Gamma <- eigen(Sigma)$vectors
> elps <- t(t(ellipse(Sigma, level=0.5, npoints=1000))+mu)
> chi <- qchisq(0.5,df=2)
> c <- sqrt(chi)
> factor <- c*sqrt(lambda)
> plot(X[,1],X[,2])
> lines(elps)
> points(mu[1], mu[2])
> segments(mu[1],mu[2],factor[1]*Gamma[1,1],factor[1]*Gamma[2,1]+mu[2])
> segments(mu[1],mu[2],factor[2]*Gamma[1,2],factor[2]*Gamma[2,2]+mu[2])
```

R code for Problem 4.23.

```
> x <- c(-0.6, 3.1, 25.3, -16.8, -7.1, -6.2, 25.2, 22.6, 26.0)
> # (a)
> qqnorm(x)
> qqline(x)
> # (b)
> y <- sort(x)
> n <- length(y)
> p <- (1:n)-0.5)/n
> q <- qnorm(p)
> rQ <- cor(y,q)
```

R code for Problem 4.26.

```
> n <- 10
> x1 <- c(1,2,3,3,4,5,6,8,9,11)
> x2 <- c(18.95, 19.00, 17.95, 15.54, 14.00, 12.95, 8.94, 7.49, 6.00, 3.99)
> X <- cbind(x1,x2)
> Xbar <- colMeans(X)
> S <- cov(X)
> Sinv <- solve(S)
>
> # (a)
> d <- diag(t(t(X)-Xbar)%*%Sinv%*%(t(X)-Xbar))
>
> # (b)
> library(ellipse)
```

```
> p <- 2
> elps <- t(t(ellipse(S, level=0.85, npoints=1000))+Xbar)
> plot(X[,1],X[,2],type="n")
> index <- d < qchisq(0.5,df=p)
> text(X[,1][index],X[,2][index],(1:n)[index],col="blue")
> text(X[,1][!index],X[,2][!index],(1:n)[!index],col="red")
> lines(elps,col="blue")
>
> # (c)
> names(d) <- 1:10
> sort(d)
> qqplot(qchisq(ppoints(500),df=p), d, main="",
+ xlab="Theoretical Quantiles", ylab="Sample Quantiles")
> qqline(d,distribution=function(x){qchisq(x,df=p)})
```