**STA 546 ASSIGNMENT #2**

Suruchi Jaikumar Ahuja

1. The SwissBankNotes data was loaded in to the R environment. The data consists of six variables measured on 200 old Swiss 1,000-franc bank notes. The first 100 are genuine and the second 100 are counterfeit. The six variables are length of the bank note, height of the bank note, measured on the left, height of the bank note, measured on the right, distance of the inner frame to the lower border, distance of inner frame to upper border, and length of the diagonal.

Principal Component Analysis is one of the ways to analyze the structure of a given correlation matrix. Biplots are a type of exploratory graph used in [statistics](https://en.wikipedia.org/wiki/Statistics), a generalization of the simple two-variable scatterplot. A Biplot is an enhanced scatterplot that uses both points and vectors to represent structure. A biplot uses points to represent the scores of the observations on the principal components, and it uses vectors to represent the coefficients of the variables on the principal components. In a biplot, the length of the lines approximates the variances of the variables.

The longer the line, the higher the variance. The closer the angle is to 90, or 270 degrees, the smaller the correlation. An angle of 0 or 180 degrees reflects a correlation of 1 or −1, respectively. The cutpoint of a perpendicular from a specific point to a variable line approximates the value of that observation on the variable that the line represents. If the cutpoint falls on the origin, the value of the observation is approximately the average of the respective variable. Cutpoints far off in the direction of the variable line indicate high values, while cutpoints far off on the variable line, which has been extended through the origin, and represent low values.

The distance between two points approximates the Euclidean distance between two observations in the multivariate space. Observations that are far away from each other have a high Euclidean distance, and vice versa.

PCA was performed on the 200 bank notes combined (pcr.fit), 100 genuine notes (pcr.fit1) and 100 counterfeit notes (pcr.fit2).

Figure 1- PCA on all 200 notes.

The height.right - height.left and inner.upper – inner.lower have a smaller correlation. The diagonal and length have a correlation of -1.

Figure 2 – PCA on 100 genuine notes.

The height.right - height.left and length have a small correlation. The diagonal – inner.lower have a smaller correlation when compared to the diagonal – inner.upper.

Figure 3 – PCA on 100 counterfeit notes.

The height.right - height.left and length have a small correlation. The diagonal – inner.lower have a smaller correlation when compared to the inner.lower – inner.upper which has a correlation of -1.

**OUTPUT:**

PCA for 200 notes combined – pcr.fit

Importance of components:

PC1 PC2 PC3 PC4 PC5 PC6

Standard deviation 1.7163 1.1305 0.9322 0.67065 0.51834 0.43460

Proportion of Variance 0.4909 0.2130 0.1448 0.07496 0.04478 0.03148

Cumulative Proportion 0.4909 0.7039 0.8488 0.92374 0.96852 1.00000

PCA for 100 genuine notes - pcr.fit1

Importance of components:

PC1 PC2 PC3 PC4 PC5 PC6

Standard deviation 1.4845 1.3026 0.9827 0.76348 0.57156 0.47340

Proportion of Variance 0.3673 0.2828 0.1610 0.09715 0.05445 0.03735

Cumulative Proportion 0.3673 0.6501 0.8111 0.90820 0.96265 1.00000

PCA for the 100 counterfeit notes - pcr.fit2

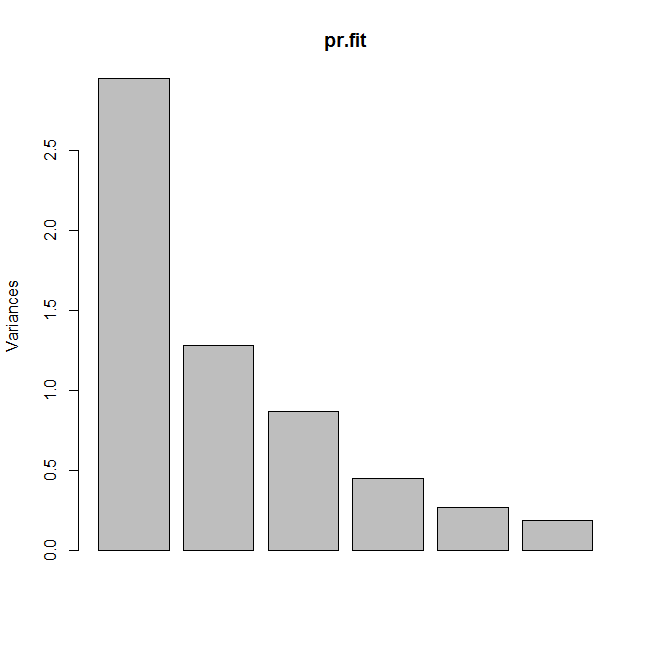
Importance of components:

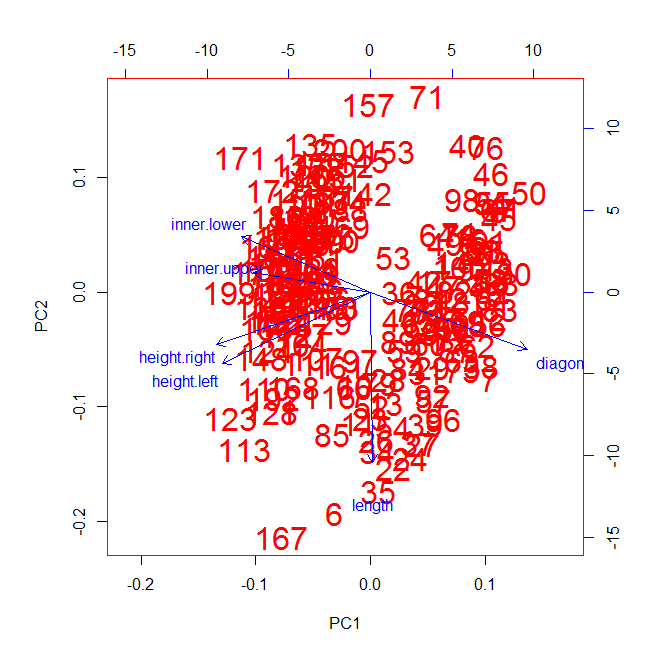
PC1 PC2 PC3 PC4 PC5 PC6

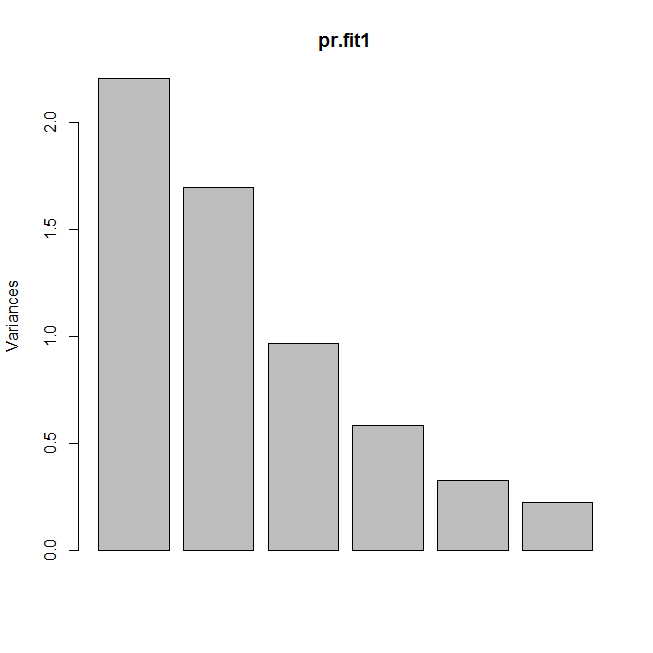
Standard deviation 1.3915 1.3285 0.9941 0.8823 0.56755 0.45840

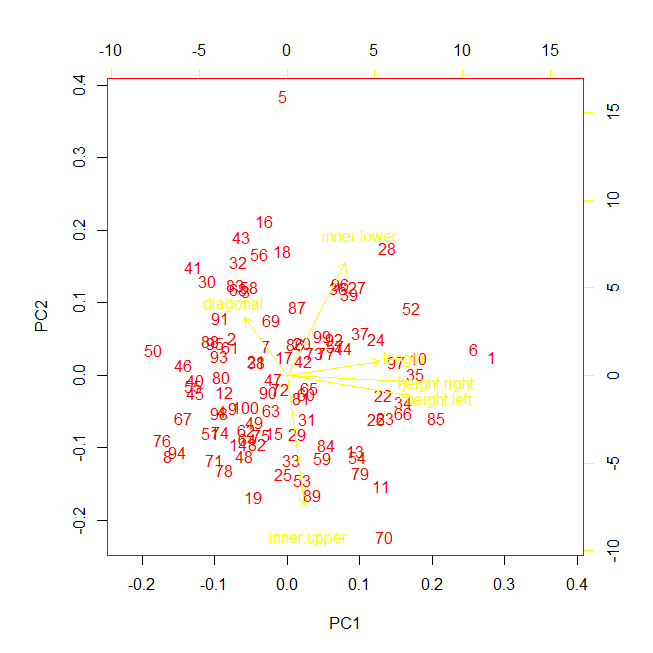
Proportion of Variance 0.3227 0.2941 0.1647 0.1297 0.05368 0.03502

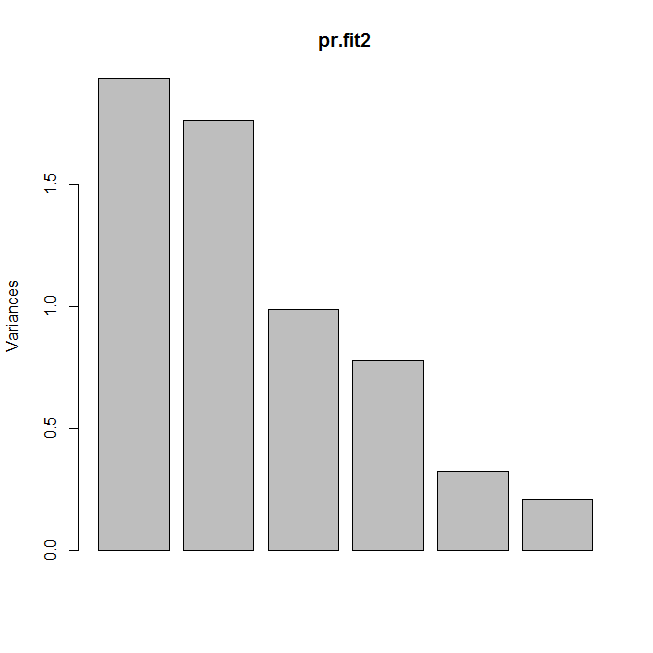
Cumulative Proportion 0.3227 0.6169 0.7816 0.9113 0.96498 1.00000

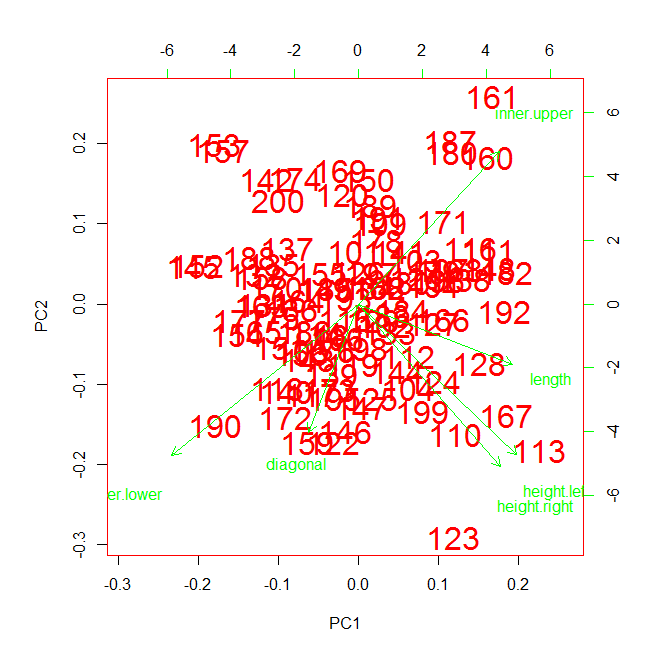












2)

a) A simulated data set is generated with 20 observations in each of three classes (i.e. 60 observations total), and 50 variables using the rnorm matrix function.

b) Then Principal component analysis is performed on the 60 observations and the first two principal component score vectors were plotted.

The three classes were separated amongst two dimensions.

> summary(pca.fit)

Importance of components:

PC1 PC2 PC3 PC4 PC5 PC6 PC7 PC8 PC9 PC10

Standard deviation 2.6988 1.1421 0.001837 0.001795 0.001633 0.001581 0.001553 0.001526 0.001486 0.001412

Proportion of Variance 0.8481 0.1519 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000

Cumulative Proportion 0.8481 1.0000 0.999990 1.000000 1.000000 1.000000 1.000000 1.000000 1.000000 1.000000

PC11 PC12 PC13 PC14 PC15 PC16 PC17 PC18 PC19

Standard deviation 0.001396 0.001323 0.001271 0.001257 0.001241 0.001176 0.00116 0.001118 0.001103

Proportion of Variance 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.00000 0.000000 0.000000

Cumulative Proportion 1.000000 1.000000 1.000000 1.000000 1.000000 1.000000 1.00000 1.000000 1.000000

PC20 PC21 PC22 PC23 PC24 PC25 PC26 PC27

Standard deviation 0.001046 0.0009855 0.0009754 0.0009535 0.0008699 0.000864 0.0008552 0.0008496

Proportion of Variance 0.000000 0.0000000 0.0000000 0.0000000 0.0000000 0.000000 0.0000000 0.0000000

Cumulative Proportion 1.000000 1.0000000 1.0000000 1.0000000 1.0000000 1.000000 1.0000000 1.0000000

PC28 PC29 PC30 PC31 PC32 PC33 PC34 PC35

Standard deviation 0.000775 0.0007688 0.0007352 0.0007196 0.0006782 0.0006499 0.0006217 0.0005897

Proportion of Variance 0.000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000

Cumulative Proportion 1.000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000

PC36 PC37 PC38 PC39 PC40 PC41 PC42 PC43

Standard deviation 0.0005174 0.0004684 0.0004377 0.0004037 0.0003869 0.0003545 0.000338 0.0003038

Proportion of Variance 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.000000 0.0000000

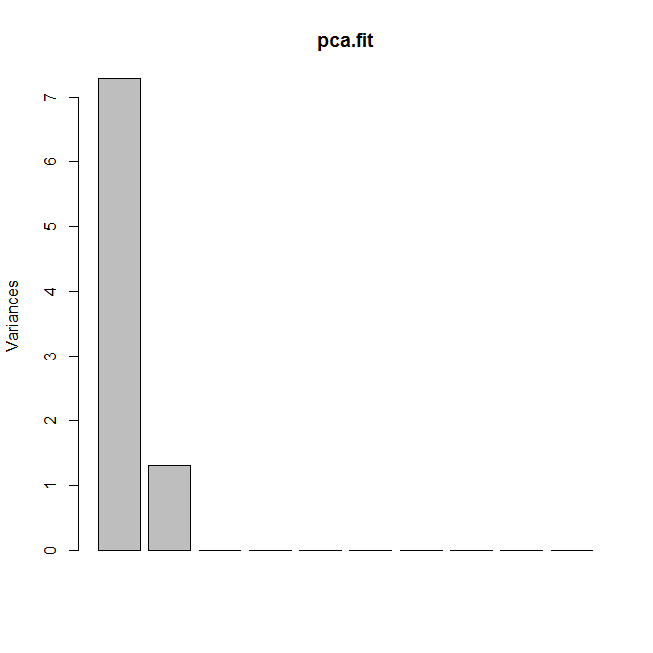
Cumulative Proportion 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.000000 1.0000000

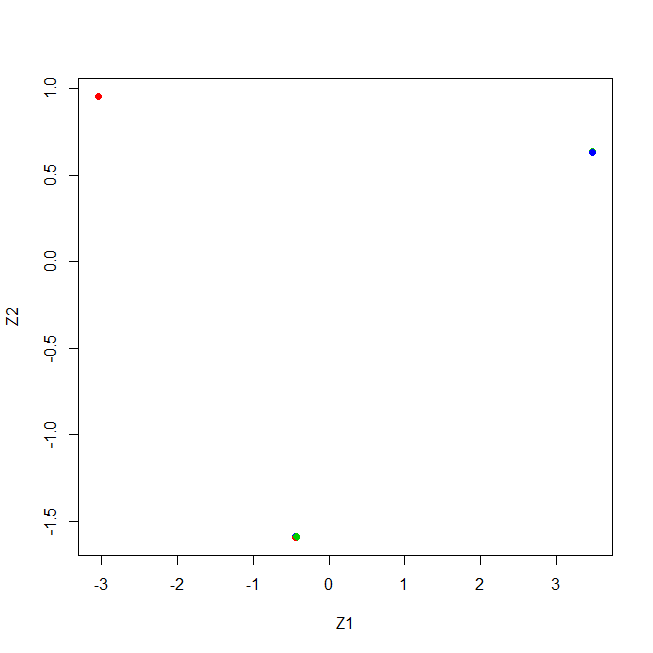
PC44 PC45 PC46 PC47 PC48 PC49 PC50

Standard deviation 0.0002983 0.0002516 0.0002185 0.0001896 0.0001695 0.0001502 6.54e-05

Proportion of Variance 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.00e+00

Cumulative Proportion 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.00e+00





The three classes appear separated.

c) K-means clustering is performed with **k = 3**

**Output :**

K-means clustering with 3 clusters of sizes 20, 20, 20

Cluster means:

[,1] [,2] [,3] [,4] [,5] [,6] [,7]

1 5.4995556691 -0.0001130635 0.0002187017 0.0001345330 -0.0001029434 -7.047398e-05 -3.164311e-04

2 -0.0005770699 3.5002011111 0.0002357237 0.0002245423 0.0002482695 -8.380968e-05 -8.143271e-05

3 0.9997493359 0.0002871451 -0.0001017394 -0.0000334065 0.0002686169 1.239387e-04 7.125200e-05

[,8] [,9] [,10] [,11] [,12] [,13] [,14]

1 -1.755812e-04 1.422970e-04 -6.168786e-05 -1.295236e-04 -6.592217e-05 2.217174e-04 1.891016e-04

2 1.881360e-04 -6.889968e-05 -3.594279e-04 9.252549e-05 1.217535e-05 2.070529e-05 -1.942211e-05

3 -7.141589e-05 2.663629e-04 -3.366666e-04 3.333376e-05 -4.478199e-04 8.416392e-05 -2.648504e-04

[,15] [,16] [,17] [,18] [,19] [,20] [,21]

1 3.408722e-04 -2.953151e-04 -0.0002675800 -0.0002994502 -0.0001454027 -0.0003728389 1.397908e-04

2 -4.927852e-05 -1.537389e-05 0.0003704810 0.0001021034 0.0001108084 0.0000392188 8.434349e-05

3 -3.706510e-04 -2.761375e-04 -0.0001222273 -0.0002475101 0.0002914270 0.0003305047 2.989011e-04

[,22] [,23] [,24] [,25] [,26] [,27] [,28]

1 -1.837680e-04 9.454004e-05 -1.487155e-04 1.466122e-04 -9.975076e-05 -2.004699e-04 -4.406043e-05

2 1.355016e-04 2.425147e-04 2.230595e-04 -1.283606e-05 5.536242e-05 -5.209517e-05 -1.869655e-04

3 9.628414e-06 3.430629e-04 -6.764444e-05 -5.428870e-05 -1.897886e-05 -2.397714e-04 -2.681083e-04

[,29] [,30] [,31] [,32] [,33] [,34] [,35]

1 -0.0003331215 0.0001205176 0.0001865945 -5.647941e-05 -1.318039e-04 5.176023e-05 4.515253e-04

2 0.0001686262 0.0004833994 0.0001242470 2.779779e-04 2.659215e-06 -3.008496e-04 -9.878963e-05

3 -0.0004224605 0.0003861232 -0.0001176832 1.747618e-04 1.259342e-05 1.123157e-04 1.755353e-04

[,36] [,37] [,38] [,39] [,40] [,41] [,42]

1 2.221272e-04 2.019379e-04 7.428501e-05 1.384942e-04 -1.074317e-04 -8.158249e-05 -0.0004816672

2 -5.181785e-05 1.199431e-04 3.798469e-04 8.711286e-05 2.856175e-05 -5.578925e-05 -0.0001168167

3 -2.482336e-04 -5.547728e-05 -2.005418e-04 1.605739e-06 2.451343e-04 2.823288e-04 0.0001288931

[,43] [,44] [,45] [,46] [,47] [,48] [,49]

1 -6.137097e-05 4.385944e-04 1.776743e-04 -1.933974e-04 2.356829e-04 -1.753904e-04 -5.551896e-05

2 3.579308e-05 -1.065742e-04 -1.876941e-04 -3.715238e-06 9.752815e-05 2.529988e-05 -9.447102e-05

3 -1.863985e-04 5.772604e-05 1.915198e-05 1.337425e-04 4.679337e-06 -2.857036e-05 1.545212e-04

[,50]

1 2.823891e-05

2 2.950203e-04

3 4.691068e-05

Clustering vector:

[1] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1

[53] 1 1 1 1 1 1 1 1

Within cluster sum of squares by cluster:

[1] 0.0009623654 0.0009181103 0.0009701958

(between\_SS / total\_SS = 100.0 %)

Available components:

[1] "cluster" "centers" "totss" "withinss" "tot.withinss" "betweenss" "size"

[8] "iter" "ifault"

Summary of k-means clustering with k = 3

Length Class Mode

cluster 60 -none- numeric

centers 150 -none- numeric

totss 1 -none- numeric

withinss 3 -none- numeric

tot.withinss 1 -none- numeric

betweenss 1 -none- numeric

size 3 -none- numeric

iter 1 -none- numeric

ifault 1 -none- numeric

Using the table function to compare the k-means clustering and true class labels. The table below indicates a perfect match. The observations are perfectly clustered.

true\_labels 1 2 3

1 0 0 20

2 0 20 0

3 20 0 0

d) Now k-means clustering is performed using **k = 2 and k = 4.**

When k = 2, all observations of one of the clusters is now absorbed in one of the two clusters.

> k.out$cluster

[1] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1

[53] 1 1 1 1 1 1 1 1

true\_labels 1 2

1 0 20

2 0 20

3 20 0

When k = 4, the first cluster is split into two clusters.

> k.out$cluster

[1] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 4 4 4 1 4 1 4 4 4 4 4 4

[53] 4 4 4 4 1 4 1 4

>table(true\_labels,k.fit$cluster)

true\_labels 1 2 3 4

1 10 0 10 0

2 0 0 0 20

3 0 20 0 0

e) Now k-means clustering with **k = 3 is performed on the first two principal score vectors** rather than on the raw data. All observations are perfectly clustered now.

> k.fit$cluster

[1] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 1 1 1 1 1 1 1 1 1 1 1

[53] 1 1 1 1 1 1 1 1

> table(true\_labels,k.fit$cluster)

true\_labels 1 2 3

1 0 0 20

2 0 20 0

3 20 0 0

f) Now the scale function is used, and k-means clustering is performed with k = 3 on the data after scaling each

variable to have standard deviation one. The results obtained from the unscaled data are better than the results obtained from the scaled data as scaling the data affects the distance between the observations.

> table(true\_labels,k.fit$cluster)

true\_labels 1 2 3

1 10 5 5

2 15 4 1

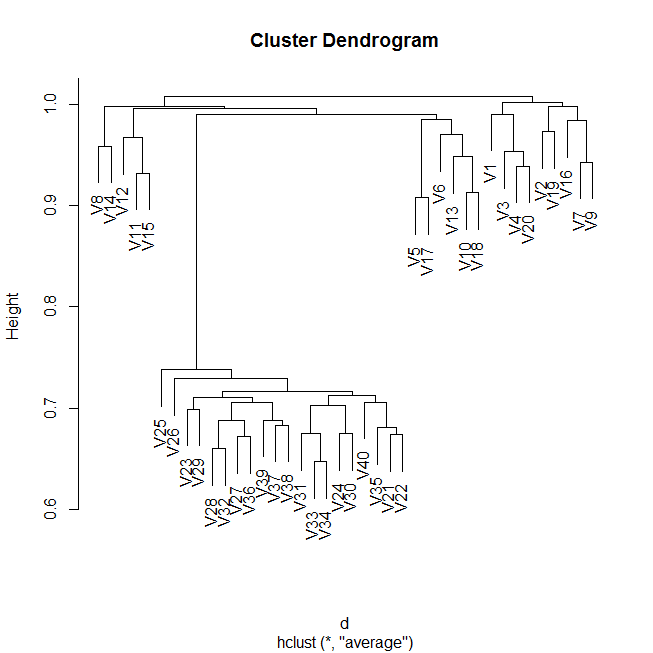
3 7 6 7

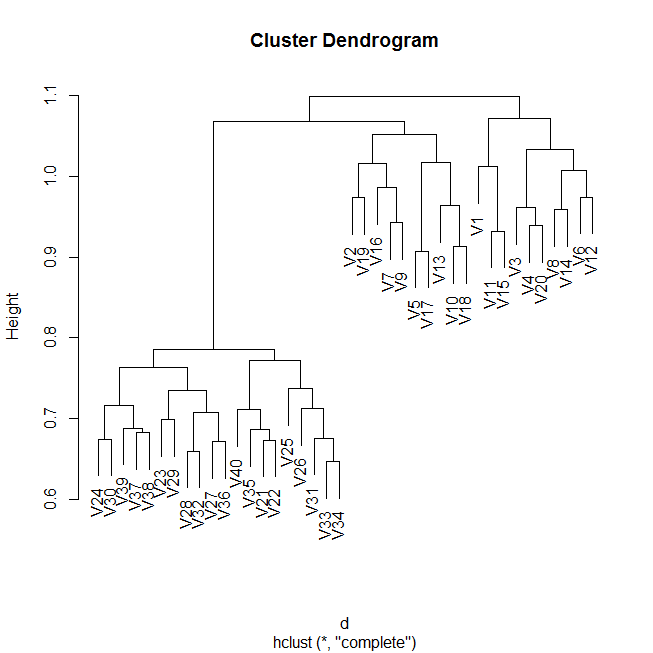
3)

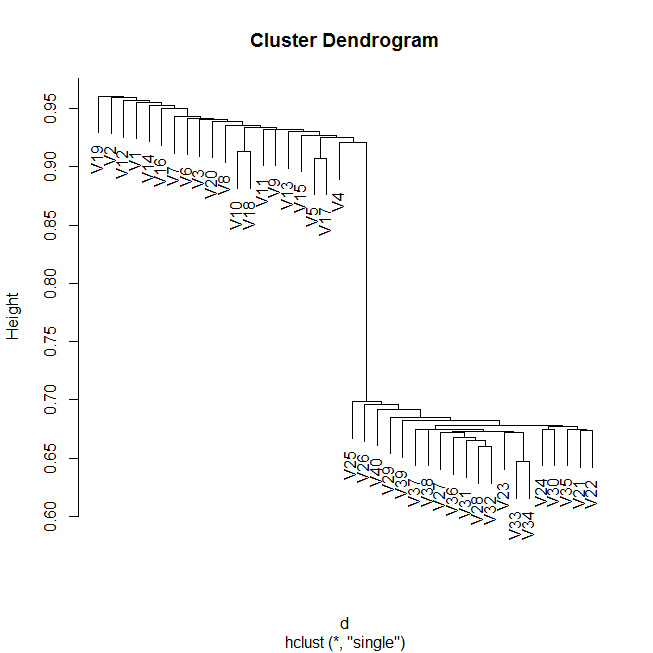
a) Ch10Ex11.csv is a gene expression data set that consists of 40 tissue samples with measurements on 1,000 genes. The first 20 samples are from healthy patients, while the second 20 are from a diseased group. The data was loaded into the R environment using the read.csv().

b) Hierarchical clustering was applied to the samples using correlation-based distance. The correlation of the data was found using the cor function then the distance was used to perform hierarchical clustering. Hcclust was performed using the methods average, complete and single.

Three clusters are obtained when the average clust is used. Two clusters are obtained when the complete and single clusts are used. The results depend on the type of linkage used and they are quite different from each other as seen in the graphs below. Only the single and complete method divide the data into two groups.







c)

Principal Component Analysis can be used to see which genes differ the most across the two groups of healthy and diseased patients. The PCA gives the absolute values of the loading vectors for the genes which can be used to describe the variance. Also using PCA will help characterize the weight of each gene.

Importance of components:

PC1 PC2 PC3 PC4 PC5 PC6 PC7 PC8 PC9 PC10

Standard deviation 11.9409 6.06818 5.93476 5.83115 5.75209 5.70031 5.63448 5.57726 5.54943 5.50625

Proportion of Variance 0.1267 0.03271 0.03129 0.03021 0.02939 0.02887 0.02821 0.02764 0.02736 0.02694

Cumulative Proportion 0.1267 0.15939 0.19068 0.22089 0.25029 0.27915 0.30736 0.33499 0.36236 0.38929

PC11 PC12 PC13 PC14 PC15 PC16 PC17 PC18 PC19 PC20

Standard deviation 5.48852 5.46025 5.40230 5.33441 5.27756 5.21594 5.20000 5.15140 5.11600 5.05591

Proportion of Variance 0.02676 0.02649 0.02593 0.02528 0.02475 0.02417 0.02402 0.02358 0.02325 0.02271

Cumulative Proportion 0.41605 0.44254 0.46847 0.49375 0.51850 0.54267 0.56669 0.59027 0.61352 0.63623

PC21 PC22 PC23 PC24 PC25 PC26 PC27 PC28 PC29 PC30

Standard deviation 5.03836 5.01868 4.95965 4.91393 4.86397 4.81796 4.80811 4.73485 4.70098 4.65564

Proportion of Variance 0.02255 0.02238 0.02185 0.02145 0.02102 0.02062 0.02054 0.01992 0.01963 0.01926

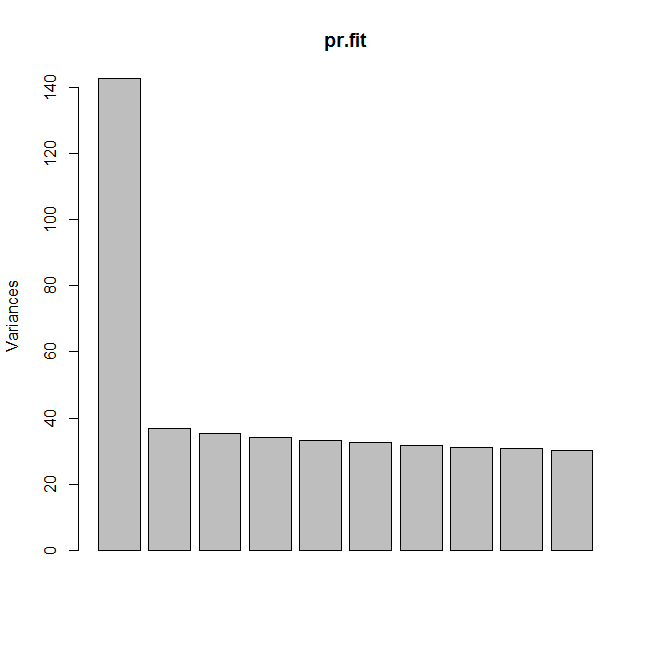
Cumulative Proportion 0.65878 0.68116 0.70301 0.72447 0.74548 0.76611 0.78665 0.80656 0.82620 0.84545

PC31 PC32 PC33 PC34 PC35 PC36 PC37 PC38 PC39 PC40

Standard deviation 4.61621 4.56733 4.53032 4.49528 4.36502 4.35858 4.26700 4.20277 4.13922 5.251e-15

Proportion of Variance 0.01893 0.01853 0.01823 0.01795 0.01693 0.01688 0.01618 0.01569 0.01522 0.000e+00

Cumulative Proportion 0.86439 0.88292 0.90115 0.91910 0.93603 0.95291 0.96909 0.98478 1.00000 1.000e+00



Index

[1] 865 68 911 428 624 11 524 803 980 822

These are the ten most different genes across the two groups. It represents the top 1% of the differing genes.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*The End\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*