MAST90138 – Semester 2, 2023. Assignment #1

Instructions:

- This assignment contains 2 problems with a total of 24 +1 marks which count towards 15% of the final mark for the subject.
- Assignment is due by 5pm on Monday 11 September, 2023. You must complete the online plagiarism form on LMS by 5pm on Monday 11 September, 2023.
- **[1 mark]** Your assignment should clearly show your name and student ID number, your tutor's name and the time and day of your tutorial class. Your answers must be clearly numbered and in the same order as the assignment questions. Your answers must be easy to read (marks may be deducted for illegible handwriting). Include all of your working out in your answers. All R outputs, including graphs and tables, must be accompanied by your concise and clearly written R code used to produce it. Any graph, table or R code must be accompanied by clear and concise comments.
- Use tables, graphs and concise text explanations to support your answers. All tables and graphs must be clearly commented and identified.
- All R code should be clearly written and commented. Uncommented R code is not acceptable.
- Comments should be brief and concise: marks will be awarded for clarity.
- No late submission is allowed.
- Your lecturer may not help you directly with assignment questions, but may provide some appropriate guidance.

Data: In the assignment you will analyse some wheat data. The dataset is available in `.txt` format on the LMS web page within the *Assignments* menu. 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 which each code the variety of wheat the kernel comes from (there are three varieties).

Problem 1 [10 marks]

(a) **[3 marks]** Give 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:

$$\Sigma = \begin{pmatrix} 1 & a \\ 3 & b \end{pmatrix}.$$

(b) **[5 marks]** Compute explicitly and without using any software, all the orthonormal eigenvectors and the eigenvalues of the matrix

$$\Sigma = \begin{pmatrix} 4 & -\sqrt{3} \\ -\sqrt{3} & 2 \end{pmatrix}.$$

You must prove how you obtain the eigenvalues and eigenvectors using detailed mathematical derivations and using only results seen in chapter 1 in class.

Again using only results seen in chapter 1 in class, give explicitly an orthogonal matrix Γ and a diagonal Λ such that we can write

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

(c) **[1 mark]** 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, for all n kernels, the variables X1 to X7 described above. Also 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 menus in R studio to read your data, please print out the corresponding instructions (they are given by R studio). Your matrix and vector may not contain NAs; you are allowed to manipulate the data file before using it but then you need to describe here what you did to it.

(d) **[1 mark]** Using R, for the covariance matrix S of `X` at (c), give explicitly an orthogonal matrix Γ and a diagonal matrix Λ such that we can write

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

Problem 2 [14 marks]

(a) **[3 marks]** Perform a principal component analysis of the wheat data. Explain why only X1 to X7 (and not X8) can be included in the principal component analysis. Store the eigenvalues of the covariance matrix in a vector called `lambda` and the eigenvectors in a matrix called `gamma`. What percentage of the variability of the data does each principal component explain? Also compute the cumulative percentages of variance $\psi_1, \dots \psi_7$ defined in class and draw a screeplot for these data. How many principal components does this suggest we should keep?

(b) **[3 marks]** Give explicitly the linear combinations of the seven variables used in the wheat example to create the first and second principal components and describe which variables have the most weight in the construction of those two PCs. Do the coefficients of these linear combinations depend on the wheat kernel (i.e. depend on the index i of the kernel)? Do the principal components depend on the wheat kernel (i.e. depend on the index i of the kernel)?

(c) **[3 marks]** Draw a scatterplot of the first two principal components of all individuals, using colours to identify different groups of data (include a clear legend). Describe what you can extract from this graph. Which groups are visible on the graph? What do they correspond to? How do the original variables contribute to those groups?

(d) **[5 marks]** Using the formula given in class, but replacing each population quantity by its empirical estimator, compute the correlation matrix that contains the correlations between each principal component and the 7 original variables X1 to X7. Draw the correlation graph showing the correlations between those variables and the first two PCs, using for each an arrow to represent the correlations with the first two principal components as in the correlation circles shown and discussed in the slides, and indicate the names of the variables near each arrow as done in the examples shown in class. Add to your graph a circle of radius 1 centered at the origin. Use this and the other results of your PC analysis to describe and interpret further the results of the principal component analysis, explicitly discussing the original variables, the groups of individuals, and the connection between these two.

Hints: 1) To draw an arrow in R, use the command `arrows` 2) To add some text to a graph in R, used the command `text(x,y,yourtext)` where x and y are the x and y coordinates of where to write your text and yourtext is the text you want to write there. 3) To add a circle to a graph, use

```
radius <- 1
theta <- seq(0, 2 * pi, length = 200)
lines(x = radius * cos(theta), y = radius * sin(theta))
```