HOMEWORK -2

QUESTION1:

USING CODE LIKE THAT DEMONSTRATED IN CLASS, DOWNLOAD THE .PNG FILE CON- TAINING AN IMAGE OF DAVID UMINSKY IN THE FILES/DATA FOLDER ON THE COURSE WEBSITE. NOTE, YOU MAY HAVE TO DOWNLOAD THIS FIRST AND THEN OPEN IT FROM YOUR OWN COMPUTER. SET X TO BE THE PIXEL INTENSITY ASSOCIATED WITH THE RED COLOR IN THE IMAGE USING CODE LIKE THAT PERFORMED IN CLASS. ANSWER THE FOLLOWING QUESTIONS:

A. WHAT ARE THE DIMENSIONS OF X? PLOT A HISTOGRAM OF THE PIXEL INTENSITIES WITHIN THE IMAGE.

The dimension of the image matrix is 322 X 451. The histogram of the image is obtained and the plot is given below.

```
##############################################################################

directory <- "/MSAN_USF/courses_fall/621-
MachineLearning/David_Uminsky.png"
#import the image as an array where each dimension of the array
represents a different level of the RGB color scheme
David <- readPNG(directory)
#convert the image to a matrix whose entries are strings that
represent color identifiers. This can be used for plotting
David_plot <- as.raster(David)
dim(David_plot)

#plot the image
grid.raster(David_plot)

David_red_only <- David[, , 1]
hist(David_red_only,main="Histogram for pixel intensity",
     xlab="pixel",
     border="blue",
     col="green",
     las=1,
     breaks=15)
##############################################################################
```

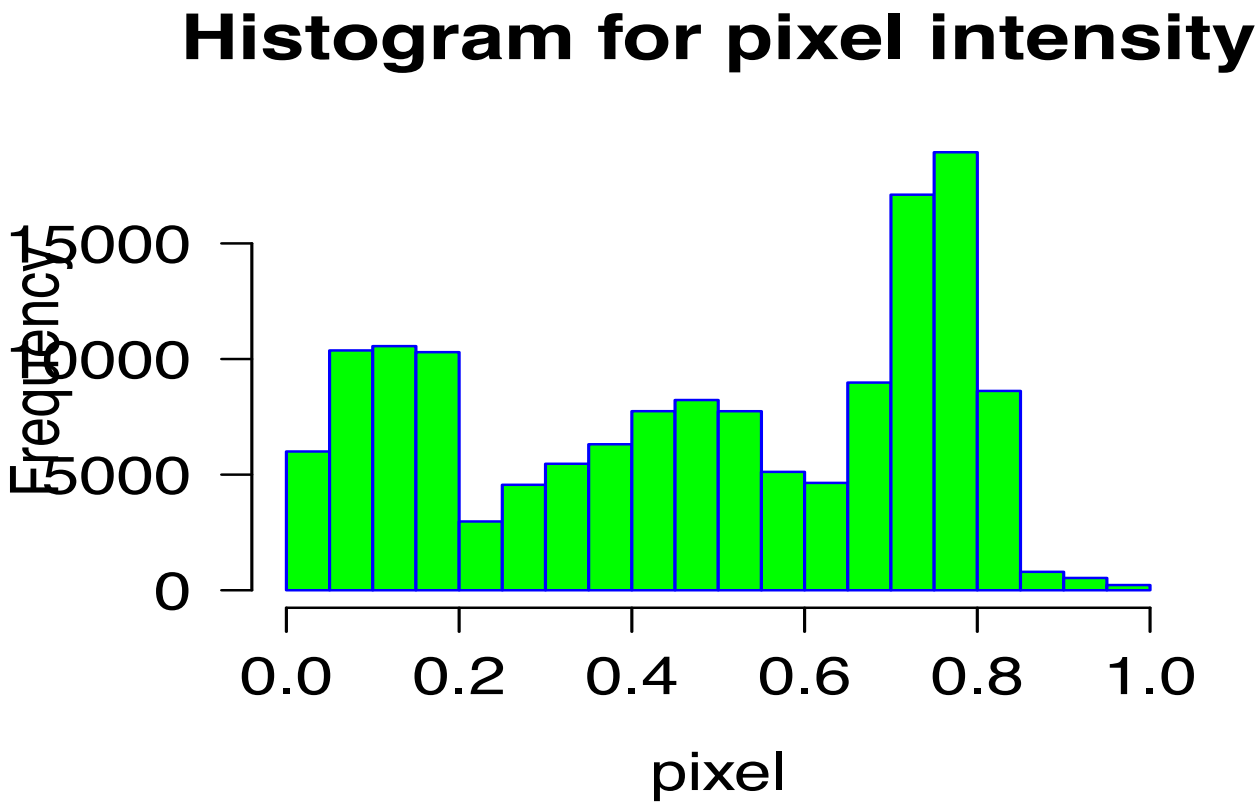

Figure 1: Histogram of pixel intensity.

B. ENSURE THAT THE COLUMNS X ARE CENTERED AND PERFORM PCA ON THIS IMAGE. PLOT THE SCREE PLOTS FOR THIS DATA, WHICH ILLUSTRATE THE PERCENT- AGE VARIATION EXPLAINED AGAINST THE NUMBER OF PRINCIPAL COMPONENTS AND THE CUMULATIVE PERCENTAGE VARIATION EXPLAINED AGAINST THE NUMBER OF PRINCIPAL COMPONENTS. HOW MANY PCS ARE NEEDED TO EXPLAIN 90% OF THE TOTAL VARIATION OF X?

From the scree plots we notice that the number of principal components needed are around 7 to explain 90% of the total variation of the X.

```
################################################################################

pr.out <- prcomp(David_red_only, scale = TRUE)
dim(pr.out$rotation)
pr.var <- pr.out$sdev^2
pve <- pr.var / sum(pr.var)
par(mfrow = c(1,2))

ret = 0
for(i in 1:length(cumsum(pve))){
  if(round(cumsum(pve)[i], 5) >= 0.90){
```

```
    ret = i
    break
  }
}
plot(pve, xlab = "Principal Component", ylab = "Proportion of
Variance Explained", ylim = c(0, 0.05), xlim = c(1, 250))

plot(cumsum(pve), xlab = "Principal Component", ylab =
"Cumulative Proportion of
     Variance Explained", ylim = c(0, 1), type = "b", xlim =
c(1, 50))
abline(h=0.9, v=ret,col = "red")
################################################################################
```

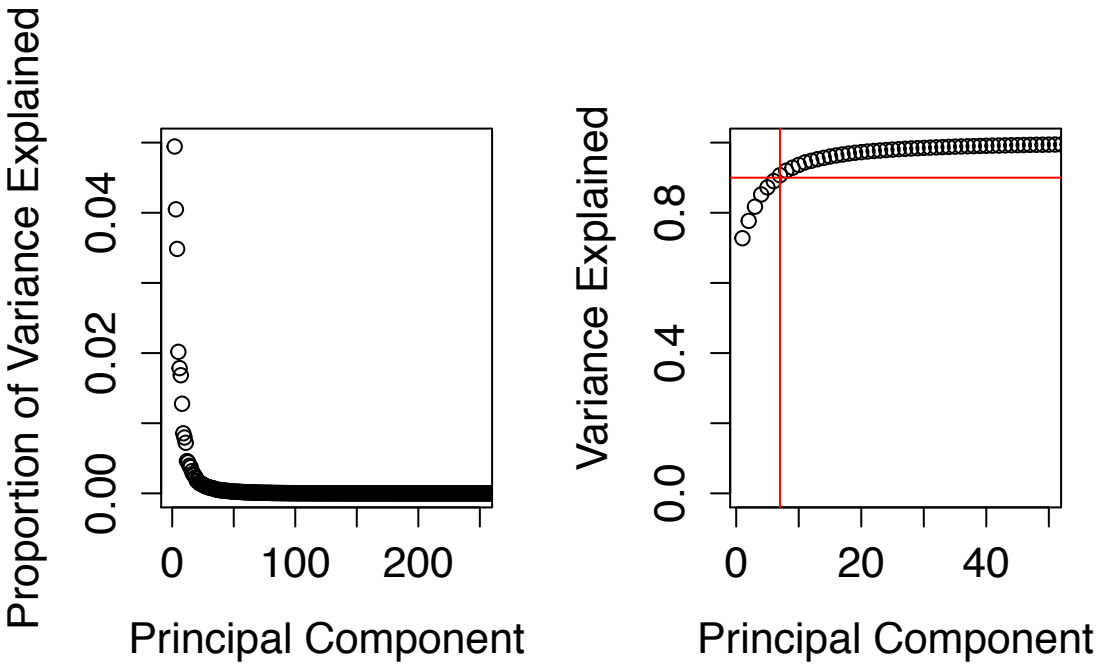

Figure:2 Scree plot of the percent variation in X. Left image is the percent variation explained by each Principal component(PC). The right side plot is cumulative sum of the percent variation explained by each PC.

C. For d=1,5,10,15,20,30,50,100,200 project the image onto the first d principal components and plot the resulting compressed image for each d. For each of the nine plots, include the cumulative percentage variation explained by the projection.

The percentage variation explained by each of the PC's is shown below.

```
> cumsum(pve)[[1]]
[1] 0.7273913
> cumsum(pve)[[5]]
[1] 0.8723462
> cumsum(pve)[[10]]
[1] 0.9363417
> cumsum(pve)[[15]]
[1] 0.9602688
> cumsum(pve)[[20]]
[1] 0.9728533
> cumsum(pve)[[30]]
[1] 0.9849385
> cumsum(pve)[[50]]
[1] 0.9942903
> cumsum(pve)[[100]]
[1] 0.9992247
> cumsum(pve)[[200]]
[1] 0.999965
```

Figure:3 : cumulative sum percent variation explained by X

```
################################################################################

plot(pve, xlab = "Principal Component", ylab = "Proportion of
W <- pr.out$rotation #the loading matrix
pc.image <- list()
num.pcs <- c(1,5,10,15,20,30,50,100,200)

#scale the original image
Image <- scale(David_red_only)
for(j in 1:length(num.pcs)){
  u.proj <- W
  #we will only use the first num.pcs PC loadings so set the
remaining to 0
  u.proj[, (num.pcs[j] + 1) : 322] <- 0

  #Make the projection
  projection <- (Image%*%u.proj)%*%t(u.proj)

  #to draw an image, values need to be between 0 and 1
  scaled <- (projection - min(as.numeric(projection)))
  scaled <- scaled / max(as.numeric(scaled))
  pc.image[[j]] <- as.raster(scaled)
```

```
}
#plot each of the images
grid.raster(pc.image[[1]])
grid.raster(pc.image[[2]])
grid.raster(pc.image[[3]])
grid.raster(pc.image[[4]])
grid.raster(pc.image[[5]])
grid.raster(pc.image[[6]])
grid.raster(pc.image[[7]])
grid.raster(pc.image[[8]])
grid.raster(pc.image[[9]])
##############################################################################
```

Figure:4 Images with increasing number of principal components
d=1,5,10,15,20,30,50,100,200

QUESTION2:

WE WILL NOW USE PCA AND PCR TO ANALYZE DATA FROM THE CANCER GENOME ATLAS (TCGA). THE DATA CONTAINS THE GENE EXPRESSION OF 217 PATIENTS WHO WERE CLASSIFIED AS EITHER HAVING A "LUMINAL A" BREAST CANCER TUMOR OR A "BASAL" BREAST CANCER TUMOR. THE DATA WE WILL EXAMINE CONTAINS A SAMPLE OF 2000 RANDOMLY SELECTED GENES ASSOCIATED WITH EACH PATIENT. LOAD THE TCGA DATA FROM THE TCGA EXAMPLE .TXT FILE IN THE FILES/DATA FOLDER ON THE COURSE WEBSITE. SET THE FIRST COLUMN OF THE MATRIX ASIDE AS A VARIABLE TUMOR.TYPE. FURTHER, SET Y = TO EXPRESSIONS OF THE FIRST GENE. KEEP THE REMAINING AS THE DESIGN MATRIX X. ANSWER THE FOLLOWING QUESTIONS.

A. FIRST, RANDOMLY SELECT 80% OF THE ROWS (PATIENTS) AND KEEP THEM AS THE TRAINING SET. SET THE REMAINING 20% ASIDE AS THE TEST SET. PERFORM PCA ON THE TRAINING SET. PLOT THE SCREE PLOTS FOR THE RESULTING PCS. WHAT NUMBER OF PCS ARE NEEDED TO EXPLAIN 85% OF THE VARIATION IN THE TRAINING SET?

The number of PC's needed to explain 85% of the variation in X of the training data is around 63. This is shown in the scree plot.

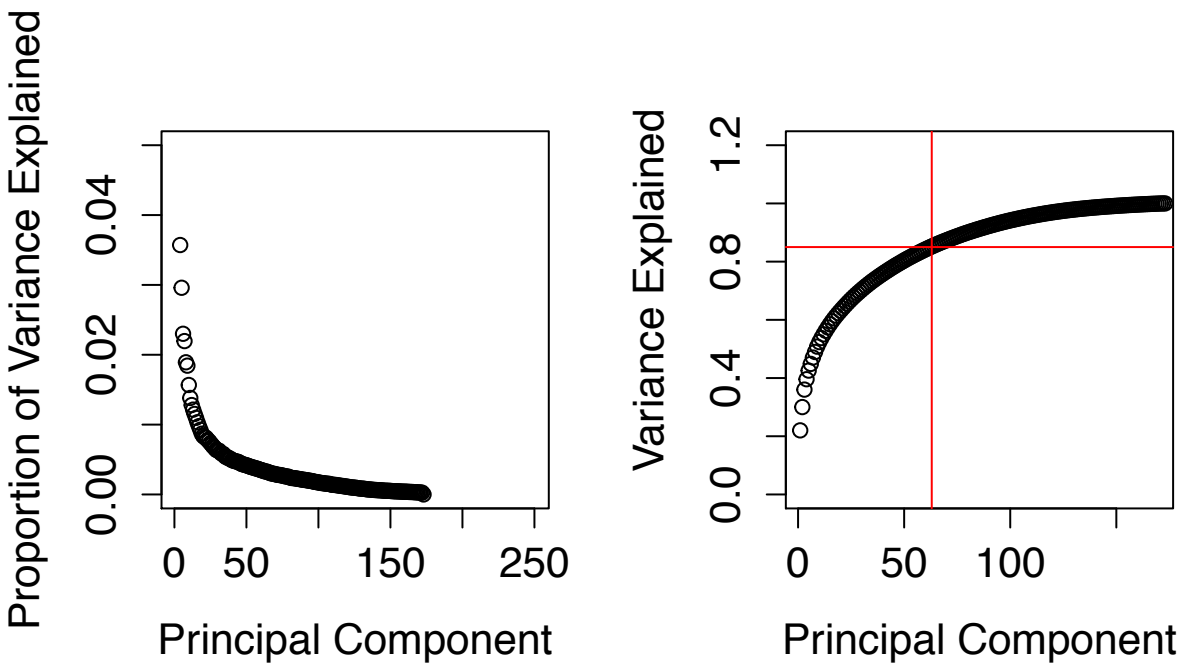

Figure:5 : Percent variation and cumulative sum percent variation explained by X

##########

```
 #choose the training and test set randomly
x1 <- model.matrix(geneData$Gene.1~., data=geneData)[,-1]
dim(x1)
 set.seed(1)
y1 <- geneData$Gene.1
train1 <- sample(1:nrow(x1), nrow(x1)*0.8)
test1 <- (-train1)
gdtrain  <- y1[train1]
tumr <- x1[train1,1]
gdtest <- y1[test1]

 #performing PCA
pr.out1 <- prcomp(x1[train1,], scale = TRUE)
#look at the names of the results of applying PCA
names(pr.out1)
pr.out1$rotation
#plot of PC1 against PC2 (the so-called bi-plot)

pr.var1 <- pr.out1$sdev^2
pve <- pr.var1 / sum(pr.var1)
par(mfrow = c(1,2))
ret = 0
for(i in 1:length(cumsum(pve))){
  if(round(cumsum(pve)[i], 5) >= 0.85){
    ret = i
    break
  }
}
plot(pve, xlab = "Principal Component", ylab = "Proportion of
Variance Explained", ylim = c(0, 0.05), xlim = c(1, 250))
plot(cumsum(pve), xlab = "Principal Component", ylab =
"Cumulative Proportion of
     Variance Explained", ylim = c(0, 1.2), type = "b", xlim =
c(1, 170))
abline(h=0.85, v=ret,col = "red")
dim(pr.out1$x)


################################################################################
```

B. PLOT A PAIRWISE SCATTER PLOT OF THE FIRST 4 PCS ON THE TRAINING DATA AND COLOR THE SCORES ACCORDING TO THE BREAST CANCER TUMOR. TYPE OF THE PATIENT. DISCUSS ANY TRENDS THAT THE PAIRWISE SCATTER PLOT REVEAL, IF ANY.

The pairwise scatter plot of the first 4 PC's is shown below . We can see clustering of the data for each of the tumor types. The first PC and subsequent shows clustering. So the first PC itself is able to segregate the two types of data well.

```
##########################################################################
pairs(pr.out1$x[,c(1,2,3,4)], pch = 21, bg = c("red",
"blue")[as.factor(tumr)])

par(xpd=TRUE)

legend(0.8, 0.1,as.vector(as.factor(c("Basal", "Normal"))),
       fill = c("red","blue"))

##########################################################################
```

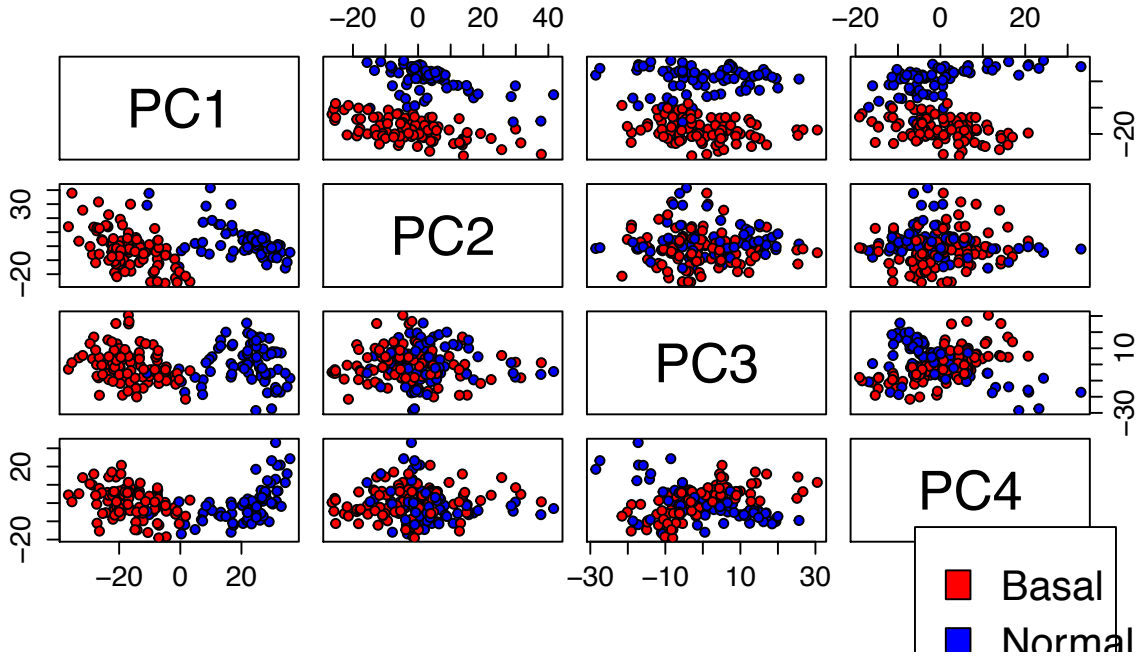

Figure6 : Pairwise scatter plot of first 4 PC's

C. Run a principal component regression of the first gene on the remaining genes in the training set. Use 10-fold cross-validation to determine the number of PCs to use, and plot the associated MSPE of the cross validation trials against the number of PCs. What is your chosen model? (i.e. how many principal components are you using, and what are the associated parameter estimates?)

Using the gene data to run a PCR of the first gene against the remaining genes. Also performing 10 fold cross validation on the training gene data set. Looking at the validation plot, the minimum MSEP occurs at number of PC's around 114 components. The r output is summarized below for the pcr.fit.

The parameter estimates of MSEP = 0.2718 and % training variance explaind = 96.16 for

114 PC's.

```
##############################################################################

fit PCR across a number of PCs

pcr.fit <- pcr(Gene.1 ~., data = geneData, subset = train1,
scale = TRUE, validation = "CV")

#look at the validation plot to choose the number of PCs

validationplot(pcr.fit, val.type = "MSEP")

summary(pcr.fit)

##############################################################################
```

	104 comps	105 comps	106 comps	107 comps	108 comps	109 comps	110 comps
CV	0.2760	0.2753	0.2758	0.2755	0.2754	0.2742	0.2744
adjCV	0.2681	0.2675	0.2681	0.2679	0.2667	0.2660	0.2660
	111 comps	112 comps	113 comps	114 comps	115 comps	116 comps	117 comps
CV	0.2732	0.2726	0.2725	0.2718	0.2732	0.2741	0.2737
adjCV	0.2642	0.2644	0.2651	0.2644	0.2663	0.2677	0.2680
	118 comps	119 comps	120 comps	121 comps	122 comps	123 comps	124 comps
CV	0.2736	0.2741	0.2743	0.2741	0.2748	0.2775	0.2800
adjCV	0.2674	0.2682	0.2687	0.2689	0.2697	0.2717	0.2739
	125 comps	126 comps	127 comps	128 comps	129 comps	130 comps	131 comps
CV	0.2795	0.2790	0.2791	0.2794	0.2790	0.2798	0.2798

Figure 7 : Summary of the PCR fit in terms of Root Mean squared Error predicted for 114 PC'

	105 comps	106 comps	107 comps	108 comps	109 comps	110 comps	111 comps
X	94.87	95.03	95.18	95.33	95.48	95.62	95.76
Gene.1	89.02	89.02	89.08	89.64	89.67	89.86	90.27
	112 comps	113 comps	114 comps	115 comps	116 comps	117 comps	118 comps
X	95.90	96.03	96.16	96.29	96.41	96.53	96.65
Gene.1	90.28	90.30	90.36	90.45	90.45	90.45	90.78
	119 comps	120 comps	121 comps	122 comps	123 comps	124 comps	125 comps
X	96.76	96.87	96.98	97.09	97.19	97.29	97.40
Gene.1	90.78	90.85	90.91	90.94	91.23	91.51	92.26
	126 comps	127 comps	128 comps	129 comps	130 comps	131 comps	132 comps
X	97.5	97.59	97.68	97.77	97.85	97.94	98.02
Gene.1	92.3	92.30	92.86	92.90	92.90	92.99	93.31

Figure 8 : Summary of the PCR fit in terms of TRAINING: % variance explained

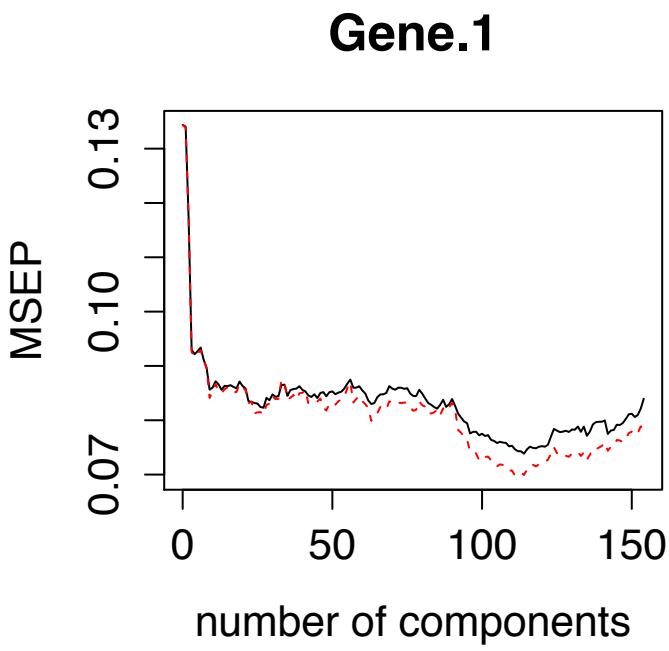

Figure 9: Validation plot

D. APPLY YOUR MODEL FROM (C) ON THE TEST SET AND CALCULATE THE MSPE.

Applying the model on the test set with 114 PC's yields MSEP = 0.1228.

```
##############################################################################
pcr.pred <- predict(pcr.fit, x1[test1, ], ncomp = 114)
#calculate the MSPE
mean((pcr.pred - gdtest)^2
##############################################################################
```

```
> #predict the values on the test set
> pcr.pred <- predict(pcr.fit, x1[test1, ], ncomp = 114)
>
> #calculate the MSPE
> mean((pcr.pred - gdtest)^2)
[1] 0.1228376
>
```

The final model is fit with all the data and with 114 Principal components.

```
##############################################################################

#final model

pcr.final.model <- pcr(y1 ~ x1, scale = TRUE, ncomp = 114)

summary(pcr.final.model)

coef(pcr.final.model)

##############################################################################
```

R output:

```
> summary(pcr.final.model)
Data:   X dimension: 217 2000
        Y dimension: 217 1
Fit method: svdpc
Number of components considered: 114
TRAINING: % variance explained
    1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps  9 comps
X    21.595    29.16    34.84    38.71    41.81    44.13    46.19    48.01    49.69
y1    1.239    22.58    30.48    33.75    33.91    34.17    35.10    37.07    37.24
    10 comps  11 comps  12 comps  13 comps  14 comps  15 comps  16 comps  17 comps
X      51.17     52.57     53.79     54.91     56.00     57.03     58.01     58.94
y1     37.46     37.46     37.46     37.81     37.86     41.13     42.38     42.59
    18 comps  19 comps  20 comps  21 comps  22 comps  23 comps  24 comps  25 comps
X      59.81     60.60     61.36     62.10     62.83     63.52     64.20     64.86
y1     44.40     44.42     45.48     45.59     47.02     49.68     49.87     50.75
    26 comps  27 comps  28 comps  29 comps  30 comps  31 comps  32 comps  33 comps
X      65.51     66.14     66.76     67.36     67.95     68.52     69.07     69.60
y1     50.93     51.12     51.23     52.55     52.55     52.59     53.05     53.05
    34 comps  35 comps  36 comps  37 comps  38 comps  39 comps  40 comps  41 comps
X      70.12     70.63     71.12     71.61     72.09     72.57     73.03     73.49
y1     53.05     53.35     53.56     56.09     58.95     59.02     59.25     59.37
    42 comps  43 comps  44 comps  45 comps  46 comps  47 comps  48 comps  49 comps
X      73.94     74.38     74.81     75.24     75.65     76.06     76.46     76.85
y1     59.49     59.57     60.08     60.12     60.14     60.48     61.14     64.46
    50 comps  51 comps  52 comps  53 comps  54 comps  55 comps  56 comps  57 comps
X      77.24     77.62     78.00     78.37     78.73     79.08     79.42     79.77
y1     64.51     66.35     66.43     68.53     68.77     68.89     69.92     69.96
    58 comps  59 comps  60 comps  61 comps  62 comps  63 comps  64 comps  65 comps
X      80.11     80.44     80.77     81.09     81.41     81.72     82.03     82.32
y1     70.03     70.20     70.30     70.57     70.65     70.78     70.82     70.84
    66 comps  67 comps  68 comps  69 comps  70 comps  71 comps  72 comps  73 comps
X      82.62     82.91     83.20     83.48     83.76     84.03     84.30     84.57
y1     70.95     71.19     71.42     71.42     71.42     71.78     71.93     72.07
    74 comps  75 comps  76 comps  77 comps  78 comps  79 comps  80 comps  81 comps
X      84.84      85.1     85.36     85.61     85.86     86.10     86.34     86.58
y1     72.10      72.1     72.24     72.31     72.68     73.25     73.38     73.62
    82 comps  83 comps  84 comps  85 comps  86 comps  87 comps  88 comps  89 comps
X      86.81     87.04     87.27     87.49     87.72     87.93     88.15     88.36
y1     74.24     74.39     74.48     74.51     75.14     76.96     77.45     77.59
    90 comps  91 comps  92 comps  93 comps  94 comps  95 comps  96 comps  97 comps
X      88.57     88.78     88.98     89.18     89.38     89.57     89.76     89.95
y1     78.02     78.04     78.12     80.12     80.50     80.66     81.76     81.76
    98 comps  99 comps  100 comps  101 comps  102 comps  103 comps  104 comps  105 comps
X      90.14     90.33      90.51      90.69      90.87      91.04      91.21      91.39
y1     82.19     82.47      83.34      83.79      83.80      83.81      84.08      84.10
```

```
    106 comps  107 comps  108 comps  109 comps  110 comps  111 comps  112 comps
X       91.56      91.72      91.89      92.05      92.21      92.37      92.53
y1      84.10      85.25      85.29      85.30      85.31      85.31      85.36
    113 comps  114 comps
X       92.68      92.83
y1      85.77      86.13
```

Coefficient estimates of few of the genes.

```
Gene.1002     -1.297996e-03
Gene.1003      4.105571e-03
Gene.1004      2.359966e-03
Gene.1005     -1.139287e-03
Gene.1006      1.692455e-03
Gene.1007      1.204904e-03
Gene.1008     -4.976692e-04
Gene.1009      1.575444e-03
Gene.1010      1.458762e-03
Gene.1011     -2.803582e-04
Gene.1012     -3.438495e-03
Gene.1013      5.601285e-04
Gene.1014      1.083239e-03
Gene.1015      1.576666e-03
Gene.1016      5.222527e-04
```

Conceptual Question 3d.

D. In terms of Σ, what is the total variation in X explained by the first d principal components.

In terms of the covariance matrix the the first d principal components are the first d diagonal elements of the covariance matrix divided by the sum of all elements of the diagonal of the covariance matrix i.e. the trace of the covariance matrix.

(please note: this question has been also answered in the conceptual section hand written document)