35459 Multivariate Statistics — Assignment 2

- The assignment is worth 40% of the marks for the subject.
- Note that it may not be possible to accept a late assignment, and that a late assignment may score zero.
- If a late submission is accepted, then it will be penalised by reducing the maximum attainable mark by 2% per day, or part thereof. e.g. after one day, it will be worth a maximum of 38 %.
- For each question, you are expected to perform (only) the analysis listed. Write up your analysis as a short report. Any graphics that you use to support your discussion should also be included.
- The R code that you use to produce the analysis must also be submitted with your assignment as an appendix. Your code must show all working to get full marks; it is not sufficient just to give the answer.

PART A

Question 1

(12 marks)

Take $Y = (Y_1, Y_2, Y_3)^\top$ to be the same as in the previous assignment. That is, Y has a multivariate normal distribution, with zero mean $(0, 0, 0)^\top$, and with covariance matrix

$$\Sigma = \frac{1}{5630} \begin{pmatrix} 575 & -60 & 10 \\ -60 & 300 & -50 \\ 10 & -50 & 196 \end{pmatrix}.$$

Recall that you performed a linear regression to find the coefficient $\beta_{2,1}$ in

$$Y_2 = \beta_{2,1} Y_1 + \epsilon_2$$

where $\hat{Y}_2 = \beta_{2,1} Y_1$ is the linear least squares predictor of Y_2 based on Y_1, and where $\epsilon_2 = Y_2 - \hat{Y}_2$ is the prediction error. You also performed a linear regression to find the coefficients $\beta_{3,1}$ and $\beta_{3,2}$ in

$$Y_3 = \beta_{3,1} Y_1 + \beta_{3,2} Y_2 + \epsilon_3$$

where $\hat{Y}_3 = \beta_{3,1} Y_1 + \beta_{3,2} Y_2$ is the linear least squares predictor of Y_3 based on Y_1 and Y_2, and where $\epsilon_3 = Y_3 - \hat{Y}_3$ is the prediction error.

a) Give the numbers $\beta_{2,1}$, $\beta_{3,1}$ and $\beta_{3,2}$.

b) Estimate $\sigma_2^2 = \text{var}(\epsilon_2)$.

c) Estimate $\sigma_3^2 = \text{var}(\epsilon_3)$.

d) Form a 3×3 matrix T

$$T = \begin{pmatrix} 1 & 0 & 0 \\ -\beta_{2,1} & 1 & 0 \\ -\beta_{3,1} & -\beta_{3,2} & 1 \end{pmatrix}.$$

e) Form the matrix

$$T\Sigma T^\top.$$

f) Form the matrix S^{-1} in

$$S^{-1} \equiv T^\top D^{-1} T.$$

Here D is a 3×3 diagonal matrix, with entries on the main diagonal $\sigma_1^2, \sigma_2^2, \sigma_3^2$. The number σ_1^2 is *for you to find.* Choose σ_1^2 to make S^{-1} as close to Σ^{-1} as possible.

Question 2

(6 marks)

a) Use the stock-price data in Table 8.4 of Johnson and Wichern (stockdata.csv). Perform a factor analysis for the stock-price data using the sample correlation matrix.

(i) Choose the value of m (justify your choice) and the method of extraction.

(ii) Try to interpret the factors that you get. (It may help to draw appropriate plots, or to consider rotation of the factors.)

(iii) How do these factors compare with those given in the chapter from the sample correlation matrix? (See table 9.8 on page 510 of Johnson and Wichern)

Question 3

(12 marks)

Refer to the male Egyptian skull data in egyptskull.csv. These data are from skulls from 5 epochs, as indicated. The measurements are: Maximum Breadth, Basibregmatic Height, Basialveolar Length and Nasal Height.

a) Give a brief explanation of Logistic Regression.

b) Give a brief explanation of Classification Trees.

c) Choose two of the four variables and draw a scatterplot with the Epoch of each point indicated. Comment.

d) Using the first 25 observations in each epoch as the training data set and the last 5 observations in each epoch as a test data set, generate classification rules using the following methods:

- Linear discriminant analysis
- Quadratic discriminant analysis
- Multinomial Logistic Regression
- Classification and Regression Trees (CART). You may use the R package "tree" for this purpose.
- Neural Networks

e) Using the confusion matrix and the apparent error rate, compare the effectiveness of each of the classification rules in making predictions for the test data.

f) Classify the 8 points below using the best rule that you feel you have developed above.

128, 143, 103, 50	129, 126, 91, 50	130, 127, 99, 45	130, 131, 98, 53
134, 124, 91, 55	130, 130, 104, 49	134, 139, 101, 49	136, 133, 91, 49

PART B (25 marks)

Question 1

(6 marks)

Use the same "diabetes" data from the previous assignment. Use the following R listing to experiment with the LASSO: Least Absolute Shrinkage and Selection Operator. You will need two R packages: `lars` and `glmnet`.

```
##########################################################
# Example of the LASSO:
# Least Absolute Shrinkage and Selection Operator.
##########################################################
# You need both these packages:
library(lars)
library(glmnet)
# Load the illustrative "diabetes" data
# and extract the response vector
# "yVector" and predictor matrix "Xmatrix":
data(diabetes)
Xmatrix <- diabetes$x
yVector <- diabetes$y

# Obtain the LASSO fit.
# You should experiment with different values
# of the regularisation ("lambda") paramter:
LASSOfit <- glmnet(Xmatrix, yVector, lambda=1)
betaHat <- as.numeric(LASSOfit$beta)
```

a) Experiment with different choices of the parameter 'lambda.'

b) What do you find?

c) Compare the LASSO fit you find now, with the fit that you found in the previous assignment.

Question 2

(8 marks)

Find **one** dataset using online sources that you can use to demonstrate the techniques that you

have learned in this subject. A good place is the UC Irvine Machine Learning Repository, or Kaggle:

`http://archive.ics.uci.edu/ml/`
`https://www.kaggle.com/`

a) For your chosen dataset, estimate a covariance matrix. Describe the method you used.

b) What other methods could you use to estimate a covariance matrix?

c) Suppose some entries of the inverse of the covariance matrix are known to be zero, but that you don't know precisely which entries. (i) Would this effect your choice of method to estimate the covariance matrix? (ii) Discuss any connections you see between this situation and the example of Question 1 in Part A.

Question 3

(6 marks)

Using the same dataset as in Question 2, you now need to pose one research question that you believe you can address using this dataset. You then need to select appropriate techniques, and apply those techniques, to analyse the data to address the research question that you have posed. Finally, you will need to reflect on the adequacy of the dataset to address the questions that you have posed, and make suggestions about how you might collect the data differently to better address your question (consider what to collect or how to collect, for instance).

Your answer to this question should include:

- A report that describes the data, poses the research question, analyses the research question, and reflects on the usefulness of the data to answer the question. This should be in a report format, with essential output in the report, and any other output that you use in an appendix. You should also indicate where you obtained the data from (e.g. paper reference or URL).
- A .R file containing your code.
- A .csv file containing the data set