STAT 602 Homework 12

Yuchi Hu

April 12, 2019

1. Question 10.7.8 pg 416

In Section 10.2.3, a formula for calculating PVE was given in Equation 10.8. We also saw that the PVE can be obtained using the **sdev** output of the **prcomp()** function.

On the **USArrests** data, calculate PVE in two ways:

(a)

Using the **sdev** output of the **prcomp()** function, as was done in Section 10.2.3.

Answer

We perform principal components analysis using **prcomp()**. The variance explained by each principal component is obtained by squaring the standard deviation of each principal component (**sdev** output of **prcomp()**). The proportion of variance explained (PVE) by each principal component is computed by dividing the variance explained by each principal component by the total variance explained by all four principal components:

```
## [1] 0.62006039 0.24744129 0.08914080 0.04335752
```

We see that the first principal component explains 62% of the variance in the data, the second principal component explains 24.7% of the variance, and so forth.

(b)

By applying Equation 10.8 directly. That is, use the **prcomp()** function to compute the principal component loadings. Then, use those loadings in Equation 10.8 to obtain the PVE.

These two approaches should give the same results.

Answer

Equation 10.8 gives the PVE of the mth principal component:

$$\frac{\sum_{i=1}^{n}(\sum_{j=1}^{p}\phi_{jm}x_{ij})^2}{\sum_{j=1}^{p}\sum_{i=1}^{n}x_{ij}^2}$$

where ϕ_{jm} is the loading for the jth variable on the mth principal component and x_{ij} is the value of the jth variable for the ith observation.

In part (a), we performed PCA using standardized variables with mean zero and standard deviation one, so first we use **scale()** to standardize the variables. Then, we use **prcomp()** to compute the principal component loadings. Using equation 10.8 with the standardized variables and the loadings, we calculate PVE:

```
##         PC1         PC2         PC3         PC4
## 0.62006039 0.24744129 0.08914080 0.04335752
```

We see that the results are equal to those from part (a).

2. Question 10.7.9 pg 416

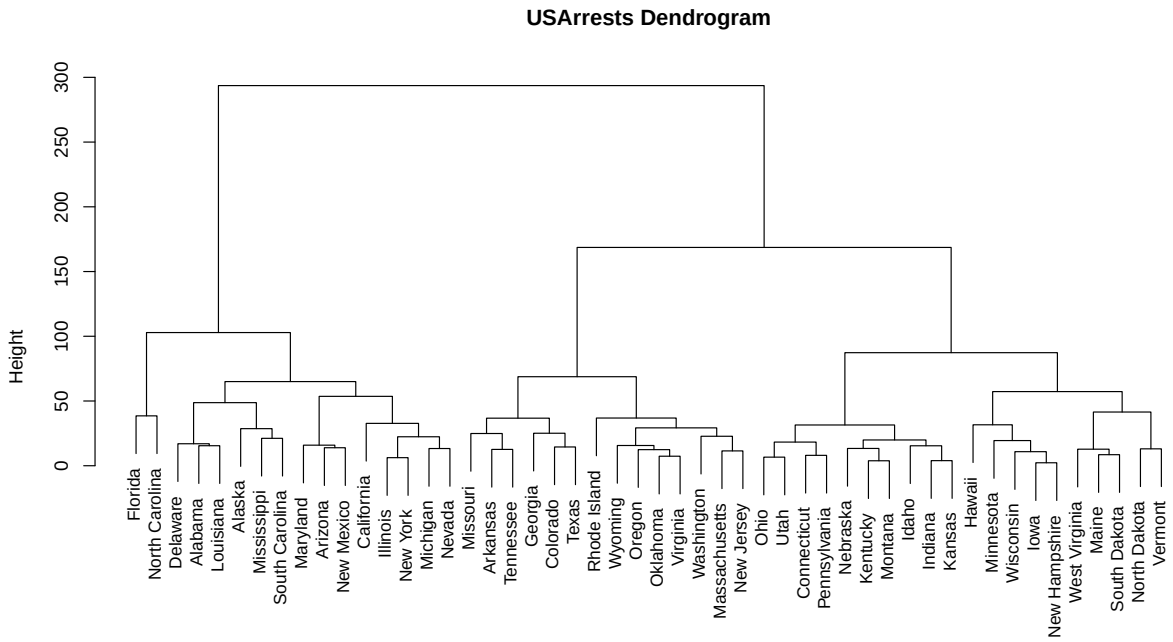
Consider the **USArrests** data. We will now perform hierarchical clustering on the states.

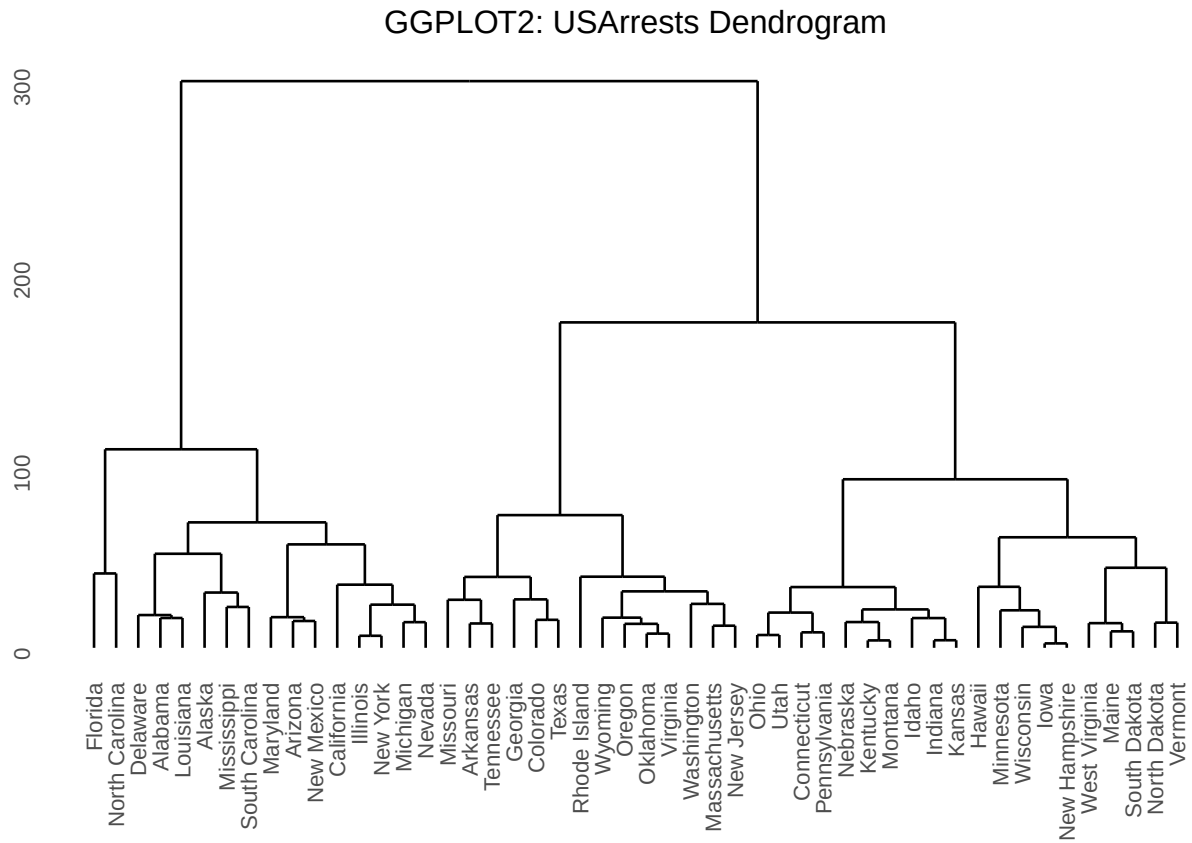
(a)

Using hierarchical clustering with complete linkage and Euclidean distance, cluster the states.

Answer

We use hierarchical clustering with complete linkage and Euclidean distance to cluster the states. **hclust()** implements hierarchical clustering, and **plot()** plots the resulting dendrogram. To plot the dendrogram in ggplot2, we use the **ggdendro** library. The dendrogram is shown below:




GGPLOT2: USArrests Dendrogram
300
200
100
0
Florida
North Carolina
Delaware
Alabama
Louisiana
Alaska
Mississippi
South Carolina
Maryland
Arizona
New Mexico
California
Illinois
New York
Michigan
Nevada
Missouri
Arkansas
Tennessee
Georgia
Colorado
Texas
Rhode Island
Wyoming
Oregon
Oklahoma
Virginia
Washington
Massachusetts
New Jersey
Ohio
Utah
Connecticut
Pennsylvania
Nebraska
Kentucky
Montana
Idaho
Indiana
Kansas
Hawaii
Minnesota
Wisconsin
Iowa
New Hampshire
West Virginia
Maine
South Dakota
North Dakota
Vermont

(b)

Cut the dendrogram at a height that results in three distinct clusters. Which states belong to which clusters?

Answer

We use **cutree()** to cut the dendrogram at a height that results in three distinct clusters. The table below shows which states belong to which clusters.

Table 1: States in Each Cluster

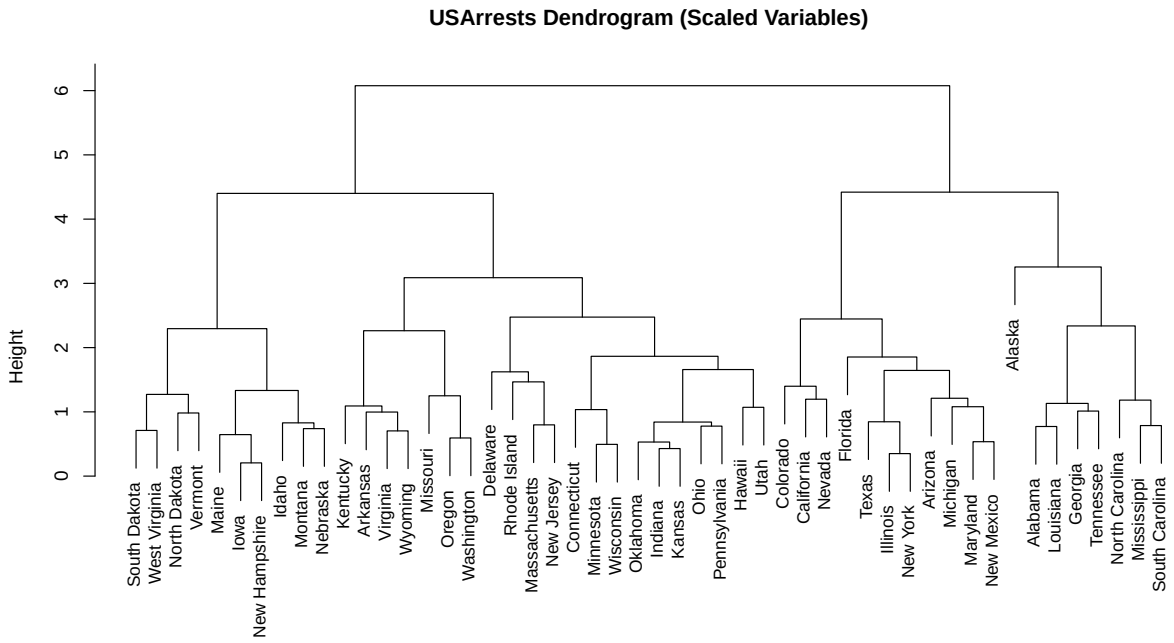
Cluster 1	Cluster 2	Cluster 3
Alabama	Arkansas	Connecticut
Alaska	Colorado	Hawaii
Arizona	Georgia	Idaho
California	Massachusetts	Indiana
Delaware	Missouri	Iowa
Florida	New Jersey	Kansas
Illinois	Oklahoma	Kentucky
Louisiana	Oregon	Maine
Maryland	Rhode Island	Minnesota
Michigan	Tennessee	Montana
Mississippi	Texas	Nebraska
Nevada	Virginia	New Hampshire
New Mexico	Washington	North Dakota
New York	Wyoming	Ohio
North Carolina		Pennsylvania
South Carolina		South Dakota
		Utah
		Vermont
		West Virginia
		Wisconsin

(c)

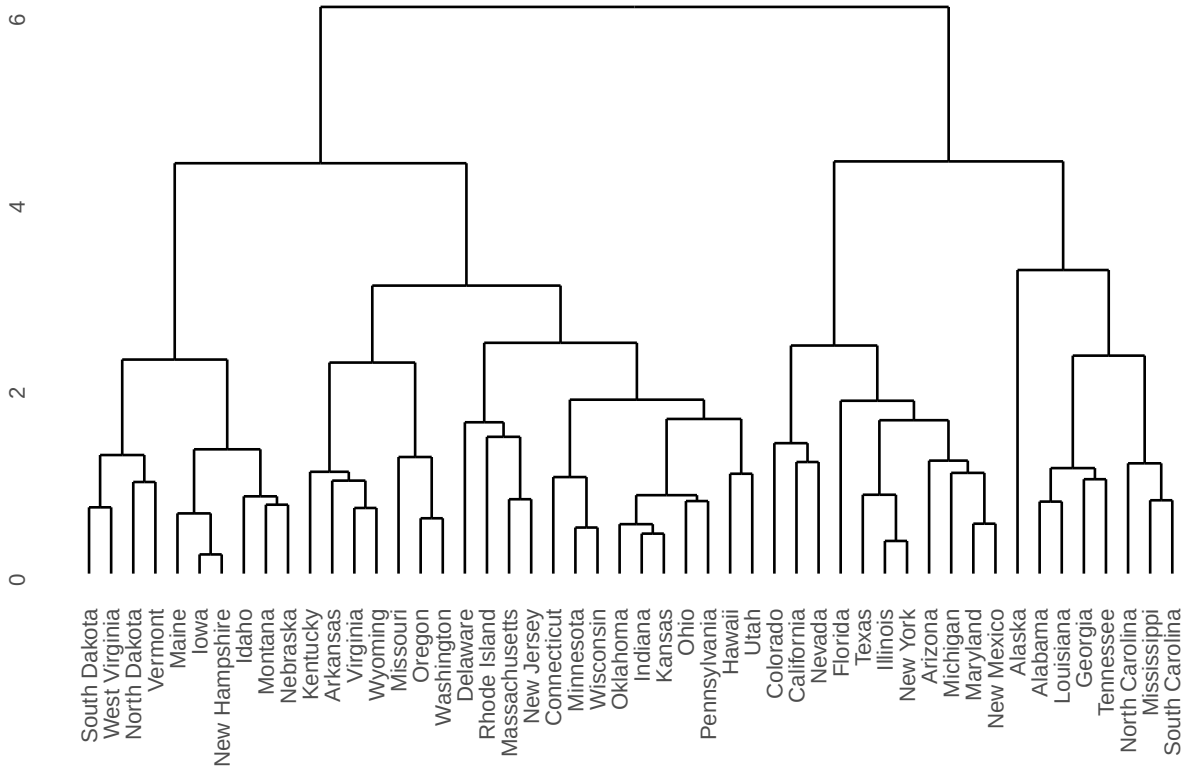
Hierarchically cluster the states using complete linkage and Euclidean distance, *after scaling the variables to have standard deviation one.*

Answer

We use **scale()** to scale the variables to have standard deviation one, then hierarchically cluster the states using complete linkage and Euclidean distance. The dendrogram is shown below:



GGPLOT2: USArrests Dendrogram (Scaled Variables)



(d)

What effect does scaling the variables have on the hierarchical clustering obtained? In your opinion, should the variables be scaled before the inter-observation dissimilarities are computed? Provide a justification for your answer.

Answer

The table below shows which states belong to which clusters after scaling the variables.

Table 2: States in Each Cluster (Scaled Variables)

Cluster 1	Cluster 2	Cluster 3
Alabama	Arizona	Arkansas
Alaska	California	Connecticut
Georgia	Colorado	Delaware
Louisiana	Florida	Hawaii
Mississippi	Illinois	Idaho
North Carolina	Maryland	Indiana
South Carolina	Michigan	Iowa
Tennessee	Nevada	Kansas
	New Mexico	Kentucky
	New York	Maine
	Texas	Massachusetts
		Minnesota
		Missouri
		Montana
		Nebraska
		New Hampshire
		New Jersey
		North Dakota
		Ohio
		Oklahoma
		Oregon
		Pennsylvania
		Rhode Island
		South Dakota
		Utah
		Vermont
		Virginia
		Washington
		West Virginia
		Wisconsin
		Wyoming

Without scaling the variables, the three clusters are somewhat balanced; however, we see that after scaling the variables, cluster 3 has many more states in it than cluster 1 and cluster 2. Thus, we obtain very different clusters after scaling the variables.

The variables should be scaled before the inter-observation dissimilarities are computed since the variables are measured in different units. *Murder*, *Assault*, and *Rape* are rates calculated per 100,000 people, while *UrbanPop* is the percentage of the population living in urban areas. If we do not scale the variables, the first

principle component loading vector would place most of its weight on *Assault* since that variables has by far the highest variance.

3. Question 10.7.11 pg 417

On the book website, www.StatLearning.com, there is a gene expression data set (**Ch10Ex11.csv**) that consists of 40 tissue samples with measurements on 1,000 genes. The first 20 samples are from healthy patients, while the second 20 are from a diseased group.

(a)

Load in the data using **read.csv()**. You will need to select **header=F**.

Answer

We load in the data using **read.csv()**.

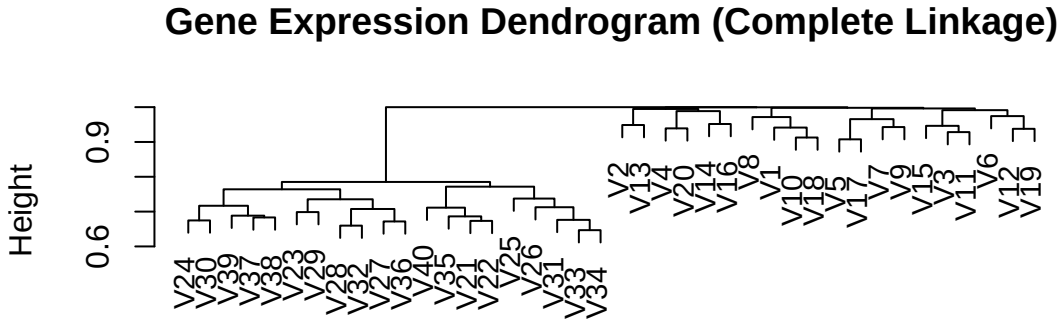
(b)

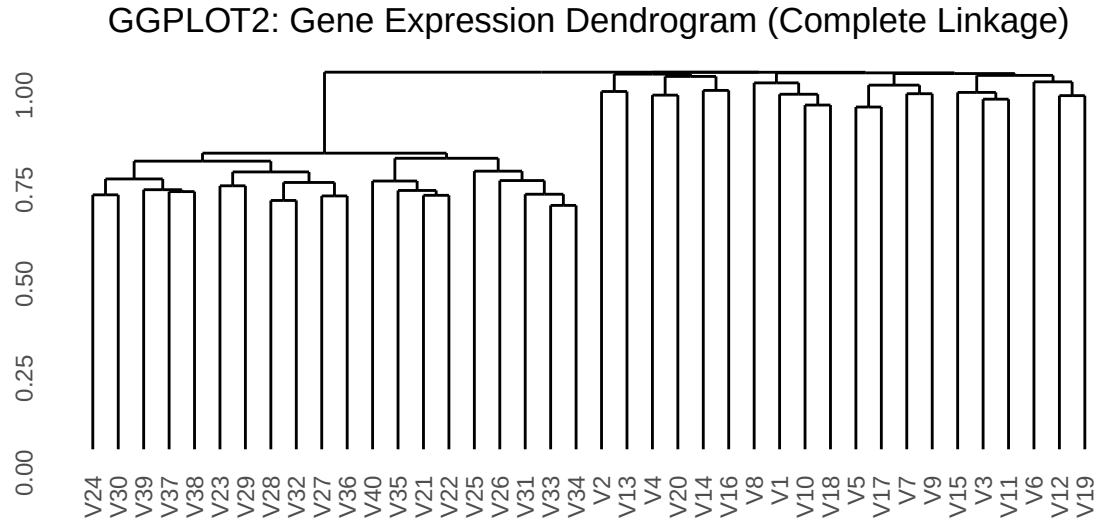
Apply hierarchical clustering to the samples using correlation-based distance, and plot the dendrogram. Do the genes separate the samples into the two groups? Do your results depend on the type of linkage used?

Answer

We apply hierarchical clustering to the samples using correlation-based distance, i.e. 1 - Abs(Correlation). We plot the resulting dendrograms with different types of linkage: complete, average, and single.

Complete Linkage



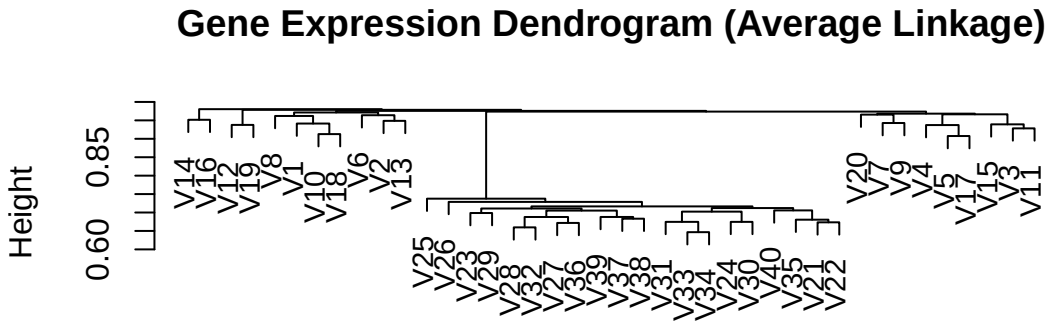


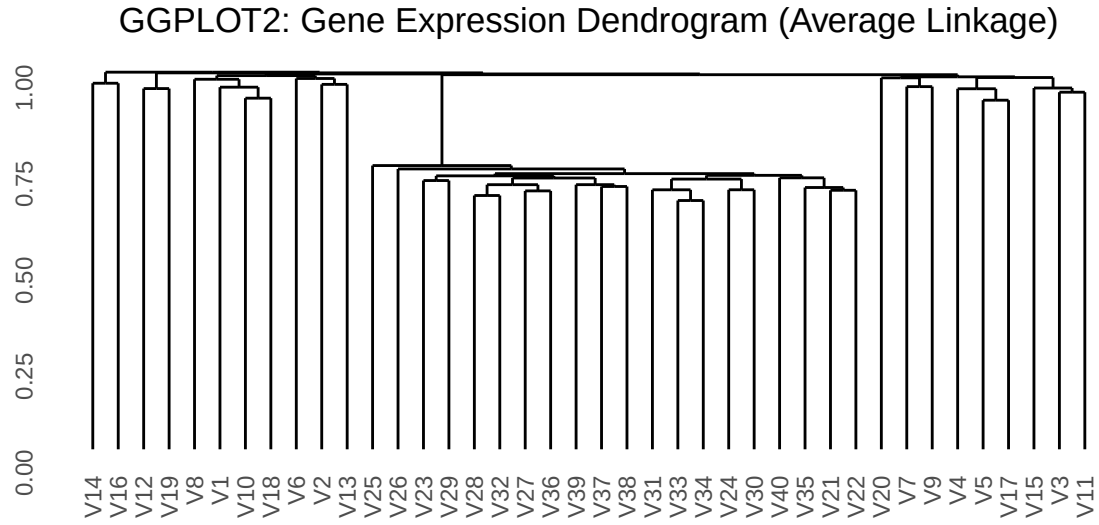
We can use **cutree()** to cut the dendrogram at a height that results in two distinct clusters. The cluster labels for each sample are:

```
##  V1  V2  V3  V4  V5  V6  V7  V8  V9 V10 V11 V12 V13 V14 V15 V16 V17 V18
##   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
## V19 V20 V21 V22 V23 V24 V25 V26 V27 V28 V29 V30 V31 V32 V33 V34 V35 V36
##   1   1   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
## V37 V38 V39 V40
##   2   2   2   2
```

The first 20 samples (V1-V20) are from healthy patients, while the second 20 (V21-V40) are from a diseased group. With complete linkage, we see that all of the healthy samples are in one cluster, while all of the diseased samples are in another cluster. Hence, we have perfect separation.

Average Linkage



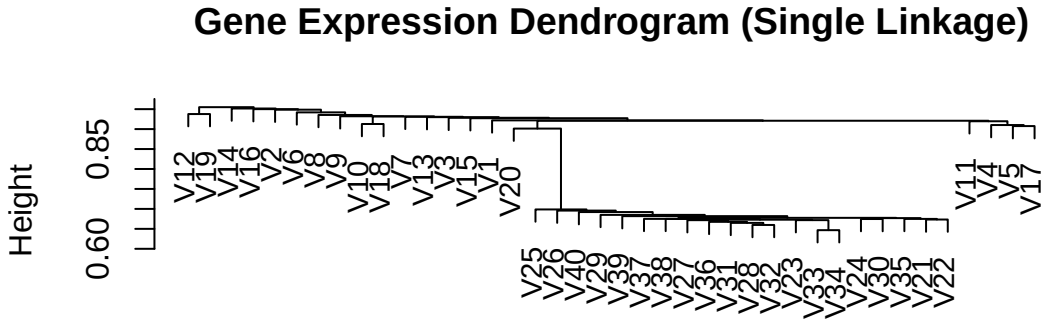


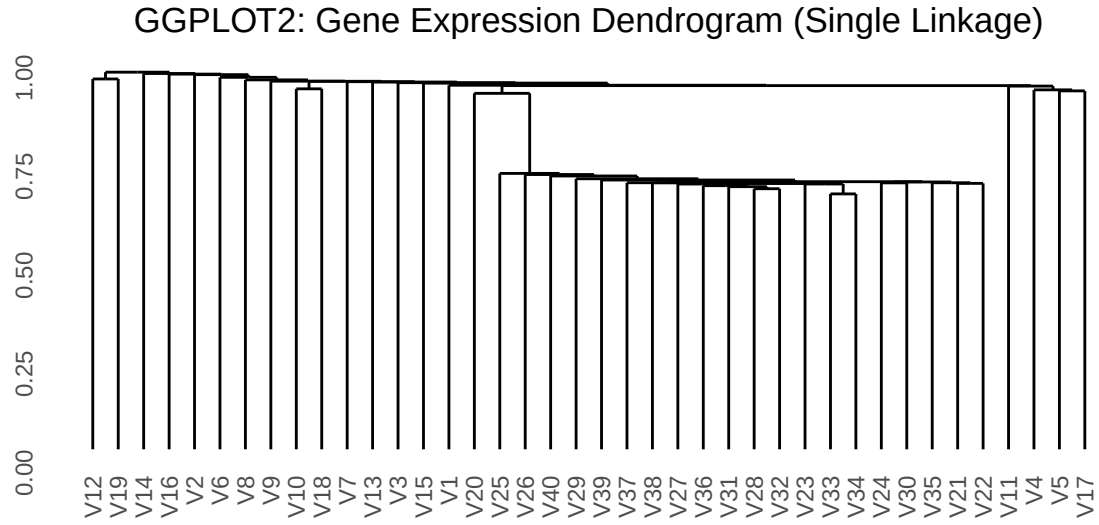
The cluster labels for each sample are:

```
##  V1  V2  V3  V4  V5  V6  V7  V8  V9 V10 V11 V12 V13 V14 V15 V16 V17 V18
##   1   1   1   1   1   1   1   1   1   1   1   1   1   2   1   2   1   1
## V19 V20 V21 V22 V23 V24 V25 V26 V27 V28 V29 V30 V31 V32 V33 V34 V35 V36
##   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
## V37 V38 V39 V40
##   1   1   1   1
```

With average linkage, we see that V14 and V16 belong to their own cluster, while all of the other samples are in another cluster.

Single Linkage





The cluster labels for each sample are:

```
##  V1  V2  V3  V4  V5  V6  V7  V8  V9 V10 V11 V12 V13 V14 V15 V16 V17 V18
##   1   1   1   1   1   1   1   1   1   1   1   2   1   1   1   1   1   1
## V19 V20 V21 V22 V23 V24 V25 V26 V27 V28 V29 V30 V31 V32 V33 V34 V35 V36
##   2   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
## V37 V38 V39 V40
##   1   1   1   1
```

With single linkage, we see that V12 and V19 belong to their own cluster, while all of the other samples are in another cluster.

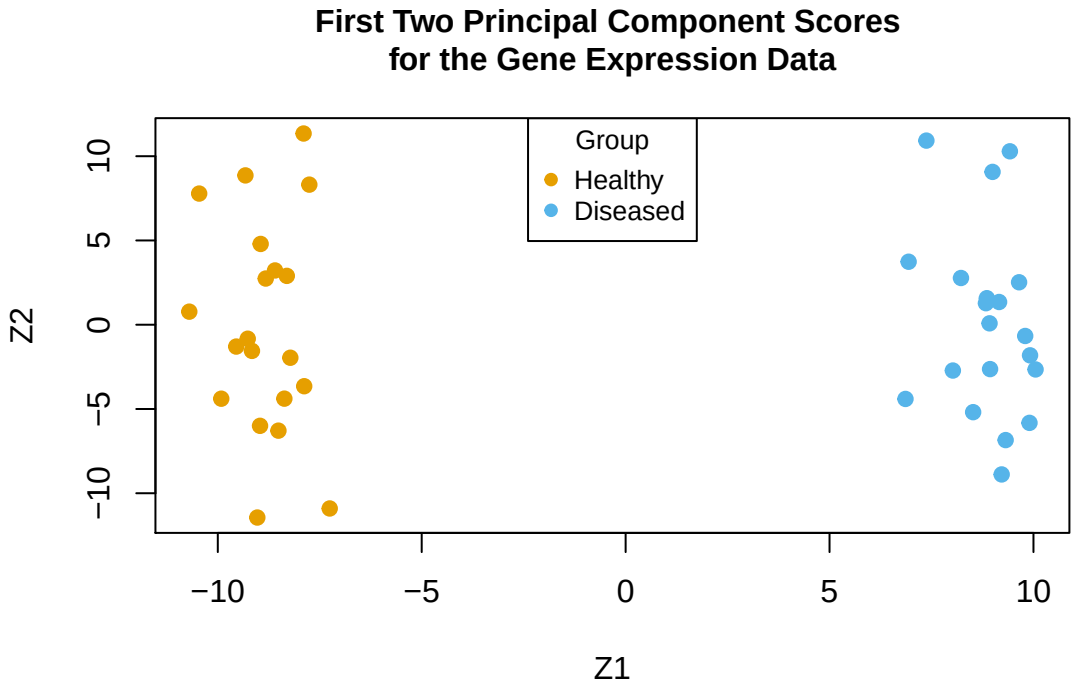
Thus, the results certainly do depend on the type of linkage. Complete linkage perfectly separates the healthy and diseased samples into their correct clusters, while average and single linkage found two samples that belong to their own cluster.

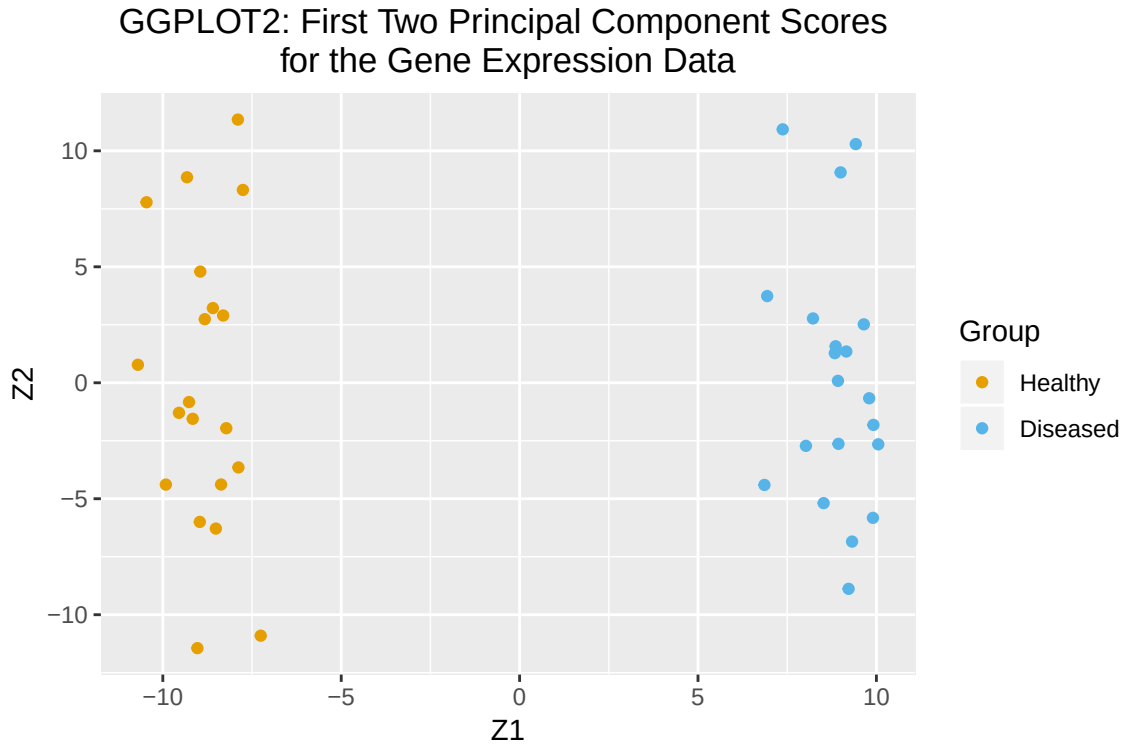
(c)

Your collaborator wants to know which genes differ the most across the two groups. Suggest a way to answer this question, and apply it here.

Answer

Since the first 20 samples are healthy and the second 20 are diseased, we can create labels for each sample. Then, we project the samples onto the first two principal components (plot the scores for the first two principal components) to see if there is good separation of the samples.





We see that there is perfect separation of the samples on the first principal component.

Next, we calculate the loading for each gene on the first principal component and then sort the absolute values of these loadings. The larger the loading for a gene on the first principal component, the higher its variance is among the samples. The genes with the 10 largest loadings are:

```
##  [1] 502 589 565 590 600 551 593 538 584 509
```

Thus, these are the 10 genes that differ the most across the two groups.