MAST90138 S2 Assignment 1

Instructions:

- This assignment counts towards 15% of the final mark for the subject. If you LaTeXand `knitr` your assignment in a nice way, you will potentially get up to a maximum of 0.75% towards the final mark for the subject as extra credits. (You may use R Markdown as well.)
- Use tables, graphs and concise text explanations to support your answers. Unclear answers may not be marked at your own cost. All tables and graphs must be clearly commented and identified.
- No late submission is allowed.

Data: In this assignment you will analyze some wheat data. The dataset is available in .txt format on the LMS along with this assignment. The data come from three different varieties of wheat denoted by 1 to 3 in the dataset. Each row of the dataset corresponds to a different wheat kernel. Seven numerical characteristics were measured on the data: X1: area, X2: perimeter X3: compactness X4: length of kernel, X5: width of kernel, X6: asymmetry coefficient X7: length of kernel groove, whereas the eighth variable X8 contains values 1, 2 or 3 dependent on the variety of wheat the kernel comes from.

Problem 1 [12 marks]:

(a) State, explicitly, all possible values that a and b can take in order for the following matrix to be a covariance matrix. Give arguments that justify your answer :

$$\Sigma = \begin{pmatrix} 1 & 2 \\ a & b \end{pmatrix}.$$

[3 marks]

(b) Compute explicitly and without using R, all the orthonormal eigenvectors and the eigenvalues of the matrix

$$\Sigma = \begin{pmatrix} 13 & -4 \\ -4 & 7 \end{pmatrix}.$$

Give explicitly an orthogonal matrix Γ and a diagonal Λ such that we can write

$$\Sigma = \Gamma \Lambda \Gamma^T.$$

[3 marks]

(c) Read the wheat data in R and create a data matrix X of size $n \times p$, where $n = 210$ and $p = 7$, which contains the seven attributes $X1$ to $X7$ described above from all n kernels; please explicitly display the dimension of the data matrix X. Then create a vector of length n which contains, for each kernel, the wheat variety it comes from, coded 1 to 3 as described above. If you use the menu in R studio to read your data, please print out the corresponding instructions (they are given by R studio). Be careful with the "separator" you should use to read your file. [3 marks]

(d) Using R, for the unbiased sample covariance matrix S of X at (c), give explicitly an orthogonal matrix Γ and a diagonal matrix such that we can write

$$S = \Gamma \Lambda \Gamma^T$$

[3 marks]

Problem 2 [8 marks]:

In our lecture and Ch.5 of Härdle and Simar, we briefly discussed the Hotelling's T^2 test, which is a multivariate generalization of the univariate t-test. We should get a taste of how it is done in R. You are expected to use the `help()` function in R to learn the suggested R functions below. The data frame `pulmonary` in the `ICSNP` package of R measures the difference in pulmonary function in 12 workers after being exposed to cotton dust for 6 hours. There are three measurements: forced vital capacity, forced expiratory volume, and closing capacity. For convenience we will let $Y_i = (Y_{i1}, Y_{i2}, Y_{i3})^T$ be the vector of observations for each worker. We will apply the Hotelling's T^2 test to see whether the means of the three variables are all zero, i.e. $E[Y] = 0$.

(a) Install the R package, load the data frame and make a scatterplot for it. The function `data()` in R can be used to load the dataset. [2 marks]

(b) One assumption of the Hotelling's T^2 test is that the data come from a multivariate normal distribution. If this assumption is valid, we would expect the squared Mahalanobis distances

$$(Y_i - \bar{Y})' S^{-1} (Y_i - \bar{Y})$$

to roughly follow a chi-squared distribution with 3 degree of freedom, where S is the unbiased sample covariance matrix of the Y_i's. Use this fact to check with R whether normality holds. What is your conclusion? [3 marks]

(Suggested functions to use: `mahalanobis`, `qqnorm`, `pchisq`, `pnorm`. Recall that for a random variable X with a continuous cumulative distribution function $C(\cdot)$, the variable $C(X)$ is uniformly distributed.)

(c) Regardless of your conclusion above, we will proceed with the Hotelling test. Do this by using the function `HotellingsT2` in R which automatically gives a p-value for the test. Report this p-value. Next, compute this same p-value "manually" as follows: Compute the T^2 statistic using elementary matrix operations in R, and calibrate the p-value using the function `pf()`, based on Theorem 5.9 in Härdle and Simar. Be careful with the degrees of freedom. [3 marks]