

Exercise 1

Tutorial Exercise.

1. Let $\mathbf{S} = \begin{pmatrix} 6.0 & 4.8 \\ 4.8 & 6.0 \end{pmatrix}$ be a sample covariance matrix.
 - (a) Determine the eigenvalues and eigenvectors of \mathbf{S} .
 - (b) If the first variable is a measurement and we record the observations in millimetres rather than centimetres then the covariance matrix becomes $\mathbf{S}_1 = \begin{pmatrix} 600 & 48 \\ 48 & 6 \end{pmatrix}$. Find the eigenvalues and eigenvectors of \mathbf{S}_1 .
 - (c) Determine the proportion of total variability explained by the first principal component in each of the above cases. Comment on the nature of the first principal components.

2. The output attached gives a principal components analysis for five anatomical variates of 49 female sparrows. The body measurements, in mm, are total length (x_1), alar extent (x_2), length of beak and head (x_3), length of humerus (x_4) and length of keel of the sternum (x_5). Birds numbered 1-21 survived the period of observation while birds 22-49 did not.
 - (a) Comment on why the principal components analysis is carried out using the correlation matrix.
 - (b) Calculate the eigenvalues of the correlation matrix. Check these values sum to 5.
 - (c) Give the proportion of variability explained by the first two principal components.
 - (d) What variables are highly correlated with the first two principal components?
 - (e) Use the correlations to find the missing values in the loadings output.
 - (f) What can you say about the survivors given the plot of the scores for the first two principal components?

3. An analysis was conducted of a data set consisting of 40 independent observations on \mathbf{X} , where \mathbf{X} is a vector consisting of six random variables. The correlation matrix had eigenvectors and eigenvalues given below.

Eigenvalues:

2.907 1.259 0.941 0.492 0.240 0.161

First 3 eigenvectors:

0.133	0.787	0.115
0.504	0.170	0.246
0.464	0.179	0.354
0.400	-0.492	0.278
0.333	0.160	-0.797
-0.497	0.230	0.297

- (a) What proportion of the total variability in the standardised data set is accounted for by the first principal component?
- (b) Calculate the correlations between the second principal component scores and the measurements on each of the six variables. What do you conclude from these?

Computer Exercise.

Protein consumption data for various European countries is provided in the file **protein** that can be downloaded from

```
protein = read.csv(file=url("http://www.maths.usyd.edu.au/  
u/UG/SM/STAT3014/r/Data/protein.txt"))
```

1. Use R to calculate the first two principal components based on the correlation matrix for this data set.
2. Give the amount of variability explained by the first three principal components.
3. Find the correlations between the first two principal components and the original variables and use these to attempt to interpret the components.
4. Find the covariance matrix for the component scores (rounded to 3 decimal places). Comment.
5. Determine the principal component scores and plot the first two component scores. Does the plot indicate some clustering of the European countries? Comment.