MAST90138 Assignment 2

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 Problem 1 you will work on the same wheat data from Assignment 1. 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 [15 marks]:

(a) Perform a principal component analysis of the wheat data (only for the attributes X1 to X7!). Store the eigenvalues of the covariance matrix in a vector called `lambda` and the eigenvectors in a matrix called `gamma`; you can use `prcomp`() in R for the task. What percentage of the variability of the data does each principal component explain? Also compute the cumulative percentages of variance $\psi_1, \ldots, \psi_7$ defined in class and draw a screeplot for these data. How many principal components does this suggest we should keep (according to the screeplot)? [3]

(b) Give explicitly the linear combinations of the original data used in this example to create the first and second principal components and give an interpretation of these linear combinations, describing which variables play the biggest roles in the construction of those two PCs. [2]

(c) Draw scatterplots of the first 2 principal components, using colours to identify different groups of data. Describe what you can extract from the plots. Which groups are visible? What do they correspond to? How do the two PCs contribute to those groups? [5]

(d) 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 each original variable. Draw the correlation graph showing the correlations between the original variables $X1$ to $X7$ and the first two PCs. For each of the seven original variables, use an arrow to represent the correlations with the first two principal components as in the correlation picture shown in class, and indicate the names of the variables near each arrow as done in the example 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 further the results of the principal component analysis, explicitly discussing the original variables, the groups of individuals, and the connection between these two. [5]

Hints: To draw an arrow in R, use the command `arrows`. 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. 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))
```

Problem 2 [8 marks]:

(a) Compute the orthogonal factor model with a single factor ($k = 1$) for the covariance matrix

$$\Sigma = \begin{pmatrix} 1 & 0.9 & 0.7 \\ 0.9 & 1 & 0.4 \\ 0.7 & 0.4 & 1 \end{pmatrix},$$

i.e. find the unique (up to a rotation) factor loadings and specific variances. What's wrong with your solution? (This is Exercise 12.5 from Härdle and Simar. The moral of this exercise is that even if you can "solve" for an orthogonal factor model, the solution may not make sense. Example 12.1 in the book may help.) [3]

(b) In this problem we will work on the `Harman23.cor` in the `datasets` package. Type `help(Harman23.cor)` to learn about this dataset carefully, and look at the examples there. We will use the function `factanal` in `R` to fit a normal factor analysis model to the dataset.

(i) For this dataset, what is maximum number of factors we can fit so that we won't "overparametrize" the model? Explain your answer. [2]

(ii) Based on your answer in (i), use the `factanal` function to fit all possible normal factor models that do not overparametrize. The chi-square statistics reported are the likelihood ratio (LR) statistics with Bartlett's correction; see p.370 of Härdle and Simar. Based on these LR statistics, which factor model has the best fit? [3]