**Module 12 Home Work**

**Problem 1: (60 points) Analysis of the ALL data set**

**(a)**Define an indicator variable ALL.fac such that ALL.fac=1 for T-cell patients and ALL.fac=2 for B-cell patients.

**(b)**Plot the histograms for the first three genes’ expression values in one row.

**(c)** Plot the pairwise scatterplots for the first five genes.

**(d)**Do a 3D scatterplot for the genes “39317\_at”, “32649\_at” and “481\_at”, and color according to ALL.fac (give different colors for B-cell versus T-cell patients). Can the two patient groups be distinguished using these three genes?

**(e)** Do K-means clustering for K=2 and K=3 using the three genes in (d).Compare the resulting clusters with the two patient groups. Are the two groups discovered by the clustering analysis?

**(f)** Carry out the PCA on the ALL data set with scaled variables. What proportion of variance is explained by the first principal component? By the second principal component?

**(g)**Do a biplot of the first two principal components. Observe the pattern for the loadings. What info is the first principal component summarizing?

**(h)**For the second principal component PC2, print out the three genes with biggest PC2 values and the three genes with smallest PC2 values.

**(i)** Find the gene names and chromosomes for the gene with biggest PC2 value and the gene with smallest PC2 value. (Hint: review Module 10 on searching the annotation.)

**Answer:**

(a)

library("ALL");

data(ALL);

data <- exprs(ALL)

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

(b)

gen\_1 <- data[1,]; gen\_2 <- data[2,]; gen\_3 <- data[3,]

par(mfrow=c(1,3))

hist(gen\_1, main = "Gen 1");

hist(gen\_2, main = "Gen 2");

hist(gen\_3, main = "Gen 3")

Histograms for first three gene expression values in one row are



(c) Pairwise scatter plots for first five genes

gen\_4 <- data[4,]; gen\_5 <- data[5,]

pairs(cbind(gen\_1, gen\_2, gen\_3, gen\_4, gen\_5))



(d)

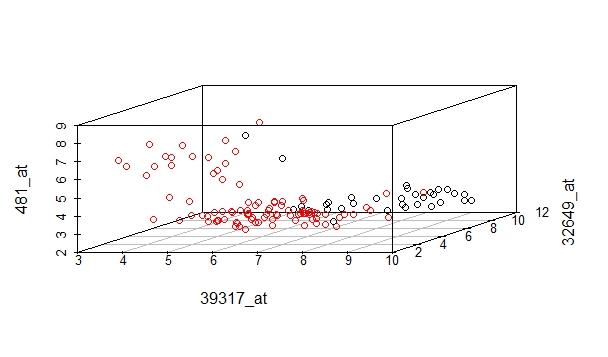
install.packages("scatterplot3d")

require(scatterplot3d)

par(mfrow=c(1,1))

d4 <- rbind(exprs(ALL[c("39317\_at","32649\_at","481\_at"),]))

scatterplot3d(t(d4),color=ALL.fac)



Yes the two patient groups can be distinguished using these 3 genes.

(e)

clu\_1 <- kmeans(t(d4),centers=2,nstart=10)

table(ALL.fac,clu\_1$cluster)

clu\_2 <- kmeans(t(d4),centers=3,nstart=10)

table(ALL.fac,clu\_2$cluster)

**Output:**

table(ALL.fac,clu\_1$cluster)

ALL.fac 1 2

1. 31 2
2. 21 74

table(ALL.fac,clu\_2$cluster)

ALL.fac 1 2 3

1 28 2 3

2 5 20 70

• ALL has 95 B-cell, 33 T-cell.

• I don’t think the clusters have been exactly divided but they are approximately 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.

(f)

> pr5.ALL <- prcomp(data, scale=TRUE)

> summary(pr5.ALL)

**Output:**

Importance of components:

PC1 PC2 PC3 PC4 PC5 PC6 PC7 PC8 PC9

Standard deviation 10.9450 1.10132 0.93237 0.75341 0.62938 0.57412 0.53197 0.5065 0.45455

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

Cumulative Proportion 0.9359 0.94536 0.95215 0.95658 0.95968 0.96225 0.96446 0.9665 0.96808

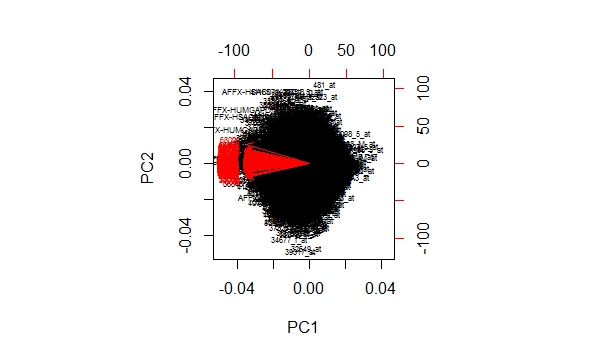
Percentage of Variance explained by first principal component = 93.6% (0.9359)

Percentage of Variance explained by second principal component =0.95% (0.0095)

(g)

par(mfrow=c(1,1))

biplot(pr5.ALL, cex=0.5)



• The first principal component is summarizing that PC1 is the average of the patients (due to large average expression values)

(h)

o7 <- order(pr5.ALL$x[,2], decreasing=T)

dimnames(data)[[1]][[o7[1]]]; dimnames(data)[[1]][[o7[2]]]; dimnames(data)[[1]][[o7[3]]]

dimnames(data)[[1]][[o7[12623]]]; dimnames(data)[[1]][[o7[12624]]]; dimnames(data)[[1]][[o7[12625]]]

**Output:**

For the second principal component PC2

Three genes with biggest loadings are:

[1] "481\_at"

[1] "38018\_g\_at"

[1] "41165\_g\_at"

Three genes with smallest loadings are:

[1] "34677\_f\_at"

[1] "32649\_at"

[1] "39317\_at"

(i)

annotation(ALL)

source("http://bioconductor.org/biocLite.R")

biocLite("hgu95av2.db");

biocLite("annotation")

library("hgu95av2.db")

library(help=hgu95av2.db)

ChrNrProbe9 <- as.list(hgu95av2CHR)

GNProbe9 <- as.list(hgu95av2GENENAME)

ChrNrProbe9[o7[1]]; GNProbe9[o7[1]]

ChrNrProbe9[o7[12625]]; GNProbe9[o7[12625]]

**Output:**

Gene name and chromosome for gene with biggest PC2 value:

> ChrNrProbe9[o7[1]]; GNProbe9[o7[1]]

$`481\_at`

[1] "3" - Chromosome Number

$`481\_at`

[1] "SNF related kinase" - Gene Name

Gene name and chromosome for gene with smallest PC2 value:

> ChrNrProbe9[o7[12625]]; GNProbe9[o7[12625]]

$`39317\_at`

[1] "6" - Chromosome Number

$`39317\_at`

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

**Problem 2 : (40 points) Variables scaling and PCA in the iris data set**

In this module and last module, we mentioned that the variables are often scaled before doing the PCA or the clustering analysis. By “scaling a variable”, we mean to apply a linear transformation to center the observations to have mean zero and standard deviation one. In last module, we also mentioned using the correlation based dissimilarity measure versus using the Euclidean distance in clustering analysis. It turns out that the correlation-based dissimilarity measure is proportional to the squared Euclidean distance on the scaled variables. We check this on the iris data set. And we compare the PCA on scaled versus unscaled variables for the iris data set.

(a)Create a data set consisting of the first four numerical variables in the iris data set (That is, to drop the last variable Species which is categorical). Then make a scaled data set that centers each of the four variables (columns) to have mean zero and variance one.

(b)Calculate the correlations between the columns of the data sets using the cor() function. Show that these correlations are the same for scaled and the unscaled data sets.

(c) Calculate the Euclidean distances between the columns of the scaled data set using dist() function. Show that the squares of these Euclidean distances are proportional to the (1-correlation)s. What is the value of the proportional factor here?

(d)Show the outputs for doing PCA on the scaled data set and on the unscaled data set. (Apply PCA on the two data sets with option “scale=FALSE”. Do NOT use option “scale=TRUE”, which will scale data no matter which data set you are using.) Are they the same?

(e) What proportions of variance are explained by the first two principle components in the scaled PCA and in the unscaled PCA?

(f) Find a 90% confidence interval on the proportion of variance explained by the second principal component, in the scaled PCA.

**Answer:**

(a)

iris\_data2 <- iris[1:4]

mean2 <- mean(iris\_data2[,1])

sd2 <- sd(iris\_data2[,1])

Sepal.Length <- NULL

for (i in 1:150){Sepal.Length[i] <- (iris\_data2[i,1]-mean2)/sd2}

mean3 <- mean(iris\_data2[,2])

sd3 <- sd(iris\_data2[,2])

Sepal.Width <- NULL

for (i in 1:150){Sepal.Width[i] <- (iris\_data2[i,2]-mean3)/sd3}

mean4 <- mean(iris\_data2[,3])

sd4 <- sd(iris\_data2[,3])

Petal.Length <- NULL

for (i in 1:150){Petal.Length[i] <- (iris\_data2[i,3]-mean4)/sd4}

mean5 <- mean(iris\_data2[,4])

sd5 <- sd(iris\_data2[,4])

Petal.Width <- NULL

for (i in 1:150){Petal.Width[i] <- (iris\_data2[i,4]-mean5)/sd5}

scaled\_data <- cbind(Sepal.Length, Sepal.Width, Petal.Length, Petal.Width)

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

(b)

co2b\_scaled <- cor(scaled\_data)

co2b\_unscaled <- cor(iris\_data2)

co2b\_scaled;co2b\_unscaled

all.equal(co2b\_scaled, co2b\_unscaled)

**Output:**

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

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

From the above output TRUE indicates same correlations for scaled and the unscaled data sets.

(c)

eu.dist <- dist(t(scaled\_data), method="euclidean")

eu.sqdist<-eu.dist^2

eu.sqdist

cr3.data<-as.dist(1-cor(co2b\_scaled))

cr3.data

prop.fac<-eu.sqdist/cr3.data

prop.fac

**Output:**

The value of the proportional factor here is 298

(d)

pc2d\_unscaled <- prcomp(iris\_data2, scale=FALSE)

pc2d\_unscaled

pc2d\_scaled <- prcomp(scaled\_data, scale=FALSE)

pc2d\_scaled

**Output:**

> pc2d\_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

> pc2d\_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 the same and this can be inferred from the output.

(e)

summary(pc2d\_unscaled)

summary(pc2d\_scaled)

**Output:**

> summary(pc2d\_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(pc2d\_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 Variance explained by first 2 components in scaled PCA = 95.81%(0.958)

• Percentage of Variance explained by first 2 components in unscaled PCA =97.77% (0.977)

(f)

data.PC <- scaled\_data

pc6 <- ncol(data.PC)

n5 <- nrow(data.PC)

nboot<-1000

de <- array(dim=c(nboot,pc6))

for (i in 1:nboot) {

dat.star <- data.PC[sample(1:n,replace=TRUE),]

de[i,] <- prcomp(dat.star, scale=TRUE)$sde

pca <-de[,2]

prop <-pca ^2/sum(de[i,]^2)

}

quantile(prop,c(0.05,0.95))

**Output:**

90% confidence interval on the proportion of variance explained by the second principal component, in the scaled PCA:

> quantile(prop,c(0.05,0.95))

5% 95%

0.2145963 0.3287979