Homework-2

1. Let $\boldsymbol{X}$ be a $p \times 1$ random vector with covariance matrix $\text{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}_x$. Let $\boldsymbol{A}$ be an $r \times p$ constant matrix and let $\boldsymbol{B}$ be a $q \times p$ constant matrix. Find $\text{Cov}(\boldsymbol{AX}, \boldsymbol{BX})$ in terms of $\boldsymbol{A}, \boldsymbol{B}$ and $\text{Cov}(\boldsymbol{X})$.

2. a) Let $\boldsymbol{\Sigma}$ be a $p \times p$ matrix with eigenvalue eigenvector pair $(\lambda, \boldsymbol{x})$. Show that $c\boldsymbol{x}$ is also an eigenvector of $\boldsymbol{\Sigma}$ where $c \neq 0$ is a real number.

   b) Let $\boldsymbol{\Sigma}$ be a $p \times p$ matrix with eigenvalue eigenvector pairs $(\lambda_1, \boldsymbol{e}_1), ..., (\lambda_p, \boldsymbol{e}_p)$. Find the eigenvalue eigenvector pairs of $\boldsymbol{A} = c\boldsymbol{\Sigma}$ where $c \neq 0$ is a real number.

3. Let $\mathbf{A} = \begin{pmatrix} 9 & -2 \\ -2 & 6 \end{pmatrix}$

   (a) Is $\mathbf{A}$ symmetric?

   (b)) Show that $\mathbf{A}$ is positive definite.

4. Use the matrix $\mathbf{A} = \begin{pmatrix} 4 & 8 & 8 \\ 3 & 6 & 9 \end{pmatrix}$

   (a) Calculate $\mathbf{A}\mathbf{A}^T$ and obtain its eigenvalues and eigenvectors.

   (b) Calculate $\mathbf{A}^T\mathbf{A}$ and obtain its eigenvalues and eigenvectors. Check that the nonzero eigenvalues are the same as those in part (a).

   (c) Obtain the spectral decomposition of $\mathbf{A}\mathbf{A}^T$.

5. Consider an arbitrary $n \times p$ matrix $\mathbf{A}$.

   (a) Show that $\mathbf{A}^T\mathbf{A}$ is a symmetric $p \times p$ matrix.

   (b) Show that $\mathbf{A}^T\mathbf{A}$ is always nonnegative definite.

6. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{60}$ be a random sample of size 60 from a five-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 $n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})^T\mathbf{S}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu})$.

   (e) The distribution of $(n-1)\mathbf{S}$.

7. Consider five measurements of normal patients and diabetics in the data *Diabetes.txt*. The variables are y_1: relative weight; y_2: fasting plasma glucose; x_1: glucose intolerance; x_2: insulin response to oral glucose; x_3: insulin resistance. The original data are from Reaven and Miller (1979, Diabetologia). Focus on x variables that are of main interest. Answer the following questions:

(a) Are the 3-dimensional data normally distributed? Why? Use all of the possible tests introduced.

(b) Compute the sample mean and covariance matrix of the x variables.

(c) Compute the eigenvalues and eigenvectors of the sample covariance matrix.

8. 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 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 $\alpha = 0.1$

9. The air pollution data set is given by *airpoll.txt*. For this problem, only focus on the first 16 observations (cities).

(a) Do a star plot to display all 7 variables. And also do faces plots. Write a short paragraph explaining what the plots tell you about the cities. You can include the "labels" argument to label the drawings for both the stars function and the faces function, e.g.: labels=city.names within the call of each function.

(b) Produce a scatterplot matrix for this air pollution data set. Write a short paragraph explaining the main conclusions from the scatterplot matrix.

(c) Do a bivariate boxplot of the pair of variables "Education" and "Mortality" from the air pollution data set. Explain what the plot tells you about the relationship between the two variables. Do you see any outliers? If so, which cities are they?

(d) Do a bubble plot with "Education" and "Mortality" on the axes and "Population Density" represented by the bubbles. Explain what the plot tells you about the relationships among the three variables. Comment on any notable cities.

(e) Use R to calculate the sample covariance matrix and the sample correlation matrix for this data subset. Identify which pairs of variables seem to be strongly associated. Write a paragraph describing the nature (strength and direction) of the relationship between these variable pairs.

(f) Use R to calculate the distance matrix for these observations (after scaling the variables by dividing each variable by its standard deviation). Write a paragraph describing some of the most similar pairs of cities and some of the most different pairs of cities, giving evidence from the distance matrix.

(g) Give a plot that will help assess whether this data set comes from a multivariate normal distribution. What is your conclusion, based on the plot? Moreover, apply all appropriate hypothesis tests, and then make your final conclusion.

10. **Reading assignments:** Chapters 3 to 5 of the textbook.