ST464/ST674
Assignment 2
Katarina Domijan
Due on Thursday 14th March 1pm

- Do all questions. Submit questions: 2, 3 and 4.
- Use R markdown to knit your results to a .pdf (or html) file. Print and submit to your tutor's box.
- Place your name and student number under author in the YAML header. E.g.

```
---
title: "Assignment 2"
output: pdf_document
author: Jane Doe 1234567
---
```

- There will be a tutorial on Monday 11th March.

1. For the data matrix below

$$\mathbf{X} = \begin{bmatrix} 4 & 2 \\ 1 & 0 \\ -1 & -1 \\ -3 & 5 \\ 1 & -1 \end{bmatrix}$$

   (a) Calculate the sample variance-covariance matrix.
   (b) Calculate the correlation matrix.
   (c) Standardize the variables to have mean 0 and standard deviation 1.
   (d) In R find the eigenvectors of the correlation matrix of x.
   (e) Using prcomp() function, find the loadings for the principal components of x.

   For a) to c) do all the calculations by hand and check your answers in R.

2. Body fat data. The data consists of observations taken on a sample of 88 males. In this question you will look at PCA of the variables variables were measured:

Neck circumference (cm)	Abdomen circumference (cm)
Knee circumference (cm)	Ankle circumference (cm)

   (a) Use pairs to construct a scatterplot matrix. Are there any outliers? If so, which cases are they?

```
bfat <- read.table("yourfolder/bodyfat.txt", header=T)
bfat <- bfat[,c("neck","abdomen", "knee", "ankle")]
```

   (b) Carry out a principal components analysis of the data. What percentage of the variability in the dataset is accounted for by the first component? What percentage of the variability in the dataset is accounted for by the first two components? Examine the scree diagram and comment. (You will find the code for the screeplot in h1code.R).
   (c) What does the first component measure? the second component? Make a biplot to assist your interpretations. Are there any outliers? What can you say about the outliers from the plot?

(d) Omiting any outliers identified, repeat parts (b) and (c).

3. A 1902 study obtained measurements on seven physical characteristics for each of 3000 criminals. The seven variables measured were (1) head length (2) head breadth (3) face breadth (4) left finger length (5) left forearm length (6) left foot length (7) height. Using the correlation matrix given below, find the principal components of the data and interpret the results. What percentage of the variability in the dataset is accounted for by the first component? What percentage of the variability in the dataset is accounted for by the first two components? Examine the scree diagram and comment.

$$\begin{bmatrix} 1.000 & & & & & & \\ 0.402 & 1.000 & & & & & \\ 0.396 & 0.618 & 1.000 & & & & \\ 0.301 & 0.150 & 0.321 & 1.000 & & & \\ 0.305 & 0.135 & 0.289 & 0.846 & 1.000 & & \\ 0.339 & 0.206 & 0.363 & 0.759 & 0.797 & 1.000 & \\ 0.340 & 0.183 & 0.345 & 0.661 & 0.800 & 0.736 & 1.000 \end{bmatrix}$$

```
# read in the correlation data as a vector

crimcorr <- matrix(c(
  1.000, 0.402, 0.396, 0.301, 0.305, 0.339, 0.340,
  0.402, 1.000, 0.618, 0.150, 0.135, 0.206, 0.183,
  0.396, 0.618, 1.000, 0.321, 0.289, 0.363, 0.345,
  0.301, 0.150, 0.321, 1.000, 0.846, 0.759, 0.661,
  0.305, 0.135, 0.289, 0.846, 1.000, 0.797, 0.800,
  0.339, 0.206, 0.363, 0.759, 0.797, 1.000, 0.736,
  0.340, 0.183, 0.345, 0.661, 0.800, 0.736, 1.000), nrow = 7, byrow = TRUE)
colnames(crimcorr)<- c("Head-L","Head-B","Face-B",
                       "L-Fing","L-Fore","L-Foot",
                       "Height")
```

4. For each of the following situations, answer, if possible: (i) Is it a classification or regression problem? (ii) Are we most intererest in inference or prediction? (iii) Provide n and p. For each predictor described state whether it is categorical or quantitative. (iv) Indicate whether we would expect the performance of a flexible learning method to be better or worse than an inflexible method.

   (a) We have a set of data on 500 worldwide tech firms. For each firm, information on profit, CEO salary, number of employees, average employee salary, and home country is recorded. We are interested in the relationship between CEO salary and other measurements.

   (b) A company wishes to launch a new product. They want to know in advance whether it will be a success or failure. They collect data on 20 similar products, and record whether they succeeded or not, price charged, marketing budget, and 10 other variables.

   (c) A dataset was collected to related the birthweight of babies to the days of gestation and gender.

   (d) Observations were collected on 56 attributes from 32 lung cancer patients belonging to one of 3 classes.

5. In this exercise you will conduct an experiment to compare the fits on a linear and flexible model fit. You will use the Auto data from the package ISLR and explore the relationship between the response mpg with weight and horsepower.

(a)
```
# install.packages("ISLR") #home computer, first time only
library(ISLR)
?Auto
Auto <-Auto[complete.cases(Auto[,c(1,4,5)]),] # to remove NAs
```

Plot the response (miles per gallon) vs weight and horsepower. What do they tell you about the relationship between mpg and the predictors?

(b)
```
# install.packages("plot3D") #home computer, first time only

library(plot3D) # install package
scatter3D(Auto$weight,Auto$horsepower,Auto$mpg)

library(plot3Drgl)
scatter3Drgl(Auto$weight,Auto$horsepower,Auto$mpg)
```

Make a 3d plot of weight, horsepower and mpg (see commands above). What do they tell you about the relationship between mpg and the predictors?

(c) Next, divide the data into a training set and a test set as follows:

```
set.seed(123)
train <- sample(nrow(Auto), round(.8*nrow(Auto)))
AutoTrain <- Auto[train,]
AutoTest <- Auto[-train,]
```

Fit a linear regression model to mpg versus weight and horsepower on AutoTrain. Call the fit f1. Examine summary(f1) and comment on the significance of the predictors.

(d) Plot the fitted surface and the data. (See lecture notes for code). Does the linear surface look like a good fit?

(e) Use loess to fit a surface to the same data. Call the fit f2. Plot the fitted surface and the data. Does the loess surface look like a good fit?

(f) Calculate the MSE for both fits on the training data. What do these numbers tell you? (See lecture notes for code.)

(g) Calculate the MSE for both fits on the test data. What do these numbers tell you?