MAM5220/MA35210 Principal Component Analysis workbook

You will be assessed on your answers to questions 2, 3 and 5

Question 1 (PRACTICE QUESTION - NOT ASSESSED)

This question makes use of a meteorological data set (Dagnelie, 1975). Data are available for 11 years, 1920-1931. The five measured variables are:

X1 rainfall in November and December (mm)

X2 average July temperature (degrees Centigrade)

X3 rainfall in July (mm)

X4 radiation in July (in mm of alcohol)

X5 average harvest yield (in quintals per hectare)

The data are contained in the data frame `meteorology` (11 rows and 5 columns)

Read in the data and perform principal component analysis (on the covariance matrix). I recommend you set up an RStudio project for this workbook with folders called `data`, `scripts` and `figs` and store the data sets in your `data` folder. You can then read in the data set and carry out principal component analysis using:

```
meteorology<-read.csv("data/meteorology.csv",row.names=1)
met.pca<-prcomp(meteorology)
```

a. Identify the following components of the `met.pca` object

   i. The centre of the data (`met.pca$center`)
   ii. The matrix of principal component scores (`met.pca$x`)
   iii. The principal component loadings (`met.pca$rotation`)
   iv. The square roots of the eigenvalues of the covariance matrix (`met.pca$sdev`)

b. Calculate the proportion of variability explained by each of the principal components (use your (squared) answer to part a) iv) combined with the result that the kth principal component explains $100\lambda_k / \sum_{i=1}^{p} \lambda_i \%$ of the variability, i.e. the ratio of the kkth eigenvalue to the sum of all of the eigenvalues). Create a vector called `proportions` that contains the five proportions explained by the five principal components.

c. Plot the scores on the first three principal components (PC2 vs. PC1), (PC3 vs. PC1), (PC3 vs. PC2) using `ggplot`.

For example, to plot PC2 against PC1, use

```
library(ggplot2)
met.pca.scores.df <- data.frame(met.pca$x)
p1 <- ggplot(met.pca.scores.df, aes(x=PC1, y=PC2)) +
  geom_point() +
  labs(x=paste0("PC1 (", 100*round(proportions[1], 2), "% of variability)"), y=paste0("PC2 (", 100*round(proportions[2], 2), "% of variability)")) +
  ggtitle("PC2 against PC1")
```

d. Repeat part a) using `scale=TRUE` as an argument to the `prcomp()` function. How do your answers to parts iii) and iv) differ from part a)? What is the scaling used here, i.e. how are the variables scaled? Part a) used the *covariance* matrix. What matrix are we using here?

Question 2

The data matrix for this question consists of the scores on five exams for a set of n=88 students on p=5 exams. The first two exams (Mechanics and Vectors) were closed-book exams, whereas the last three (Algebra, Analysis and Statistics) were open-book exams.

Have a look at the data set using the following code (you'll need to install the `bootstrap` library)

```
library(bootstrap)
head(scor)    # the data frame is called scor and the head() function shows you the top six rows – useful for getting a sense of the shape of a data set
```

a. Produce plots of the all the pairwise plots of scores on the first 3 principal components (PC2 against PC1, PC3 against PC1, PC3 against PC1). Comment on your plots. (Use `scale=TRUE` in the `prcomp()` function).

(4 marks)

b. Interpret the first two loadings vectors of the principal component analysis carried out in part a), relating the loadings vector for principal component 2 to the type of exam (open/closed)

(6 marks)

c. What are the eigenvalues of the correlation matrix of the `scor` data set?

(2 marks)

Here we consider a hypothesis testing procedure for testing for equality of eigenvalues. (Here we will assume that we are testing for equality of the final $p-kp-k$ eigenvalues, but the eigenvalues could be anywhere in the sequence.)

The null hypothesis we are interested in testing is

$$H_0 : \lambda_{k+1} = \lambda_{k+2} = \ldots = \lambda_p.$$
$$H_0 : \lambda_{k+1} = \lambda_{k+2} = \ldots = \lambda_p.$$

We write $l_1, l_2, \ldots, l_p l_1, l_2, \ldots, l_p$ for the observed eigenvalues and let

$$a_0 = (l_{k+1} + l_{k+2} + \ldots + l_p)/(p-k)$$
$$g_0 = (l_{k+1} \times l_{k+2} \times \ldots \times l_p)^{1/(p-k)},$$

where $a_0 a_0$ is the arithmetic mean of the sample estimates of the (hypothesized) repeated eigenvalue and $g_0 g_0$ is the geometric mean of the sample estimates.

If $H_0 H_0$ is true then approximately,

$$\left(n - \frac{2p+11}{6}\right)(p-k)\log\left(\frac{a_0}{g_0}\right) \sim \chi^2_{(p-k+2)(p-k-1)/2},$$
$$\left(n - \frac{2p+11}{6}\right)(p-k)\log\left(\frac{a_0}{g_0}\right) \sim \chi^2_{(p-k+2)(p-k-1)/2},$$

i.e. a chi-squared distribution on $(p-k+2)(p-k-1)/2(p-k+2)(p-k-1)/2$ degrees of freedom.

For kk taking values 3, 2 and 1, carry out appropriate hypothesis tests, clearly defining the null and alternative hypotheses in each case and stating the test statistics (the left hand side of the above equation) and their distributions. Clearly state your conclusions about which eigenvalues are significantly different.

(12 marks)
(Question = 24 marks)

Question 3

Here we present a number of ways of choosing a subset of principal components

Method 1: The cumulative proportion of variability explained.

If $l_1, l_2, \ldots, l_p l_1, l_2, \ldots, l_p$ are the observed eigenvalues, then the percentage of variability explained by the first k principal components is given by

$$100\frac{\sum_{j=1}^{k} l_j}{\sum_{j=1}^{p} l_j}\%,$$
$$100\frac{\sum_{j=1}^{k} l_j}{\sum_{j=1}^{p} l_j}\%,$$

i.e. $t_1 = 100 l_1 / \sum_{j=1}^{p} l_j \, t_1 = 100 l_1 / \sum_{j=1}^{p} l_j$, $t_2 = 100(l_1 + l_2) / \sum_{j=1}^{p} l_j \, t_2 = 100(l_1 + l_2) / \sum_{j=1}^{p} l_j$, etc.

If we choose a threshold, $t^* t^*$, e.g. $t^* = 80\% t^* = 80\%$, then we keep mm principal components, where mm is the smallest integer kk for which $t_k > t^* t_k > t^*$.

Method 2: The size of the variance of the principal components

While method 1 is equally valid whether a covariance matrix or a correlation matrix is being used, this method is specially for use with correlation matrices. The idea behind the method is that if all the variables are independent, then the principal components are the same as the original variables, and all have variance equal to one in the case of a correlation matrix. So any principal component with variance less than one contains less information than one of the original variables, and so is not worth retaining. However, sometimes rejecting principal components with variance less than 1 is a bit too strict, and may be replaced with rejecting principal components with variance less than $l^* l^*$, where $l^* l^*$ is a number less than 1. Jolliffe (1972) has suggested the threshold $l^* = 0.7 l^* = 0.7$.

Method 3: The scree graph

Here you look at a plot of $l_k l_k$ against kk (i.e. the kkth eigenvalue against kk), and decide at which value of kk the graph is 'steep' to the left of k k and 'not steep' to the right of kk. Note that this is also a subjective method for choosing the number of principal components to keep.

Use the data set `blood_chem` with each of the three methods above to identify how many principal components to keep. (You will need to use `scale=TRUE` – see what happens if you don't)

```
blood_chem<-read.csv("data/blood_chem.csv")
```

When using Method 1, what is the effect of setting the threshold at 70%, 80%, 90%?

When using Method 2, try the thresholds 1 and 0.7.

For Method 3, produce a plot of the kkth eigenvalue against kk.

(Question = 20 marks)

Question 4 (PRACTICE QUESTION - NOT ASSESSED)

This question uses the data set `Tibetan_skulls.csv`, which includes the following measurements on two sets of Tibetan skulls. Type A skulls (17 skulls) came from graves in Sikkim and neighbouring areas of Tibet. Type B (15 skulls) were picked up on a battlefield in the Lhasa district and were believed to be those of native soldiers from the eastern province of Khams.

(Note, all measurements are in mm.)

X1 Greatest length of skull

X2 Greatest horizontal breadth of skull

X3 Height of skull

X4 Upper face height

X5 Face breadth, between outermost points of cheek bones

Use the tools you have covered in this workbook to carry out an appropriate principal component analysis of the data set. Present your findings in a well-structured report, including relevant figures. Make an informed decision as to whether there are any significant differences between the two types of skull. Include an interpretation of the principal components and comment on the covariance/correlation matrices of the five variables.

Question 5

This question makes use of a data set collected in 1898 of measurements on 136 sparrows that were found freezing and brought to a laboratory at Brown University. Of the 136 sparrows, 72 survived and 64 died. The data set sparrows.csv contains the following variables:

X1 sex 1 if male, 0 if female

X2 survive 1 if sparrow survived, 0 if it died

X3 length Total length (mm)

X4 alar Alar (wings) extent (mm)

X5 weight Weight (g)

X6 lbh Length of beak and head (mm)

X7 lhum Length of humerus (mm)

X8 lfem Length of femur (mm)

X9 ltibio Length of tibio-tarsus (mm)

X10 wskull Width of skull (mm)

X11 lkeel Length of keel to sternum (mm)

Write a short (1-2 page) report carrying out a preliminary analysis of the data, based on the output of the R script `sparrows.R` which can be found on the Blackboard page, as well as a principal component analysis based on the correlation matrix. For the PCA, think carefully about which columns of the dataset you wish to include in the analysis.

(Since I am supplying you with the `R` code, there is no need to include figures generated directly from the code supplied. If you make additional figures you may include them in your answer.)

You may wish to include the following points:

- How much of the variability is explained by each PC?
- Make an informed decision about how many PCs to retain and justify your decision.
- Interpret the first two PCs. Are there differences between the survivors and those that died?

HINT: If you want to produce a plot of PC2 scores against PC1 scores coloured by whether the birds survived, use something like the following code:

```
sparrows.pca_scores.df <- data.frame(sparrows.pca$x)
sparrows.pca_scores.df$survive <- sparrows$survive
ggplot(sparrows.pca_scores.df,
       aes(x=PC1, y=PC2, col=factor(survive))) +
  geom_point() +
  labs(col = "survived")
```

You'll need to include the proportion of the variability explained by each principal component on the axis labels as in Question 1 as well as an appropriate title.

(Question = 20 marks)
(Total = 64 marks)