# HW5 --Jiayuan Guo

1. **Chapter 10, Problem 2**

   (a) Use Algorithm 10.2 to explain the different steps that lead to the dendrogram:

   For now we have:

$$\begin{pmatrix} & 0.3 & 0.4 & 0.7 \\ 0.3 & & 0.5 & 0.8 \\ 0.4 & 0.5 & & 0.45 \\ 0.7 & 0.8 & 0.45 & \end{pmatrix} \tag{1}$$

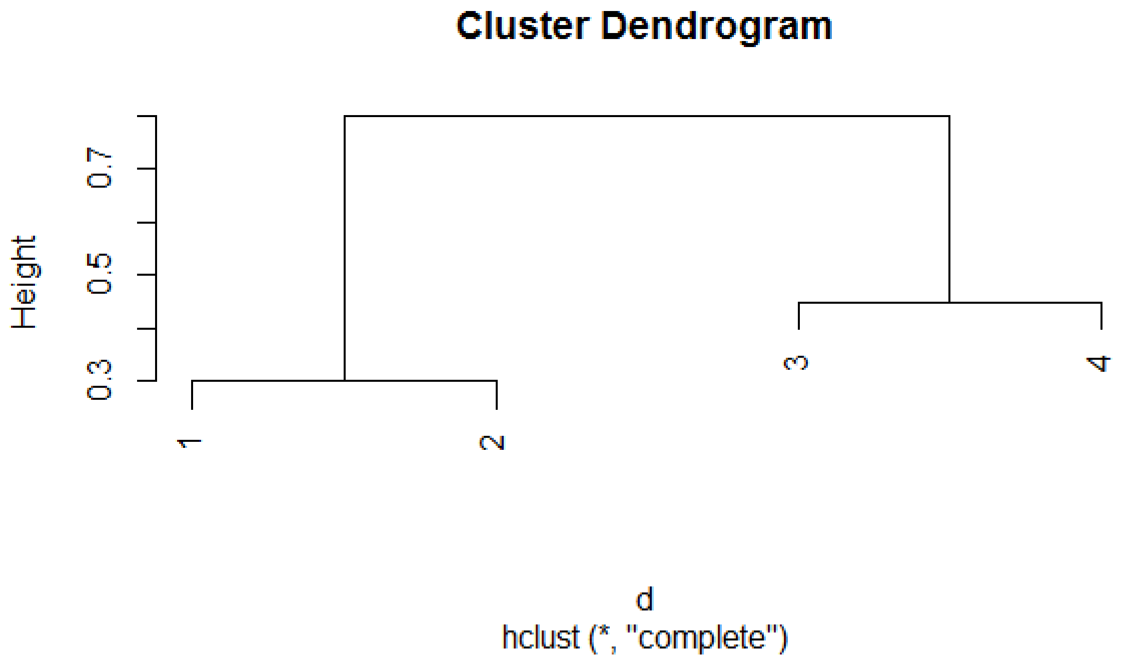
   For i=4 : 0.3 is the minimum dissimilarity, so we fuse observations 1 and 2 to form cluster (1,2) at height 0.3. And we have new dissimilarity matrix:

$$\begin{pmatrix} & 0.5 & 0.8 \\ 0.5 & & 0.45 \\ 0.8 & 0.45 & \end{pmatrix} \tag{2}$$

   For i=3 : 0.45 is the minimum dissimilarity, so we fuse observations 3 and 4 to form cluster (3,4) at height 0.45. And we have the new dissimilarity matrix:

$$\begin{pmatrix} & 0.8 \\ 0.8 & \end{pmatrix} \tag{3}$$

   It remains to fuse clusters (1,2) and (3,4) to form cluster ((1,2),(3,4)) at height 0.8.



(b) Repeat (a) using simple linkage clustering:

For now we have:

$$\begin{pmatrix} & 0.3 & 0.4 & 0.7 \\ 0.3 & & 0.5 & 0.8 \\ 0.4 & 0.5 & & 0.45 \\ 0.7 & 0.8 & 0.45 & \end{pmatrix} \tag{4}$$

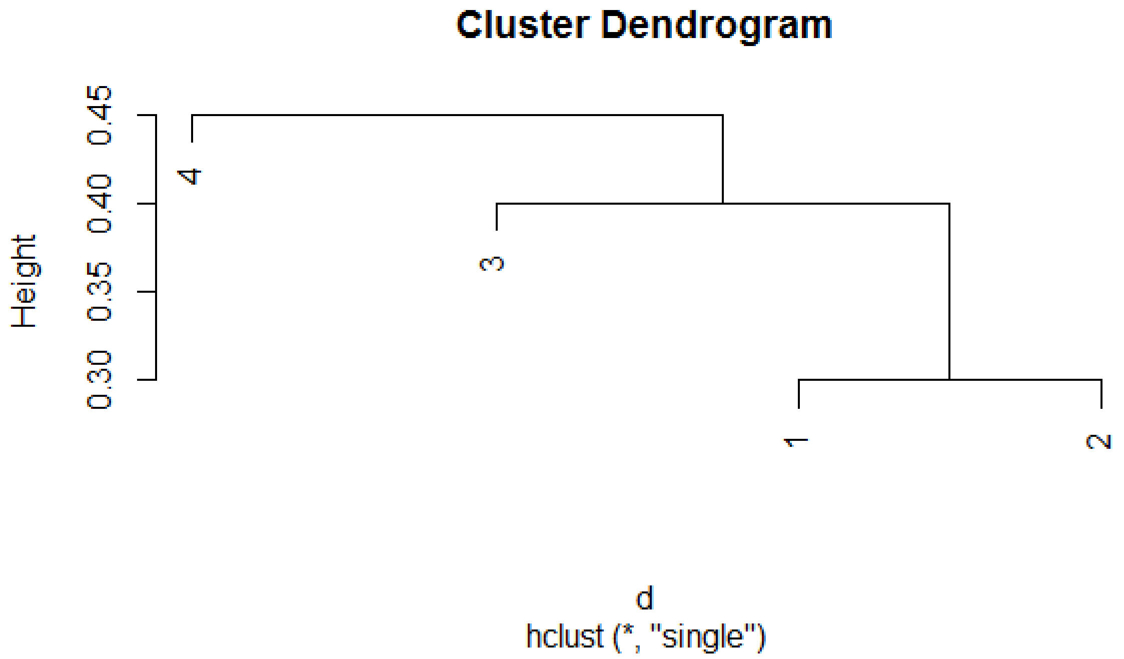
For i=4 : 0.3 is the minimum dissimilarity, so we fuse observations 1 and 2 to form cluster (1,2) at height 0.3. And we have new dissimilarity matrix:

$$\begin{pmatrix} & 0.4 & 0.4 \\ 0.4 & & 0.45 \\ 0.7 & 0.45 & \end{pmatrix} \tag{5}$$

For i=3 : 0.4 is the minimum dissimilarity, so we fuse cluster (1,2) and observation 3 to form cluster ((1,2),3) at height 0.4. And we have the new dissimilarity matrix:

$$\begin{pmatrix} & 0.45 \\ 0.45 & \end{pmatrix} \tag{6}$$

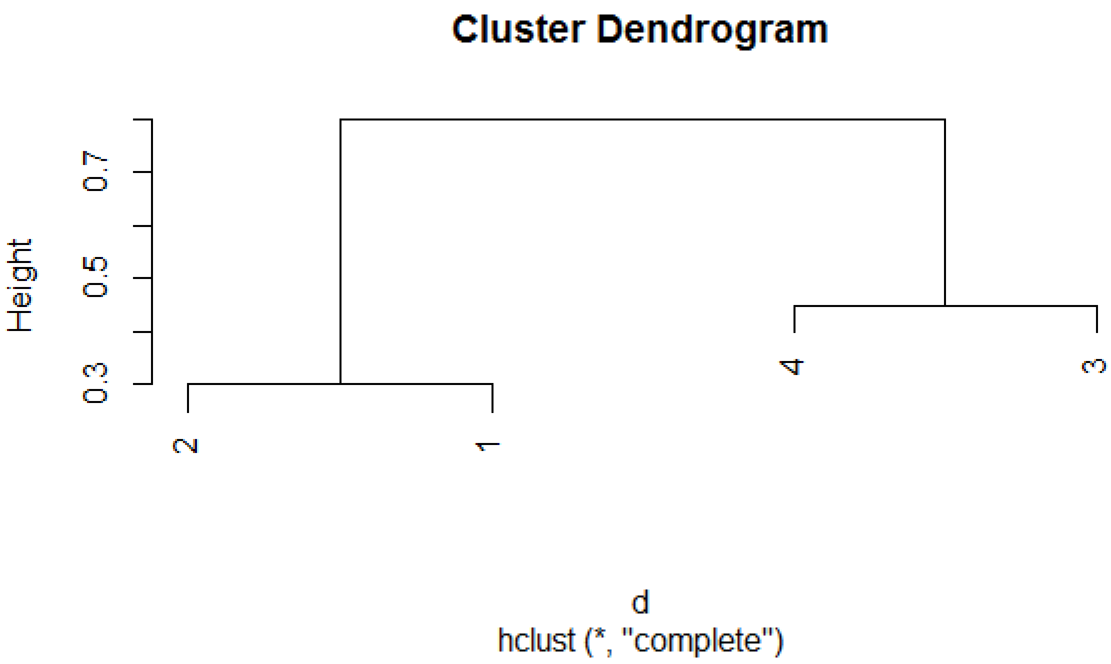
It remains to fuse clusters ((1,2),3) and observation 4 to form cluster (((1,2),3),4) at height 0.45.



(c) (1,2) and (3,4)

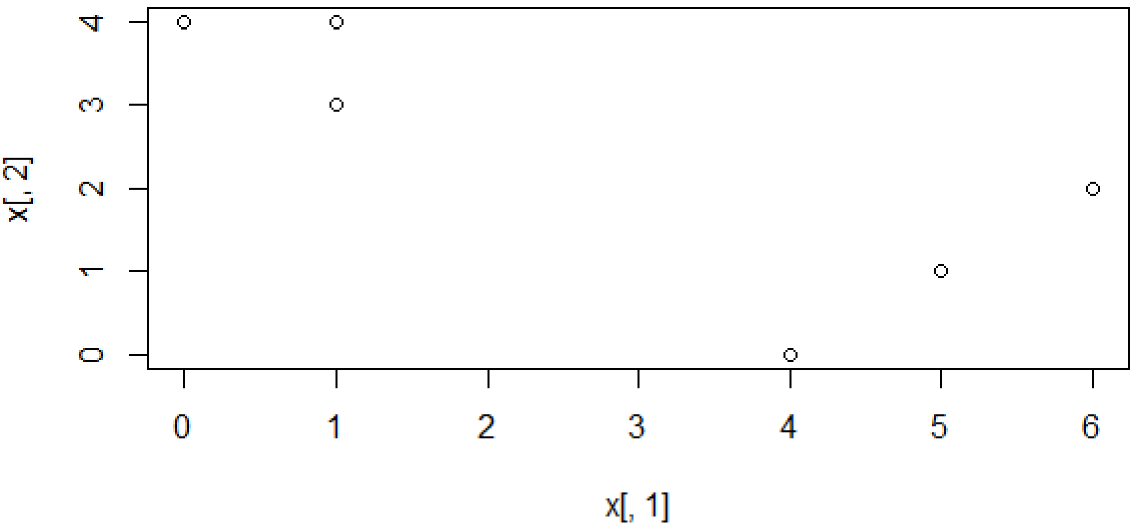
(d) (1, 2, 3) and (4)

(e)



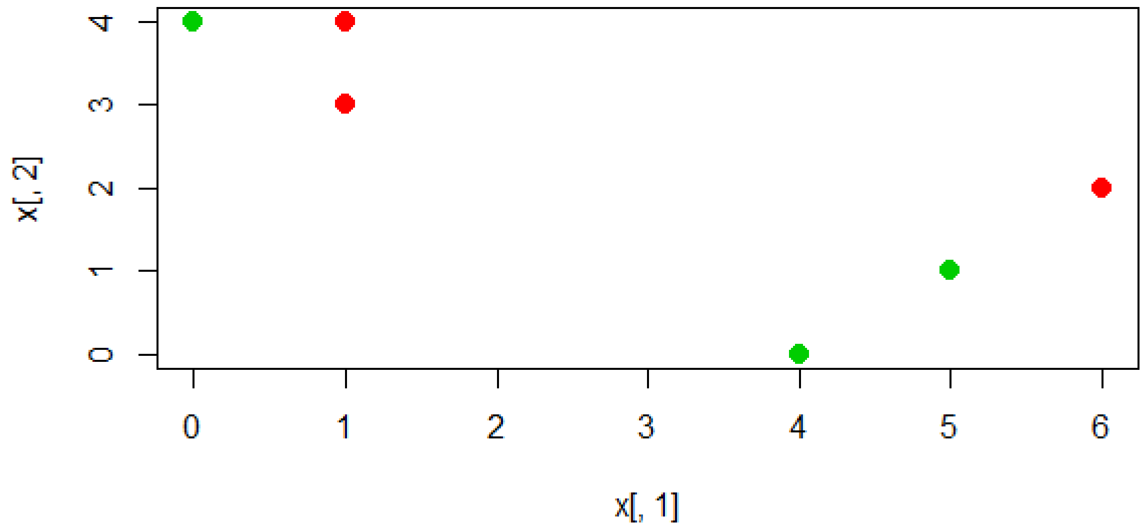
2. **Chapter 10, Problem 3**

(a)



(b)

```
Output:
> labels
[1] 1 1 2 2 1 2
```



(c)

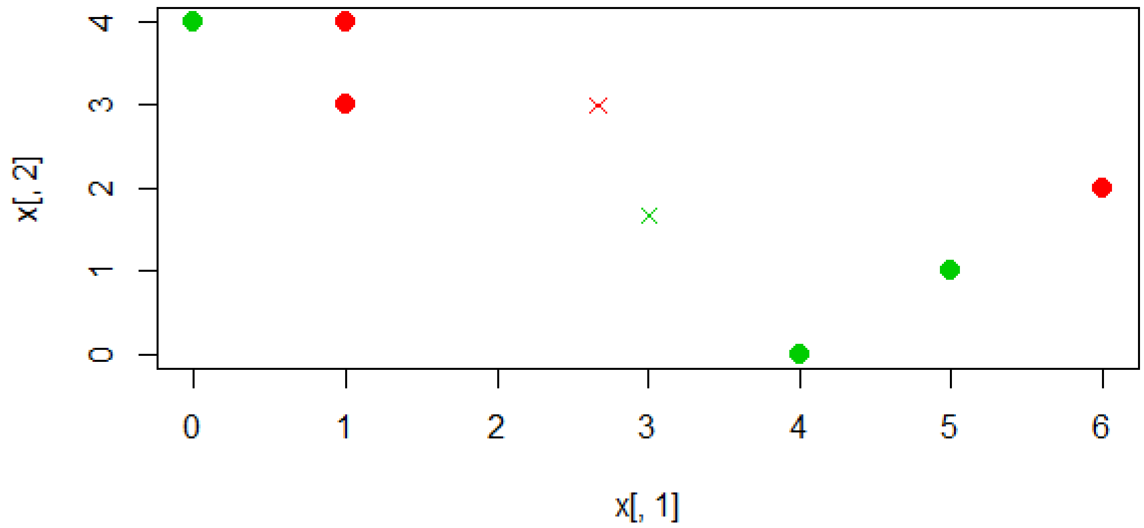
Compute the centroid for the green cluster:

$x_{11} = \frac{1}{3}(0+4+5) = 3$ and $x_{12} = \frac{1}{3}(4+0+1) = \frac{5}{3}$

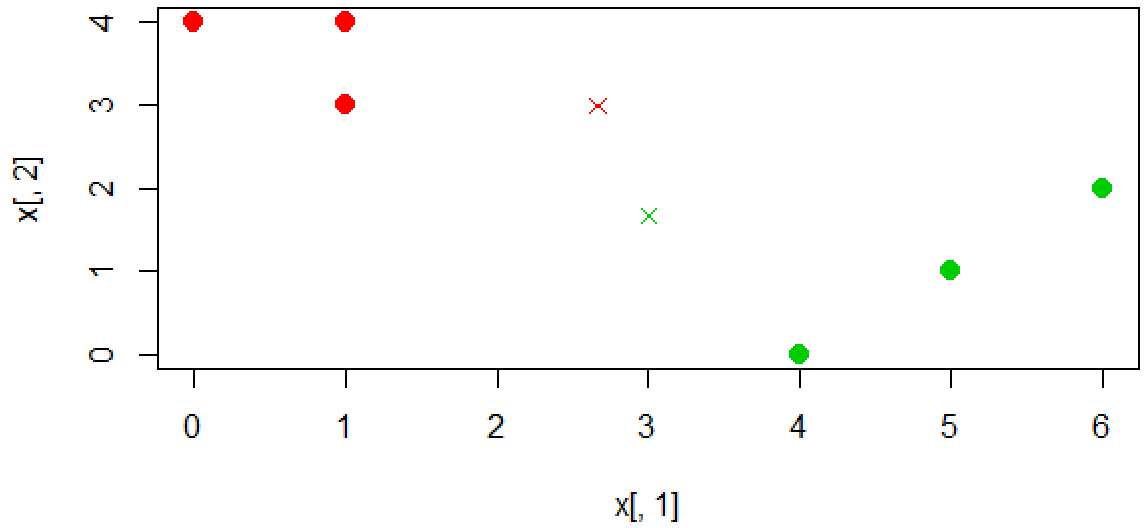
Compute the centroid for the red cluster:

$x_{21} = \frac{1}{3}(1+1+6) = \frac{8}{3}$ and $x_{22} = \frac{1}{3}(2+4+3) = 3$

```
Output:
> centroid1
[1] 2.666667 3.000000
> centroid2
[1] 3.000000 1.666667
```



(d)



(e)Repeat (c) and (d) until the answers obtained stop changing:

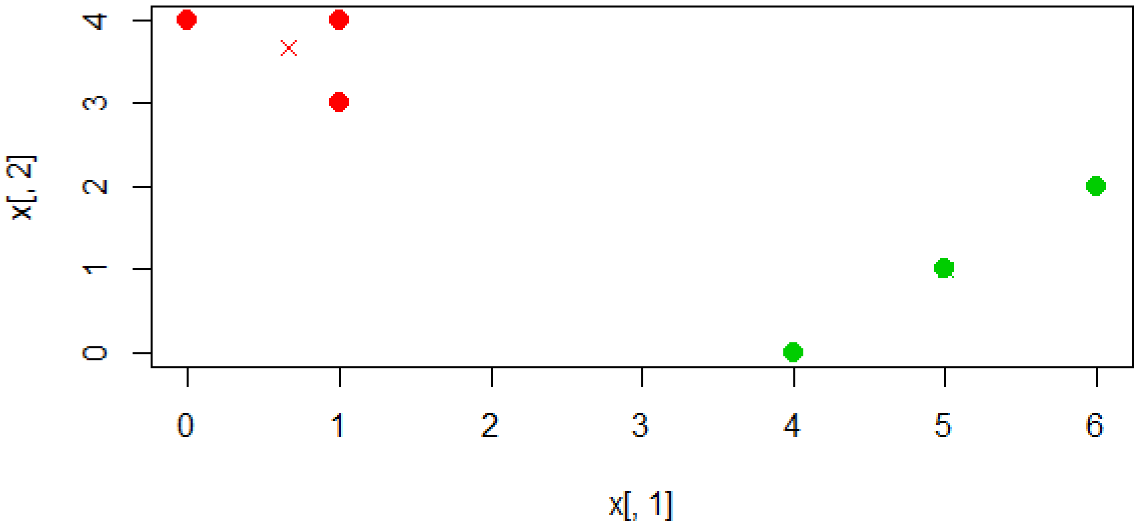
Compute the centroid for the green cluster:

$x_{11} = \frac{1}{3}(4+5+6) = 5$ and $x_{12} = \frac{1}{3}(0+1+2) = 1$

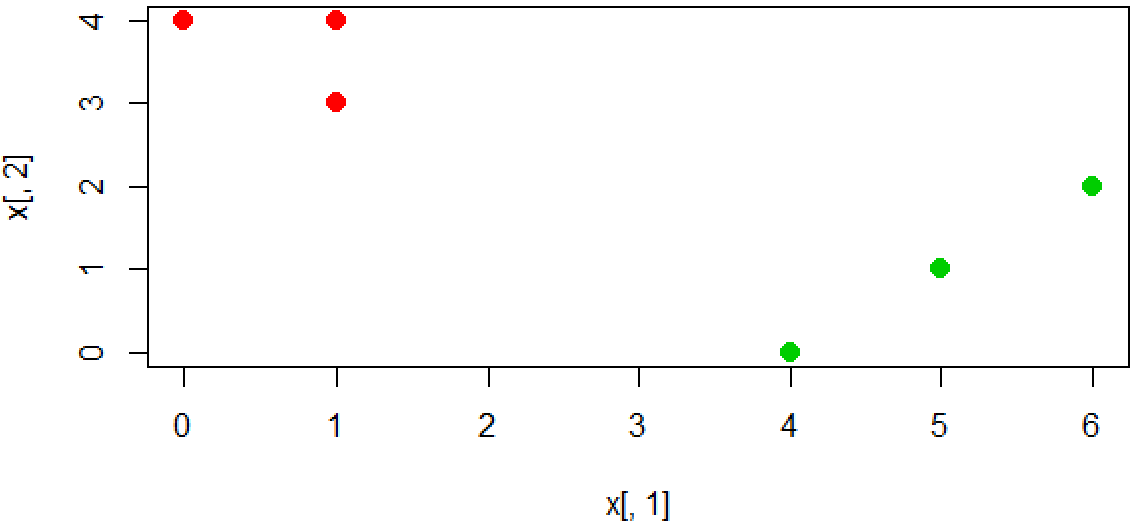
Compute the centroid for the red cluster:

$x_{21} = \frac{1}{3}(0+1+1) = \frac{2}{3}$ and $x_{22} = \frac{1}{3}(3+4+4) = \frac{11}{3}$

```
Output:
> centroid1
[1] 0.6666667 3.6666667
> centroid2
[1] 5 1
```

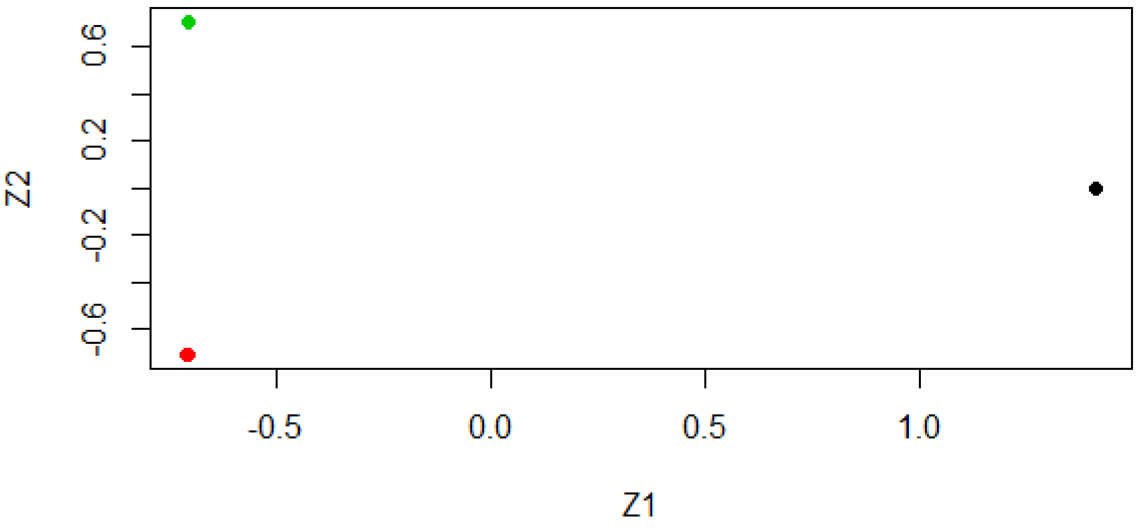


(f)



3. Chapter 10, Problem 10

(b)



(c)

```
Output:
> table(km.out$cluster, c(rep(1,20), rep(2,20), rep(3,20)))
     1  2  3
  1 20  0  0
  2  0  0 20
  3  0 20  0
```

Observations are perfectly clustered.

(d)

```
Output:
> table(true.labels, km.out$cluster)
true.labels  1  2
          1 20  0
          2  0 20
          3 20  0
```

All observations of one of the three clusters is now absorbed in one of the two clusters.

(e)

```
Output:
> table(true.labels, km.out$cluster)
true.labels  1  2  3  4
          1 10  0  0 10
          2  0 20  0  0
          3  0  0 20  0
```

One previous cluster split into two clusters.

(f)

```
Output:
> table(true.labels, km.out$cluster)
true.labels  1  2  3
          1  0 20  0
          2  0  0 20
          3 20  0  0
```

All observations are perfectly clustered again.

(g)

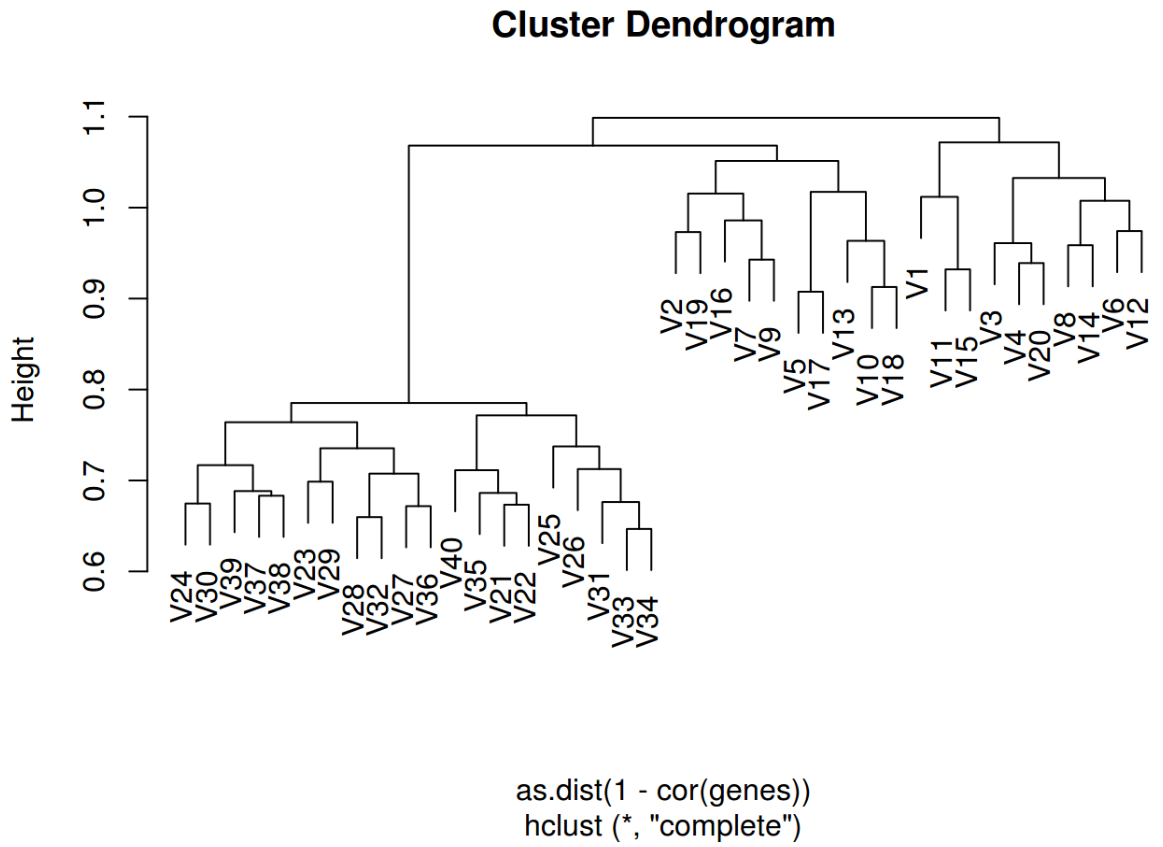
```
Output:
> table(true.labels, km.out$cluster)
true.labels  1  2  3
          1  3  8  9
          2  8  2 10
          3  8  8  4
```

Results are worse because scaling influence the distance between observations.

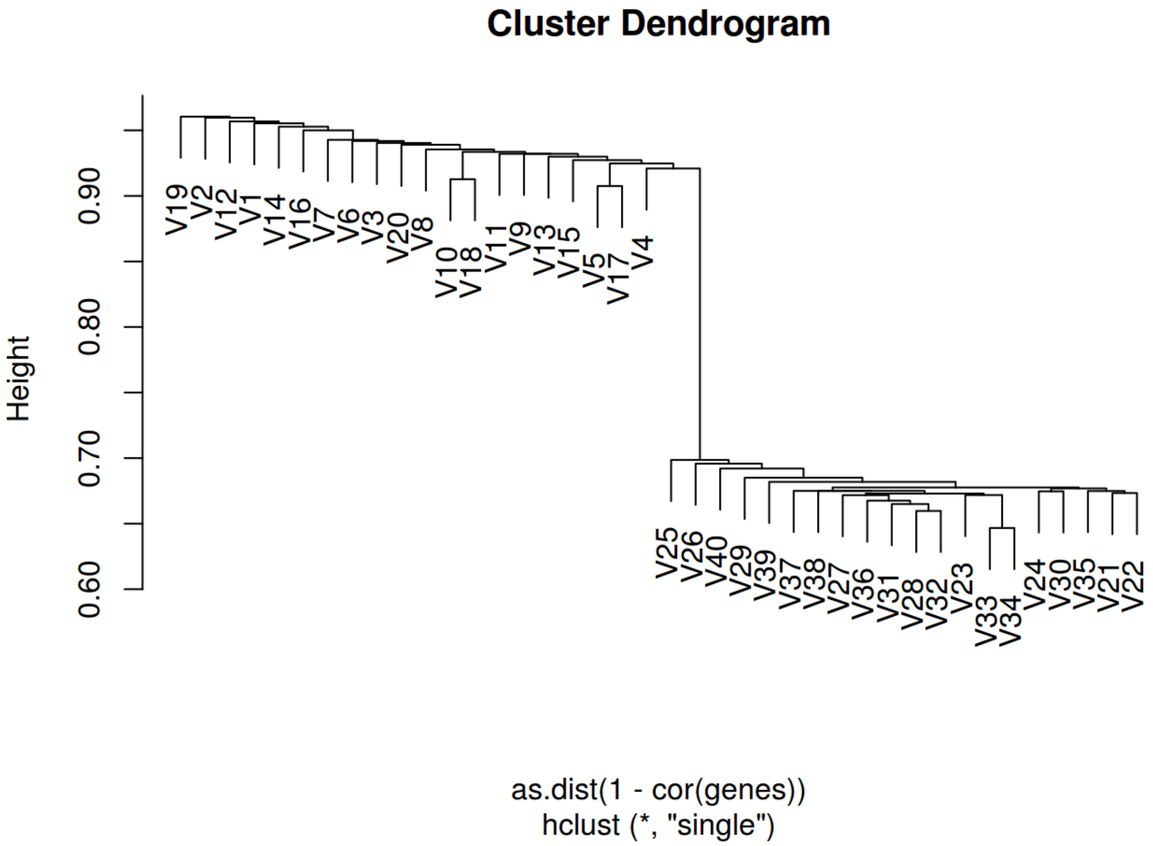
4. Chapter 10, Problem 11

(b)

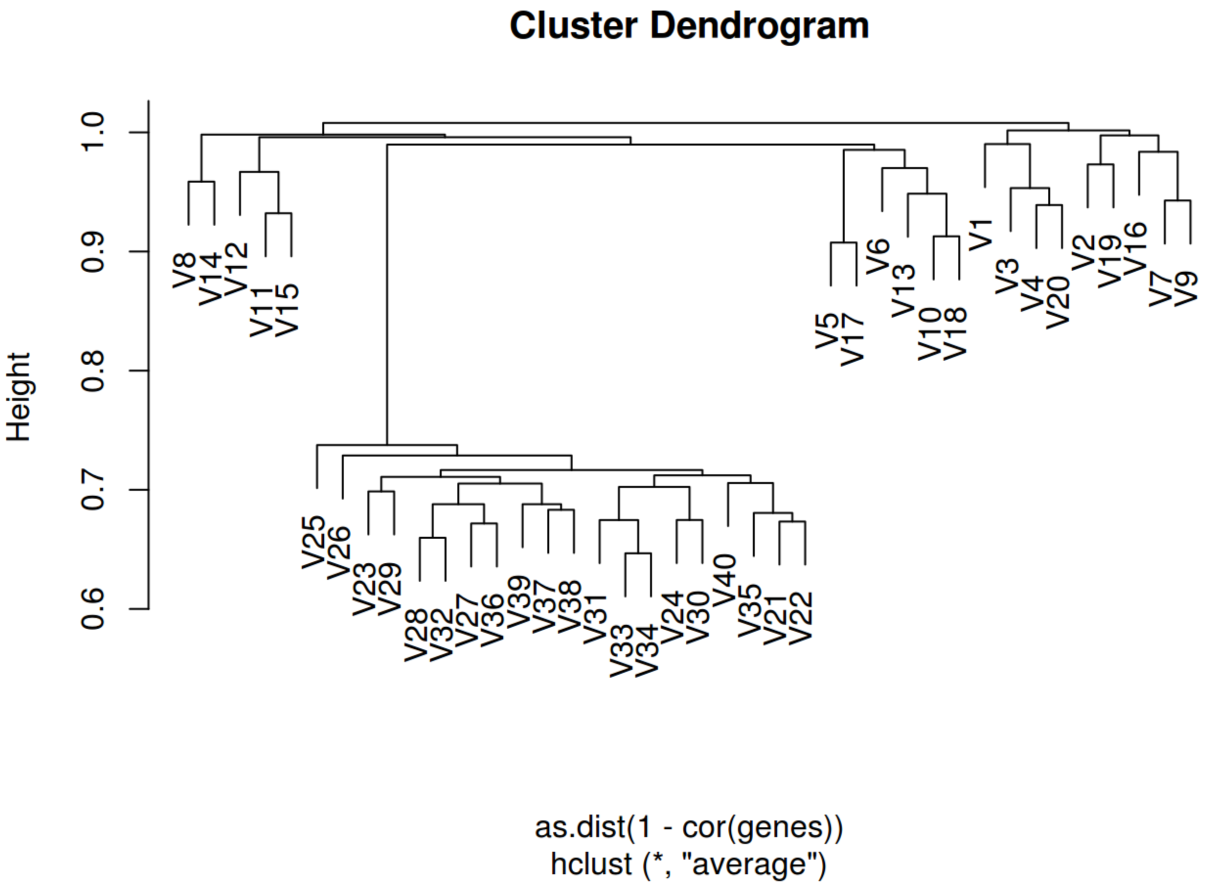
**Complete**



**Single:**


Cluster Dendrogram
Height
0.60
0.70
0.80
0.90
V19
V2
V12
V1
V14
V16
V7
V6
V3
V20
V8
V10
V18
V11
V9
V13
V15
V5
V17
V4
V25
V26
V40
V29
V39
V37
V38
V27
V36
V31
V28
V32
V23
V33
V34
V24
V30
V35
V21
V22
as.dist(1 - cor(genes))
hclust (*, "single")

**Average:**


Cluster Dendrogram
Height
0.6
0.7
0.8
0.9
1.0
V8
V14
V12
V11
V15
V25
V26
V23
V29
V28
V32
V27
V36
V39
V37
V38
V31
V33
V34
V24
V30
V40
V35
V21
V22
V5
V17
V6
V13
V10
V18
V1
V3
V4
V20
V2
V19
V16
V7
V9
as.dist(1 - cor(genes))
hclust (*, "average")

Different linkage methods lead to different results: we obtain two clusters for complete and single linkages or three clusters for average cluster.

(c)

Use PCA to see which genes differ the most across the two groups.

```
index[1:10]
Final Output:
##  [1] 865  68 911 428 624  11 524 803 980 822
```