

ISLR | Chapter 10 Exercises

Marshall McQuillen

12/27/2018

Conceptual

1

NEED TO COME BACK TOO

- A. 10.12, illustrated below, is showing the the *within-cluster variation* is equal to twice the squared distance between each data point in cluster k (C_k) and that cluster's centroid, summed across all data points.

$$\frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^P (x_{i,j} - x_{i',j})^2 = 2 \sum_{i \in C_k} \sum_{j=1}^P (x_{i,j} - \bar{x}_{k,j})^2 \quad (1)$$

$$\frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^P (x_{i,j} - x_{i',j})(x_{i,j} - x_{i',j}) = 2 \sum_{i \in C_k} \sum_{j=1}^P (x_{i,j} - \bar{x}_{k,j})(x_{i,j} - \bar{x}_{k,j}) \quad (2)$$

$$\frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^P (x_{i,j}^2 - 2x_{i,j}x_{i',j} + x_{i',j}^2) = 2 \sum_{i \in C_k} \sum_{j=1}^P (x_{i,j}^2 - 2\bar{x}_{k,j}x_{i,j} + \bar{x}_{k,j}^2) \quad (3)$$

$$\frac{|C_k|}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^P (x_{i,j}^2 - 2x_{i,j}x_{i',j} + x_{i',j}^2) = 2|C_k| \sum_{i \in C_k} \sum_{j=1}^P (x_{i,j}^2 - 2\bar{x}_{k,j}x_{i,j} + \bar{x}_{k,j}^2) \quad (4)$$

$$\sum_{i,i' \in C_k} \sum_{j=1}^P (x_{i,j}^2 - 2x_{i,j}x_{i',j} + x_{i',j}^2) = 2|C_k| \sum_{i \in C_k} \sum_{j=1}^P (x_{i,j}^2 - 2\bar{x}_{k,j}x_{i,j} + \bar{x}_{k,j}^2) \quad (5)$$

$$\sum_{i \in C_k} \sum_{j=1}^P x_{i,j}^2 - 2 \sum_{i,i' \in C_k} \sum_{j=1}^P x_{i,j}x_{i',j} + \sum_{i' \in C_k} \sum_{j=1}^P x_{i',j}^2 = 2|C_k| \sum_{i \in C_k} \sum_{j=1}^P x_{i,j}^2 - 2|C_k| \sum_{i \in C_k} \sum_{j=1}^P 2\bar{x}_{k,j}x_{i,j} + 2|C_k| \sum_{i \in C_k} \sum_{j=1}^P \bar{x}_{k,j}^2 \quad (6)$$

$$\sum_{i \in C_k} \sum_{j=1}^P x_{i,j}^2 - 2 \sum_{i,i' \in C_k} \sum_{j=1}^P x_{i,j}x_{i',j} + \sum_{i' \in C_k} \sum_{j=1}^P x_{i',j}^2 = 2|C_k| \sum_{i \in C_k} \sum_{j=1}^P x_{i,j}^2 - 4|C_k| \sum_{i \in C_k} \sum_{j=1}^P \bar{x}_{k,j}x_{i,j} + 2|C_k| \sum_{i \in C_k} \sum_{j=1}^P \bar{x}_{k,j}^2 \quad (7)$$

$$(8)$$

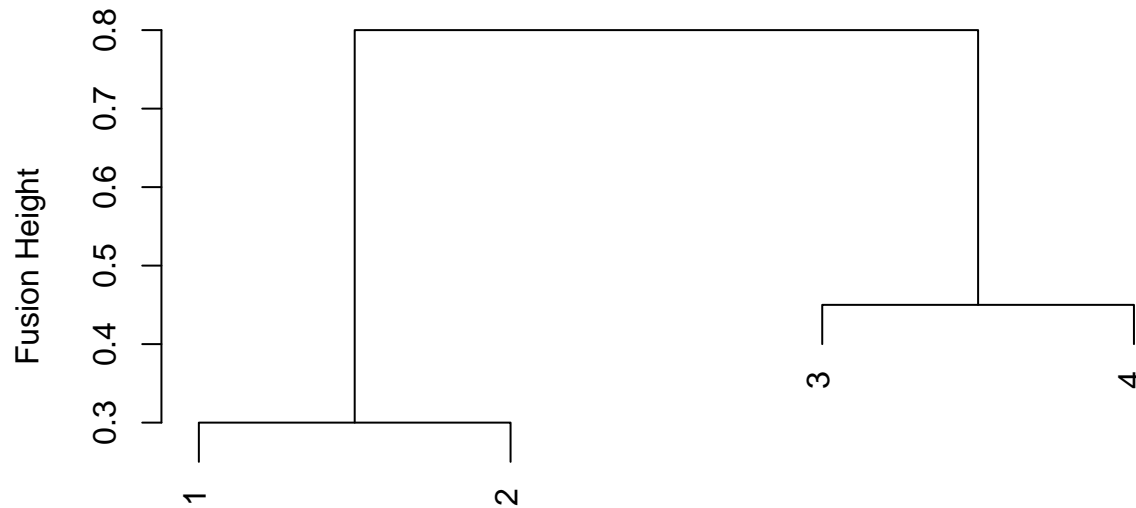
2

- A.

```
x <- matrix(c(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), ncol = 4, nrow = 4)
```

```
plot(hclust(as.dist(x), method = 'complete'),
     xlab = "",
     sub = "",
     ylab = "Fusion Height")
```

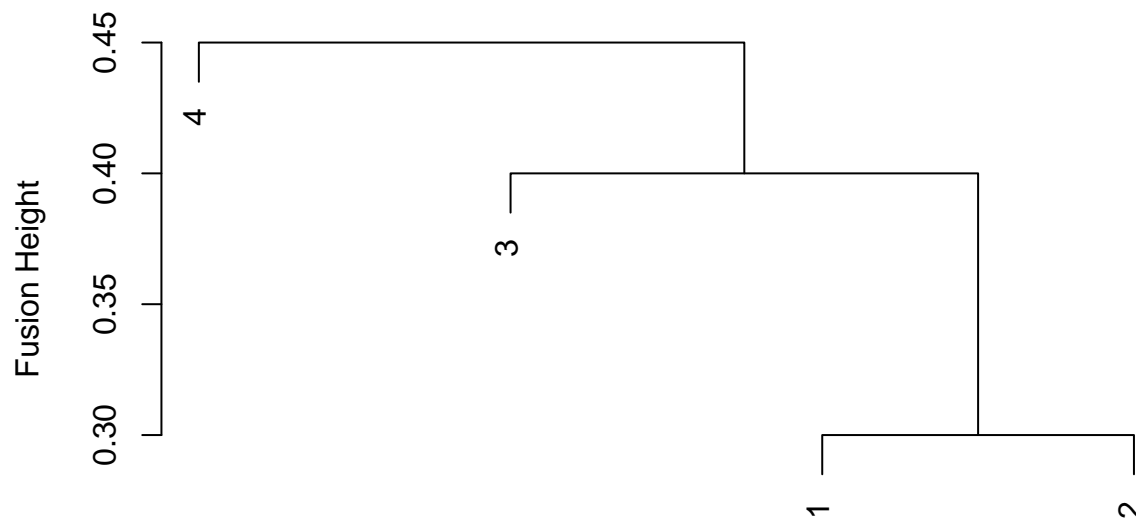
Cluster Dendrogram



- B.

```
plot(hclust(as.dist(x), method = 'single'),
     xlab = "",
     sub = "",
     ylab = "Fusion Height")
```

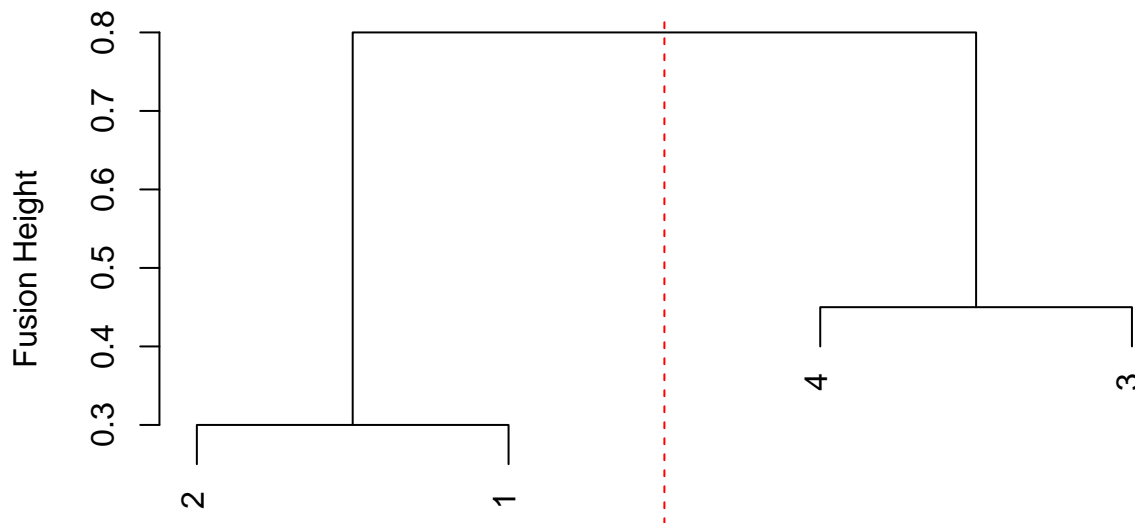
Cluster Dendrogram



- **C.** Observations 1 & 2 will be in cluster *A* and observations 3 & 4 will be in cluster *B* (assuming one cuts the dendrogram at a height greater than 0.45).
- **D.** Although the answer to this question depends on where one cuts the dendrogram, the most likely clusters would contain observations 1 & 2 in cluster *A* and observations 3 & 4 in cluster *B*. This would result from a cut at a height greater than 0.3 and less than 0.4, which is the largest vertical distance on the dendrogram. If one were to make a cut between 0.4 and 0.45, then cluster *A* would contain observations 1, 2 & 3, while cluster *B* would consist of only observation 4. However, with the distance being greater between clusters for the first grouping, that would be the more probable grouping.
- **E.** As shown below, one can simply switch the labels of the observations within each cluster to change the dendrogram without changing the meaning of the dendrogram. In addition, one could take the mirror image of the plot displayed along the dotted red line, producing a “new” dendrogram that has the same meaning.

```
plot(hclust(as.dist(x), method = 'complete'),
     xlab = "",
     sub = "",
     ylab = "Fusion Height", labels = c(2,1,4,3))
abline(v = 2.5, col = 'red', lty = 2)
```

Cluster Dendrogram



Resources

[Dendrograms in R](#)