The University of New South Wales
Department of Statistics
School of Mathematics
MATH5855 - Multivariate Analysis I
Assignment 1

Due August Monday, 27th August, 2018, 5pm (absolute deadline).
The assignment is to be handed in as a **hard copy** (no e-mails!)
Maximal number of pages: **8**.

1. Consider the joint density

$$f(x_1, x_2) = 2e^{-\frac{x_1}{x_2}}, x_1 > 0, 0 < x_2 < 1.$$

a) Compute $f_{X_2}(x_2)$ and $f_{X_1|X_2}(x_1|x_2)$.

b) Give the best approximation $g^*(X_2)$ in mean square sense for X_1 (i.e. find explicitly $g^*(X_2)$ that minimizes $E[X_1 - g(X_2)]^2$ over all possible choices of $g(X_2)$ such that $E[g(X_2)^2] < \infty$).

c) For a given realization x_2, calculate the mean square error of the best approximation (that is, calculate $E\{[X_1 - g^*(X_2)]^2|X_2 = x_2\}$.)

2. Write down the spectral decomposition of the matrix $\mathbf{\Sigma} = \begin{pmatrix} 3 & 2 \\ 2 & 3 \end{pmatrix}$. Correspondingly, find $\mathbf{\Sigma}^{\frac{1}{2}}$, $\mathbf{\Sigma}^{\frac{1}{4}}$ and $\mathbf{\Sigma}^{-\frac{1}{2}}$ (give final answers). Find the eigenvectors of $\mathbf{\Sigma}^{-1}$.

In addition, if $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ has a multivariate normal distribution with $\mathbf{\Sigma}$ as above being its covariance matrix and if $n = 25$ observations gave a vector of sample means $\begin{pmatrix} \bar{X}_1 \\ \bar{X}_2 \end{pmatrix} = \begin{pmatrix} 1.2 \\ 2.5 \end{pmatrix}$, draw a confidence ellipsoid for the mean vector $\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$ at a level of 95% as accurately as possible.

3. The table below gives an *extract* of the petal lengths and widths of two types of iris, Iris setosa and Iris versicolor.

Petal length Iris setosa	Petal width Iris Setosa	Petal length Iris versicolor	Petal width Iris versicolor
x_1	x_2	x_3	x_4
1.4	0.2	4.7	1.4
...	...	...	...
1.4	0.2	4.1	1.3

You should download the whole data set from the file `iris.dat` in `moodle`.

i) Test hypothesis about multivariate normality of the vector $\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{pmatrix}$ (use the IML program discussed in the SAS Lab 2 or some other method).

Hint: to create a **matrix x** from the vector observations x1-x4 you can proceed as follows:

```
data iris; infile 'iris.dat'; input x1 x2 x3 x4;run;
proc iml; use iris var{ x1 x2 x3 x4};
read all var _num_  into x;
```

ii) Find $\hat{\mu}$ and the sample covariance matrix. Use the IML program again or use the SAS procedure CORR. (Hint for users of CORR : proc corr cov; var x1-x4;).

iii) Estimate the conditional distribution (i.e. give estimators of mean vector and covariance matrix) of (X_3, X_4) given that $x_1 = 1.3, x_2 = 1.5$. (Use submatrices within IML.)

iv) The sample covariance matrix is

$$\mathbf{S} = \begin{pmatrix} 0.0301591837 & 0.0060693878 & -.0156326531 & -.0053183673 \\ 0.0060693878 & 0.0111061224 & -.0020000000 & -.0036693878 \\ -.0156326531 & -.0020000000 & 0.2208163265 & 0.0731020408 \\ -.0053183673 & -.0036693878 & 0.0731020408 & 0.0391061224 \end{pmatrix}$$

(or a reasonable rounding of it depending on whether you used IML or CORR). Calculate its inverse $\mathbf{S}^{-1}$ e.g. by using the IML procedure. Calculate the T^2 statistic and test the hypothesis $H_0 : E \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{pmatrix} = \begin{pmatrix} 1.2 \\ 0.6 \\ 4.0 \\ 1.6 \end{pmatrix}$ at 5% level of significance.

v) Test the hypothesis that the mean length and width of the Iris setosa equals those of the Iris versicolor at significance level 0.05. What is your conclusion? (**Hint:** Transform $X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{pmatrix}$ into $Y = \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix} X$ and reformulate the hypothesis in terms of the means of Y).

4. A shop manager is studying the sales of certain brand over periods of time. He uses different marketing methods and measures variables such as: number of sales (X_1), price (X_2), advertisement costs in local newspaper (X_3) and presence of sales assistant (X_4) in hours per period. The data distribution can be approximated by $X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{pmatrix} \sim N_4(\mu, \Sigma)$ where $\mu = \begin{pmatrix} 172 \\ 104 \\ 105 \\ 94 \end{pmatrix}$, $\Sigma = \begin{pmatrix} 1000 & -80 & 1100 & 275 \\ -80 & 500 & 90 & -90 \\ 1100 & 90 & 1500 & 200 \\ 275 & -90 & 200 & 720 \end{pmatrix}$, $R = \begin{pmatrix} 1 & -0.11317 & 0.89815 & 0.3241 \\ -0.11317 & 1 & 0.1039 & -0.15 \\ 0.89815 & 0.1039 & 1 & 0.19245 \\ 0.3241 & -0.15 & 0.19245 & 1 \end{pmatrix}$. Determine:

a) The conditional distribution of X_1 given $\begin{pmatrix} X_2 \\ X_3 \\ X_4 \end{pmatrix} = \begin{pmatrix} 100 \\ 100 \\ 100 \end{pmatrix}$. Hence find the best linear approximation (w.t. to minimal mean-squared error) of sales when $X_2 = X_3 = X_4 = 100$.

b) The squared multiple correlation of the sales with the remaining variables $(\rho^2_{1.234})$. Compare it to ρ^2_{12} and comment.

c) The conditional distribution of (X_1, X_2) given $\begin{pmatrix} X_3 \\ X_4 \end{pmatrix} = \begin{pmatrix} 100 \\ 100 \end{pmatrix}$.