STAT3064/STAT4067 Semester 2, 2021
Assignment 1

Due date: Tuesday 17 August 10pm

- Working in **teams** of up to 3 students is strongly encouraged. If you work in a team, then submit only one assignment per team, list the team members by name and student id and state what each team member has contributed to the assignment.
- Assignments to be submitted to the LMS as a **pdf** file. Up to two versions can be submitted, but only the last version will be marked.
- Do **not** provide R output as part of your assignment answers unless asked to do so. Answers hidden within code may not be easily visible and can be overlooked.
- Use `prcomp` with the raw or centred data in Questions 3 and 4.
- Marking of late assignments will follow the university rules.

Assignment Questions

Data required for this assignment can be found in the Data Sets folder.

1. (a) Why is it important to describe the mathematical model we want to simulate from and why should one not automatically choose the Gaussian model for a simulation?
   (b) Give two reasons why simulations should be reproducible.
   (c) Give details of the variance calculation $var(W_2)$ for Example 4.2 similar to those for $var(W_1)$.
2. Consider the aircraft data with the logged variables as in Question 2 of Computer Lab 1. Divide the data into the three period groups. We are interested in comparing changes that occur over time.
   (a) Show smoothed histograms of logLength and logPower separately for the three periods. Comment on the shapes of the histograms and how the change over time affects this shape.
   (b) Construct contour plots of the 2D smoothed histograms of the pairs (logPower, logWeight) and (logSpeed, logLength). Describe the shapes of the density plots and discuss how they change over time.
   (c) For which pair of variables would you expect the largest change in correlation or shape of their density over time and why?
3. Consider the aircraft data of Q2 of this assignment.
   (a) Separately for each period, carry out a principal component analysis using `prcomp` based on the raw data.
   (b) Show eigenvalue plots for each period. Interpret the results.
   (c) Show score plots of the first two PCs for each period. Comment on the results.

(d) Which logged variable contributes most to PC_1 for each period? Does this change across the three periods? Comment on the results.

(e) Based on your analysis, discuss the main changes that have occurred over time.

4. The data set `ass2pop.csv` is available in the LMS folder 'Data sets'. It contains the means and covariance matrices corresponding to two populations. The first and second column of `ass2pop.csv` are the means μ_1 and μ_2 of the first and second population respectively; columns 3:22 correspond to the covariance matrix Σ_1 of the first population, and the remaining columns correspond to the covariance matrix Σ_2 of the second populations. In this question we generate random samples from these populations as described below.

(a) Read the data into R. What is the dimension of the covariance matrix Σ_1?

(b) Generate 250 random samples from the Gaussian distribution $\mathcal{N}(\mu_1, \Sigma_1)$ and 250 samples from the Gaussian distribution $\mathcal{N}(\mu_2, \Sigma_2)$. Calculate the sample covariance matrix S of these 500 random samples, and find eigenvalues of S. Save the vector of eigenvalues into a file for later analysis.

(c) Repeat part (b) another 49 times, so you have a total of 50 vectors of eigenvalues.

(d) Calculate the mean vector of eigenvalues over the 50 repetitions and list/print this mean vector.

(e) Display the 50 vectors of eigenvalues and their mean vector in an eigenvalue or scree plot.

(f) Repeat parts (b) to (e) with 250 samples from the t-distribution $t_{10}(\mu_1, \Sigma_{01})$ and 250 samples from t-distribution $t_4(\mu_2, \Sigma_{02})$. (Hint: Σ_{0k} is the scale matrix which is obtained from the covariance matrix Σ_k using the following relationship $\Sigma_k = \frac{\nu}{\nu-2}\Sigma_{0k}$, with ν the degree of freedom of the t-distribution and $k = 1$ and 2 here.)

(g) Compare the results of the two different simulations and comment on interesting findings and differences between them the results obtained from the two sets of simulations.