STAT 3014/ 3914

| Semester 2 | Applied Statistics | 2012 |
|---|---|---|

Exercise 1

**Tutorial Exercise.**

1. Let $\boldsymbol{S} = \begin{pmatrix} 6.0 & 4.8 \\ 4.8 & 6.0 \end{pmatrix}$ be a sample covariance matrix.
   (a) Determine the eigenvalues and eigenvectors of $\boldsymbol{S}$.
   (b) If the first variable is a measurement and we record the observations in millimetres rather than centimetres then the covariance matrix becomes $\boldsymbol{S}_1 = \begin{pmatrix} 600 & 48 \\ 48 & 6 \end{pmatrix}$. Find the eigenvalues and eigenvectors of $\boldsymbol{S}_1$.
   (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?

4. Consider K-means clustering. Show that

$$\frac{1}{|C_k|} \sum_{a,b \in C_k} \sum_{j=1}^{p} (x_{aj} - x_{bj})^2 = 2 \sum_{i \in C_k} \sum_{j=1}^{p} (x_{ij} - \overline{x}_{kj})^2$$

   where $\overline{x}_{kj} = \frac{1}{|C_k|} \sum_{i \in C_k} x_{ij}$ is the mean for feature $j$ in cluster $C_k$.

5. Suppose that you have the following 6 observations and want to apply K-means.

   | Obs. | $X_1$ | $X_2$ |
   |---|---|---|
   | 1 | 1 | 4 |
   | 2 | 1 | 3 |
   | 3 | 0 | 4 |
   | 4 | 5 | 1 |
   | 5 | 6 | 2 |
   | 6 | 4 | 0 |

   (a) Plot the points.

   (b) Suppose the initial cluster assignment is $C_1 = \{1, 3, 4\}$ and $C_2 = \{2, 5, 6\}$. Calculate the cluster means.

   (c) Given the cluster means in (b) reassign the clusters.

   (d) What are the final cluster means?

6. Consider the following dissimilarity matrix for use in hierarchical clustering.

$$\mathbf{D} = \begin{bmatrix} 0 & 0.3 & 0.4 & 0.7 \\ 0.3 & 0 & 0.5 & 0.8 \\ 0.4 & 0.5 & 0 & 0.45 \\ 0.7 & 0.8 & 0.45 & 0 \end{bmatrix}$$

(a) Merge the two sets with the two points which are most similar.

(b) Recalculate the dissimilarity matrix using complete linkage after using the merge step in part (a).

(c) What are the next two sets which should be merged?

(d) Use these to construct a hierarchical clustering dendrogram.