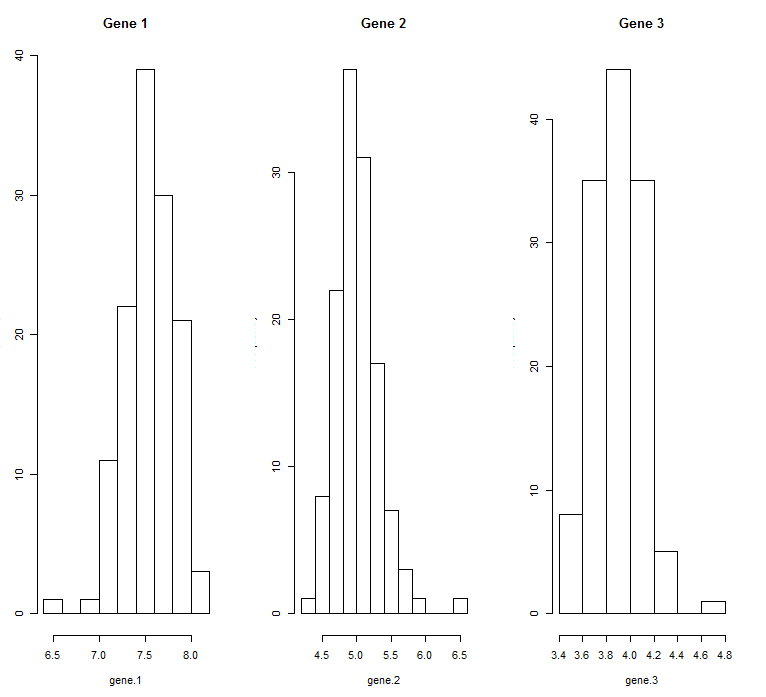


Solutions:

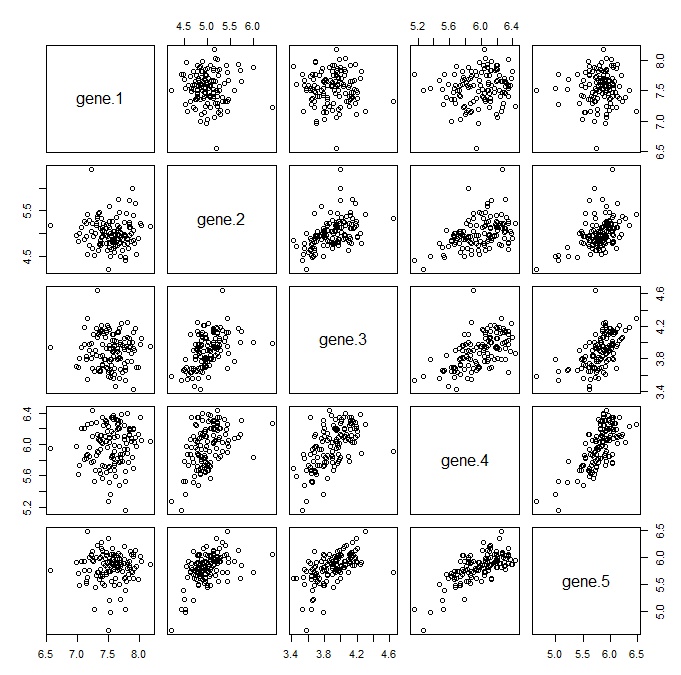
1a)

ALL.fac <- factor(ALL$BT %in% c("B","B1","B2","B3","B4"), labels=c("1","2"))

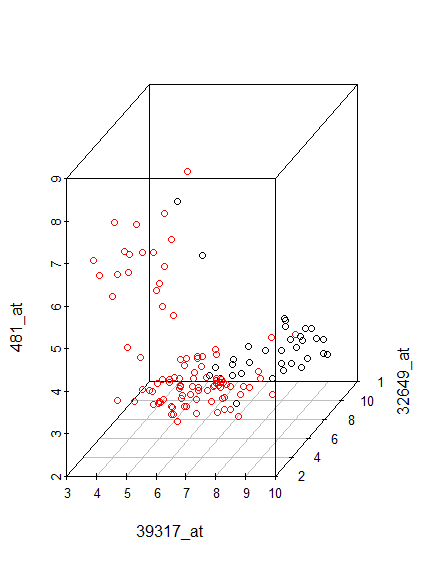
1b)



1c)



1d)



Yes, the two patient groups can be vaguely distinguished using these three genes

1e)

> table(ALL.fac,cluster1$cluster)

ALL.fac 1 2

4 31 2

3 21 74

> table(ALL.fac,cluster2$cluster)

ALL.fac 1 2 3

4 28 2 3

1. 5 20 70

* ALL has 95 B-cell, 33 T-cell
* I don’t think the clusters have been exactly divided but they are approx. divided based on B/T-cell. When k=2, 2 T-cell and 74 B-cell have been clustered into cluster-1 and in cluster-2, there are 31 T-cell and 21 from B-cell
* When k=3, 2 T-cell and 20 B-cell have been clustered into cluster-1, In cluster-2, there are 28 T-cell and 5 from B-cell , in cluster-3, there are 3 T-cell and 70 from B-cell.

1f)

> P.ALL <- prcomp(data, scale=TRUE)

> summary(P.ALL)

Importance of components:

PC1 PC2 PC3 PC4 PC5 PC6 PC7 PC8 PC9 PC10 PC11

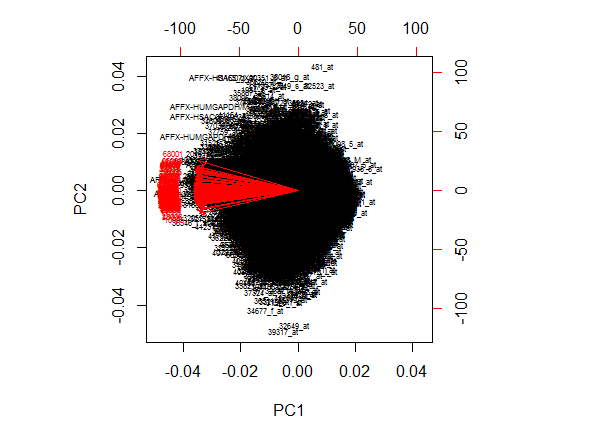
Standard deviation 10.9450 1.10132 0.93237 0.75341 0.62938 0.57412 0.53197 0.5065 0.45455 0.44240 0.41633

Proportion of Variance 0.9359 0.00948 0.00679 0.00443 0.00309 0.00258 0.00221 0.0020 0.00161 0.00153 0.00135

Cumulative Proportion 0.9359 0.94536 0.95215 0.95658 0.95968 0.96225 0.96446 0.9665 0.96808 0.96961 0.97096

* Percentage of variance explained by first principle component is 93.6%
* Percentage of variance explained by second principle component = 0.95%

1g)



The first component summarizing that PC1 is the average of the patient due to large avg expression values

1h) the genes with the biggest data loading are

> dimnames(data)[[1]][[gene.order[1]]]

[1] "481\_at"

> dimnames(data)[[1]][[gene.order[2]]]

[1] "38018\_g\_at"

> dimnames(data)[[1]][[gene.order[3]]]

[1] "41165\_g\_at"

The gemnes with the smallest data loading are

> dimnames(data)[[1]][[gene.order[12623]]]

[1] "34677\_f\_at"

> dimnames(data)[[1]][[gene.order[12624]]]

[1] "32649\_at"

> dimnames(data)[[1]][[gene.order[12625]]]

[1] "39317\_at"

1i)

Gene name and chromosomes for gene with biggest PC2 value :

> genename

[1] "SNF related kinase"

> chromosomes

3

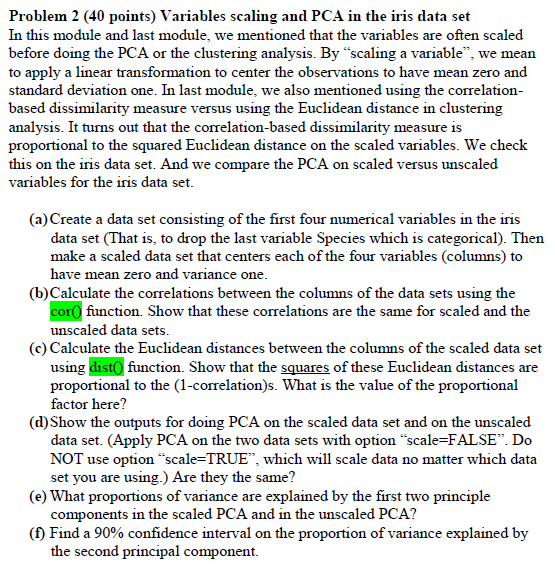
Gene name and the chromosomes for gene with smallest PC2 value:

> genenamelow

[1] "cytidine monophospho-N-acetylneuraminic acid hydroxylase, pseudogene"

> chromosomeslow

6



Solutions:

2a)

> iris.data <- iris[1:4]

> mean <- mean(iris.data[,1])

> sd <- sd(iris.data[,1])

> Sepal.Length <- NULL

> for (i in 1:150){Sepal.Length[i] <- (iris.data[i,1]-mean)/sd}

> mean <- mean(iris.data[,2])

> sd <- sd(iris.data[,2])

> Sepal.Width <- NULL

> for (i in 1:150){Sepal.Width[i] <- (iris.data[i,2]-mean)/sd}

> mean <- mean(iris.data[,3])

> sd <- sd(iris.data[,3])

> Petal.Length <- NULL

> for (i in 1:150){Petal.Length[i] <- (iris.data[i,3]-mean)/sd}

> mean <- mean(iris.data[,4])

> sd <- sd(iris.data[,4])

> Petal.Width <- NULL

> for (i in 1:150){Petal.Width[i] <- (iris.data[i,4]-mean)/sd}

> scaled.data <- cbind(Sepal.Length, Sepal.Width, Petal.Length, Petal.Width)

> xxx=data.frame(Sepal.Length, Sepal.Width, Petal.Length, Petal.Width)

2b)

> cor.scaled <- cor(scaled.data)

> cor.unscaled <- cor(iris.data)

> cor.scaled

Sepal.Length Sepal.Width Petal.Length Petal.Width

Sepal.Length 1.0000000 -0.1175698 0.8717538 0.8179411

Sepal.Width -0.1175698 1.0000000 -0.4284401 -0.3661259

Petal.Length 0.8717538 -0.4284401 1.0000000 0.9628654

Petal.Width 0.8179411 -0.3661259 0.9628654 1.0000000

> cor.unscaled

Sepal.Length Sepal.Width Petal.Length Petal.Width

Sepal.Length 1.0000000 -0.1175698 0.8717538 0.8179411

Sepal.Width -0.1175698 1.0000000 -0.4284401 -0.3661259

Petal.Length 0.8717538 -0.4284401 1.0000000 0.9628654

Petal.Width 0.8179411 -0.3661259 0.9628654 1.0000000

> all.equal(cor.scaled, cor.unscaled)

[1] TRUE

True indicates correlation

2c)

> euclidian.sqdist

Sepal.Length Sepal.Width Petal.Length

Sepal.Width 333.03580

Petal.Length 38.21737 425.67515

Petal.Width 54.25354 407.10553 11.06610

> cor.data<-as.dist(1-cor(scaled.data))

> cor.data

Sepal.Length Sepal.Width Petal.Length

Sepal.Width 1.11756978

Petal.Length 0.12824622 1.42844010

Petal.Width 0.18205887 1.36612593 0.03713457

> prop.factor<-euclidian.sqdist/cor.data

> prop.factor # propotional factor

Sepal.Length Sepal.Width Petal.Length

Sepal.Width 298

Petal.Length 298 298

Petal.Width 298 298 298

2d)

> pca.unscaled <- prcomp(iris.data, scale=FALSE)

> pca.unscaled

Standard deviations:

[1] 2.0562689 0.4926162 0.2796596 0.1543862

Rotation:

PC1 PC2 PC3 PC4

Sepal.Length 0.36138659 -0.65658877 0.58202985 0.3154872

Sepal.Width -0.08452251 -0.73016143 -0.59791083 -0.3197231

Petal.Length 0.85667061 0.17337266 -0.07623608 -0.4798390

Petal.Width 0.35828920 0.07548102 -0.54583143 0.7536574

> pca.scaled <- prcomp(scaled.data, scale=FALSE)

> pca.scaled

Standard deviations:

[1] 1.7083611 0.9560494 0.3830886 0.1439265

Rotation:

PC1 PC2 PC3 PC4

Sepal.Length 0.5210659 -0.37741762 0.7195664 0.2612863

Sepal.Width -0.2693474 -0.92329566 -0.2443818 -0.1235096

Petal.Length 0.5804131 -0.02449161 -0.1421264 -0.8014492

Petal.Width 0.5648565 -0.06694199 -0.6342727 0.5235971

They are not same

2e)

> summary(pca.unscaled)

Importance of components:

PC1 PC2 PC3 PC4

Standard deviation 2.0563 0.49262 0.2797 0.15439

Proportion of Variance 0.9246 0.05307 0.0171 0.00521

Cumulative Proportion 0.9246 0.97769 0.9948 1.00000

> summary(pca.scaled)

Importance of components:

PC1 PC2 PC3 PC4

Standard deviation 1.7084 0.9560 0.38309 0.14393

Proportion of Variance 0.7296 0.2285 0.03669 0.00518

Cumulative Proportion 0.7296 0.9581 0.99482 1.00000

Percentage of var explained by first 2 components in scaled PCA = 95.81%

Percentage of var explained by first 2 components in unscaled PCA = 97.77%

2f)

> quantile(sdevs[,2], c(0.05,0.95))

5% 95%

0.8420434 1.0434446

90% CI on the proportion of variance explained by second principle component is (0.8404, 1.0434)