Semester 2	Applied Statistics	2016

Exercise 1 Solution

Tutorial Exercise.

1. (a) The eigenvalues of $\mathbf{S}$ satisfy

$$0 = |\boldsymbol{S} - \lambda \boldsymbol{I}| = (6-\lambda)^2 - 4.8^2 = (10.8-\lambda)(1.2-\lambda)$$

Hence, $\lambda_1 = 10.8$ and $\lambda_2 = 1.2$.
Eigenvector corresponding to λ_1 is

$$\begin{pmatrix} 6 & 4.8 \\ 4.8 & 6 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = 10.8 \begin{pmatrix} a \\ b \end{pmatrix}$$

So for $a = b$, $\boldsymbol{u}_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$.
Eigenvector corresponding to λ_2 is

$$\begin{pmatrix} 6 & 4.8 \\ 4.8 & 6 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = 1.2 \begin{pmatrix} a \\ b \end{pmatrix}$$

So for $a = -b$, $\boldsymbol{u}_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix}$.

(b) The eigenvalues of $\boldsymbol{S}_1$ are

$$0 = |\boldsymbol{S}_1 - \lambda \boldsymbol{I}| = (600-\lambda)(6-\lambda) - 48^2 = (\lambda^2 - 606\lambda + 1296)$$

$$\begin{aligned} \lambda &= \frac{606 \pm \sqrt{606^2 - 4 \times 1296}}{2} \\ &= (603.8538, 2.1462) \end{aligned}$$

That is $\lambda_1 = 603.85$ and $\lambda_2 = 2.14$.

Eigenvector corresponding to λ_1 is $\begin{pmatrix} a \\ b \end{pmatrix}$ with $600a + 48b = 603.85a$.

This gives $u_1 = \begin{pmatrix} 12.4552 \\ 1 \end{pmatrix}$ or normalised $e_1 = \begin{pmatrix} 0.9968 \\ 0.08 \end{pmatrix}$.

Eigenvector corresponding to λ_2 is $\begin{pmatrix} c \\ d \end{pmatrix}$ with $600c + 48d = 2.14c$.

This gives $\boldsymbol{u}_2 = \begin{pmatrix} 1 \\ -12.4552 \end{pmatrix}$ or normalised $\boldsymbol{u}_2 = \begin{pmatrix} 0.08 \\ -0.9968 \end{pmatrix}$.

(c) The proportion of total variability explained by the first PC is
Case a:
$$\frac{10.8}{12} = 0.9$$
Case b:
$$\frac{603.85}{606} = 0.9965$$

2. (a) The original variables are quite different even through each is a measurement. Variable x_2 has variance 25.68 whereas variable x_4 only has variance 0.318. If we did not standardise the data set before doing a PC analysis x_2 and x_1 would dominate the first PC simply because they have larger sample variance. We might miss some relevant structure following this work.

   (b) The correlation matrix has eigenvalues (to 4 d.p) $1.9016^2 = 3.61, 0.532, 0.386, 0.301, 0.165$
   $3.61 + 0.53 + 0.38 + 0.30 + 0.16 = 5$.

   (c) The first two PCs explain 72.3% and 10.6% of the total variability.

   (d) The first PC is evenly loaded on all five variables (all correlations ≥ 0.75) and so is a measure of overall size. PC 2 is highly correlated with x_5 (length of the keel of the sternum)(correlation 0.64). No other variables have a high correlation with PC 2.

   (e) Since correlation $= \sqrt{\lambda_i} u_{ij}$ we can calculate the eigenvector values by dividing the correlation by $\sqrt{\lambda_i}$.
   For Component 2, the loading are
$$\frac{0.03698}{0.72904} = 0.0507 \approx 0.051$$
   For Component 4, the missing value is
$$\frac{-0.03782}{0.5491498} \approx -0.069.$$

   (f) On the PC scores plot the survivor are labelled 1. These values are clumped in the middle of the plot. They are less variable (component by component) then the non-survivors as a group. This would indicate that very large and very small birds have a lower probability of survival.

3. (a) The first PC accounts for $2.907/6 = 0.4845$ of the total variability.

   (b) Correlations for 2nd PC and the original (standardised) variables are
$$\sqrt{1.259} \times \boldsymbol{u}_2 = (0.8831, 0.1907, 0.2008, -0.5520, 0.1795, 0.2581)^T.$$
   Thus the second PC is highly correlated with x_1 and has a high negative correlation with x_4. The remaining variables have little impact on PC 2.

4. (This was done in class and is here for those who did not attend the lecture).

$$\begin{aligned}
\frac{1}{|C_k|}\sum_{a,b\in C_k}\sum_{j=1}^{p}(x_{aj}-x_{bj})^2 &= \frac{1}{|C_k|}\sum_{a,b\in C_k}\sum_{j=1}^{p}(x_{aj}-\overline{x}_{kj}+\overline{x}_{kj}-x_{bj})^2 \\
&= \frac{1}{|C_k|}\sum_{a,b\in C_k}\sum_{j=1}^{p}(x_{aj}-\overline{x}_{kj})^2-2(x_{aj}-\overline{x}_{kj})(x_{bj}-\overline{x}_{kj})+(x_{bj}-\overline{x}_{kj})^2 \\
&= \sum_{a\in C_k}\sum_{j=1}^{p}\left[\sum_{b\in C_k}\frac{1}{|C_k|}(x_{aj}-\overline{x}_{kj})^2\right] \\
&\quad -\frac{2}{|C_k|}\sum_{j=1}^{p}\left[\sum_{a\in C_k}(x_{aj}-\overline{x}_{kj})\right]\left[\sum_{b\in C_k}(x_{bj}-\overline{x}_{kj})\right] \\
&\quad +\sum_{b\in C_k}\sum_{j=1}^{p}\left[\sum_{a\in C_k}\frac{1}{|C_k|}(x_{bj}-\overline{x}_{kj})^2\right]
\end{aligned}$$

Now note that

$$\sum_{a\in C_k}(x_{aj}-\overline{x}_{kj}) = \Big(\sum_{a\in C_k}x_{aj}\Big)-\Big(|C_k|\overline{x}_{kj}\Big)=0$$

since $\overline{x}_{kj}=\frac{1}{|C_k|}\sum_{i\in C_k}x_{ij}$. Hence, the cross term above (the middle term) is zero. Next note that

$$\sum_{a\in C_k}\sum_{j=1}^{p}\left[\sum_{b\in C_k}\frac{1}{|C_k|}(x_{aj}-\overline{x}_{kj})^2\right]=\sum_{a\in C_k}\sum_{j=1}^{p}\left[(x_{aj}-\overline{x}_{kj})^2\right]$$

since the term in the square brackets is unchanged over the index b. Similarly for the 3rd term in our initial expansion. We are left with

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

The result to be proved follows by noting that the two terms on the right hand side are equal.

5. (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.

   Solution: The cluster mean for C_1 is $(2,3)$ and for C_2 is $(11/3,5/3)$.

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

   Solution: The new cluster assignments are $C_1=\{1,2,3\}$ and $C_2=\{4,5,6\}$.

   (d) What are the final cluster means?

   Solution: The final cluster mean for C_1 is $(2/3,11/3)$ and for C_2 is $(5,1)$.

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

Solution: The initial clusters are $C_1 = \{1\}$, $C_2 = \{2\}$, $C_3 = \{3\}$ and $C_4 = \{4\}$. The initial merge step will merge C_1 and C_2 since these two clusters have the smallest dissimilarity (0.3).

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

Solution: The dissimilarity matrix is

	$C_{12} = \{1,2\}$	$C_3 = \{3\}$	$C_4 = \{4\}$
$C_{12} = \{1,2\}$	0	0.5	0.8
$C_3 = \{3\}$	0.5	0	0.45
$C_4 = \{4\}$	0.8	0.45	0

Note that the bottom right two entries of the dissimilarity matrix (corresponding to C_3 and C_4) remain the same because these two clusters did not merge. The remaining two entries are obtained from

$$0.5 = \max\{d(x,y) \text{ such that } x \in \{1,2\} \text{ and } y \in \{3\}\}$$

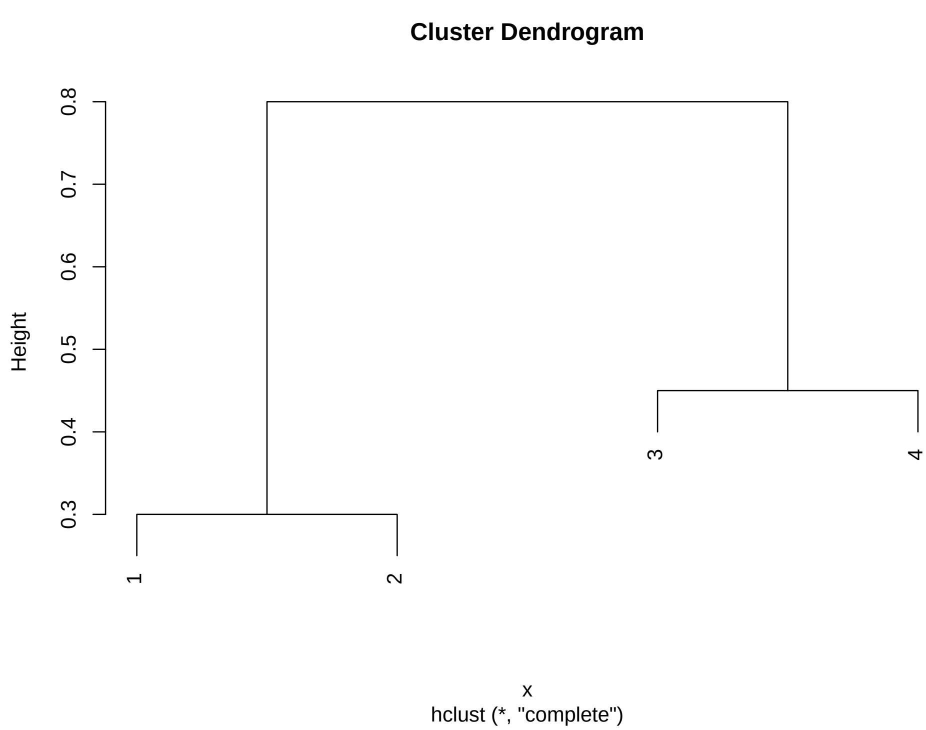
and

$$0.8 = \max\{d(x,y) \text{ such that } x \in \{1,2\} \text{ and } y \in \{4\}\}.$$

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

Solution: The next merge step will merge C_3 and C_4 since these two clusters have the smallest dissimilarity (0.45).

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



The height of the dendrogram corresponds to the minimum similarity within clusters. So the height for $C_{12} = \{1, 2\}$ is 0.3 since the similarity similarity within the cluster is 0.3. Similarly the height for $C_{12} = \{3, 4\}$ is 0.45 since the similarity similarity within the cluster is 0.45.

The dissimilarity matrix using complete linkage is

	$C_{12} = \{1, 2\}$	$C_{34} = \{3, 4\}$
$C_{12} = \{1, 2\}$	0	0.8
$C_{34} = \{3, 4\}$	0.8	0

So the two clusters "meet" at a hight of 0.8 since the maximum similarity between clusters is 0.8.