Homework 2

Ethen Liu, Sai Haran, Sophia Hoffman, Annie Didier, Arindam Bhattacharya

1/14/2017

Homework 2

```r
library(psych)
library(ggplot2)
library(data.table)
setwd('/Users/ethen/Desktop/northwestern/winter/MSIA 421 Data Mining/hw2')
```

Question 1

a. Run pca on the data and compute the fraction of variance explained by the first two principal components.

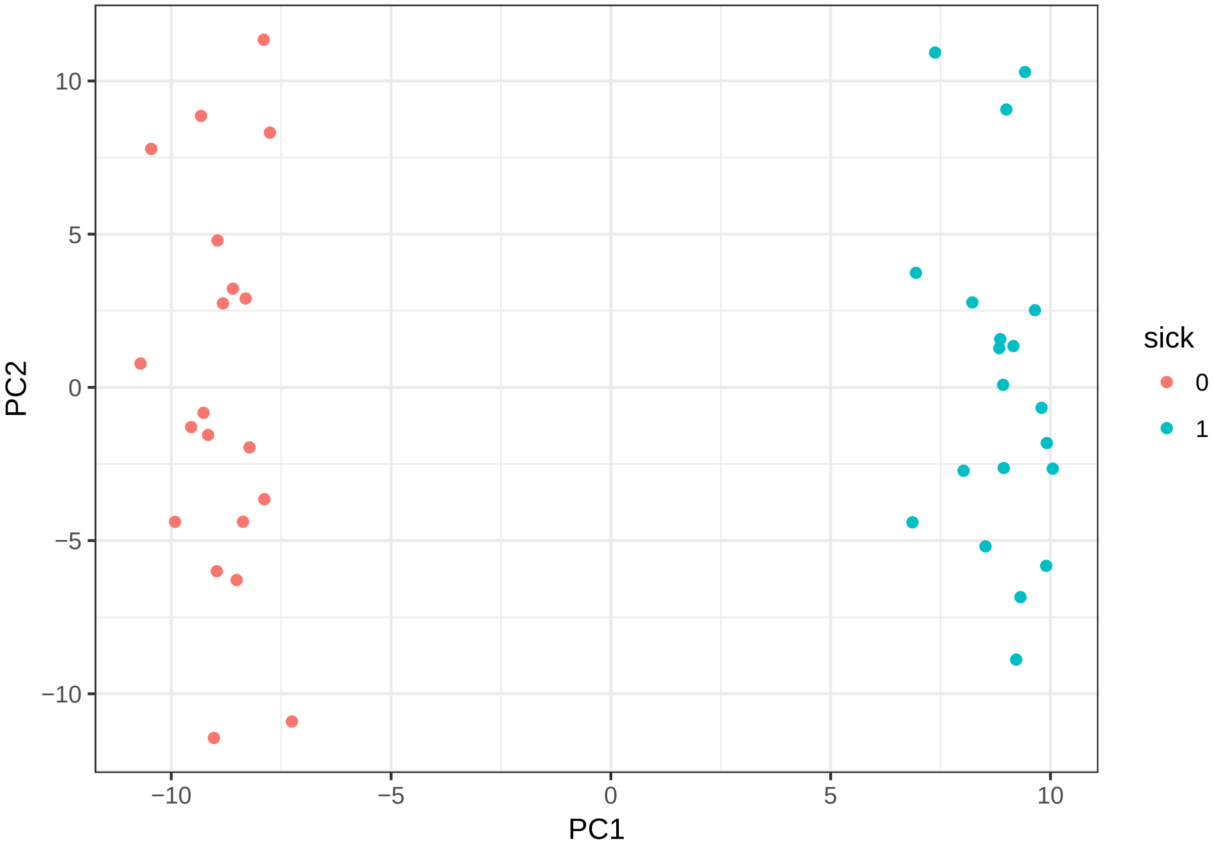
```r
gene <- fread('gene.csv')

# remove the id and the response column prior to fitting pca
gene_features <- gene[ , c(-1, -ncol(gene)), with = FALSE ]
pca <- prcomp(gene_features, scale = TRUE)
var_explained_ratio <- pca$sdev ^ 2 / ncol(gene_features)
sum(var_explained_ratio[1:2])
```

```
## [1] 0.1155751
```

b. Generate a scatter plot using the first and second principal component.

```r
gene_pca <- data.table(pca$x[, 1:2])
gene_pca[ , sick := as.factor(gene$sick) ]
```

```r
ggplot( gene_pca, aes(PC1, PC2, color = sick) ) +
geom_point() + theme_bw()
```

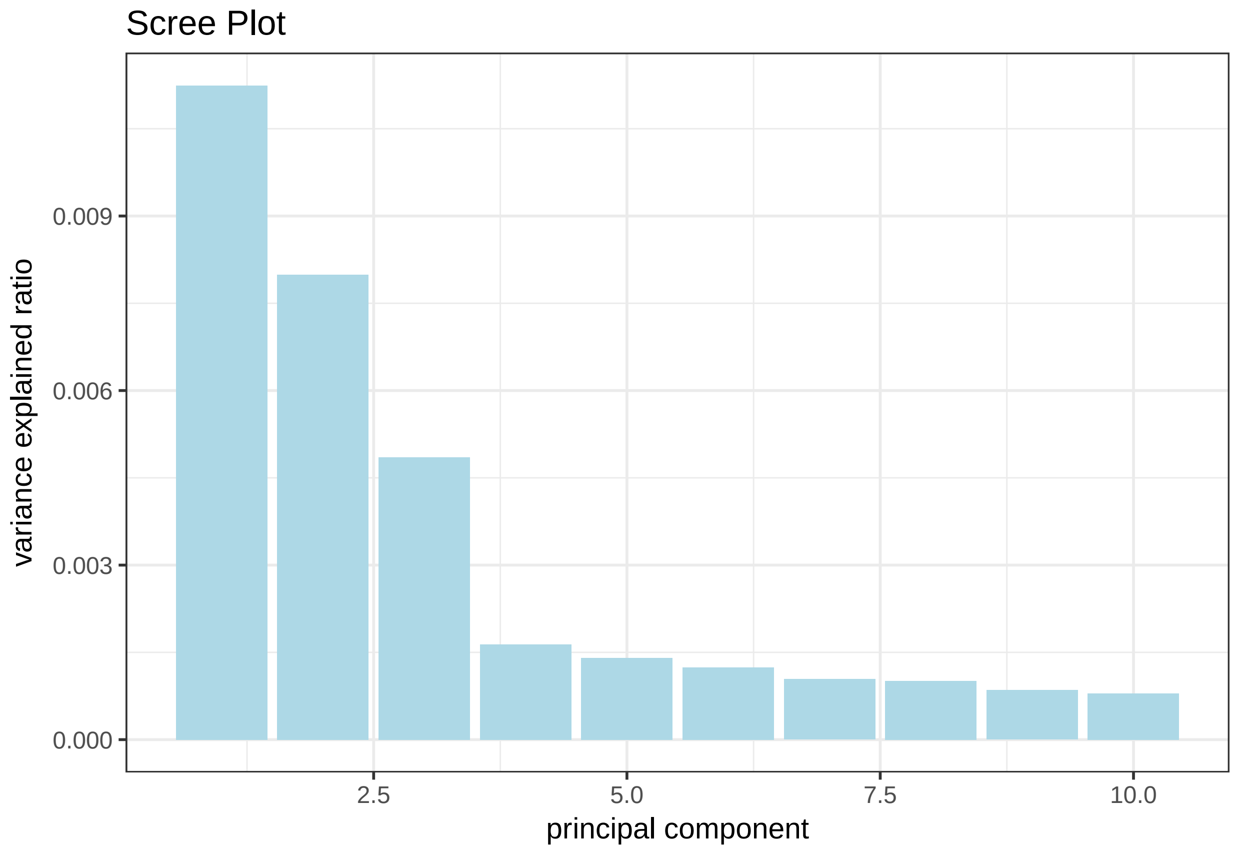


Question 2

a. Use prcomp to find the principal component analysis, plot a scree plot and comment where the elbow is.

```
news2 <- fread('news2.csv')
news2[ , V1 := NULL ]
pca1 <- prcomp(news2, scale = TRUE)
```

```
# scree plot of the pca (normalized variance explained ratio), the
# scale will be different then doing plot(pca1), since it does not
# normalize the explained variance
var_explained_ratio <- ( pca1$sdev ^ 2 / ncol(gene_features) )[1:10]
visualize_screeplot <- function(var_explained_ratio) {
    var_dt <- data.table(pca = 1:length(var_explained_ratio),
                         var = var_explained_ratio)
    scree_plot <- ggplot( var_dt, aes(pca, var) ) +
                  geom_bar(stat = 'identity', fill = 'lightblue') +
                  labs(title = 'Scree Plot', x = 'principal component',
                       y = 'variance explained ratio') +
                  theme_bw()
    return(scree_plot)
}
visualize_screeplot(var_explained_ratio)
```



The elbow of the curve is at the fourth principal component.

b. Use the psych library to perform principal component analysis and interpret the loading vectors (a.k.a rotation, principal components).

```
pca2 <- psych::principal(news2, rotate = 'none', scores = FALSE, nfactor = 3)
pca2$loadings
```

```
##
## Loadings:
##                       PC1    PC2    PC3   
## CNN                    0.274  0.742 -0.532
## CNNAndersonCooper      0.228  0.614 -0.436
## CNNAshleighBanfield    0.243  0.585 -0.495
## CNNBermanBolduan       0.232  0.635 -0.501
## CNNErinBurnett         0.197  0.554 -0.452
## CNNFareed              0.216  0.633 -0.313
## CNNHLNWeekendExp              0.101       
## CNNJakeTapper          0.235  0.659 -0.502
## CNNNewDay              0.176  0.506 -0.391
## CNNRobinMeade                             
## CNNSmerconish          0.239  0.669 -0.476
## CNNWolf                0.261  0.680 -0.538
## FoxAmerNewsroom        0.734 -0.177       
## FoxAmNewsHQ            0.897 -0.162       
## FoxandFriends          0.656 -0.168       
## FoxBretBaier           0.740 -0.188       
## FoxChrisWallace        0.807 -0.120       
## FoxMediaBuzz         0.852 -0.152
## FoxGretCarlson       0.799 -0.164
## FoxHannity           0.761 -0.159
## FoxHappeningNow      0.850 -0.179
## FoxJournalEdRpt      0.831 -0.159
## FoxKellyFile         0.785 -0.169
## FoxNeilCavuto        0.800 -0.173
## FoxOutnumbered       0.783 -0.174
## FoxRedEye            0.575 -0.104
## FoxReport            0.827 -0.149
## FoxShepardSmith      0.820 -0.156
## FoxSports
## FoxVanSusteren       0.817 -0.190
## FoxVarney            0.101
## MSNBCAlexWitt        0.141  0.654  0.527
## MSNBCAndreaMitchell  0.124  0.602  0.494
## MSNBCChrisHayes      0.113  0.630  0.535
## MSNBCHardball        0.109  0.600  0.521
## MSNBCHarrisPerry     0.104  0.574  0.521
## MSNBCLastWord               0.584  0.490
## MSNBCLive            0.147  0.678  0.571
## MSNBCRachelMaddow    0.106  0.589  0.487
## MSNBCRundown         0.114  0.501  0.486
## MSNBCUpKornacki      0.110  0.521  0.483
##
##                   PC1   PC2   PC3
## SS loadings    11.243 7.994 4.850
## Proportion Var  0.274 0.195 0.118
## Cumulative Var  0.274 0.469 0.587
```

The first principal component is simply a mix of all the features, with a much heavier weighting on FOX, effectively separating FOX viewers from non-FOX viewers. For a given channel, the first principal component averages the programs roughly equally. The second principal component contrasts FOX viewers (negative weights) vs. non-FOX viewers (positive weights). The third principal component separates viewers of all 3 channels; CNN viewers will have a large negative score, FOX viewers will be close to zero, and MSNBC viewers will have a large positive score.

Question 3

Question 4

a. Does x5, Walleye appear to group with the other fish?

```
# correlation matrix
R <- matrix(c(
    1, .4919, .2636, .4653, -.2277, .0652,
    .4919, 1, .3127, .3506, -.1917, .2045,
    .2635, .3127, 1, .4108, .0647, .2493,
    .4653, .3506, .4108, 1, -.2249, .2293,
    -.2277, -.1917, .0647, -.2249, 1, -.2144,
    .0652, .2045, .2493, .2293, -.2144, 1
), byrow = TRUE, ncol = 6)
R
```

```
##          [,1]    [,2]   [,3]    [,4]    [,5]    [,6]
## [1,]  1.0000  0.4919 0.2636  0.4653 -0.2277  0.0652
## [2,]  0.4919  1.0000 0.3127  0.3506 -0.1917  0.2045
## [3,]  0.2635  0.3127 1.0000  0.4108  0.0647  0.2493
## [4,]  0.4653  0.3506 0.4108  1.0000 -0.2249  0.2293
## [5,] -0.2277 -0.1917 0.0647 -0.2249  1.0000 -0.2144
## [6,]  0.0652  0.2045 0.2493  0.2293 -0.2144  1.0000
```

Within the centrarchid family, the pairwise correlations are all positive with magnitude greater than 0.25. However, the correlation scores of Walleye with all the other fish are either negative or very close to zero. Therefore, from a correlation of features perspective, Walleye doesn't seem to group with other types of fish.

b. Perform PCA using only x1-x4 and interpret the first two loading vectors.

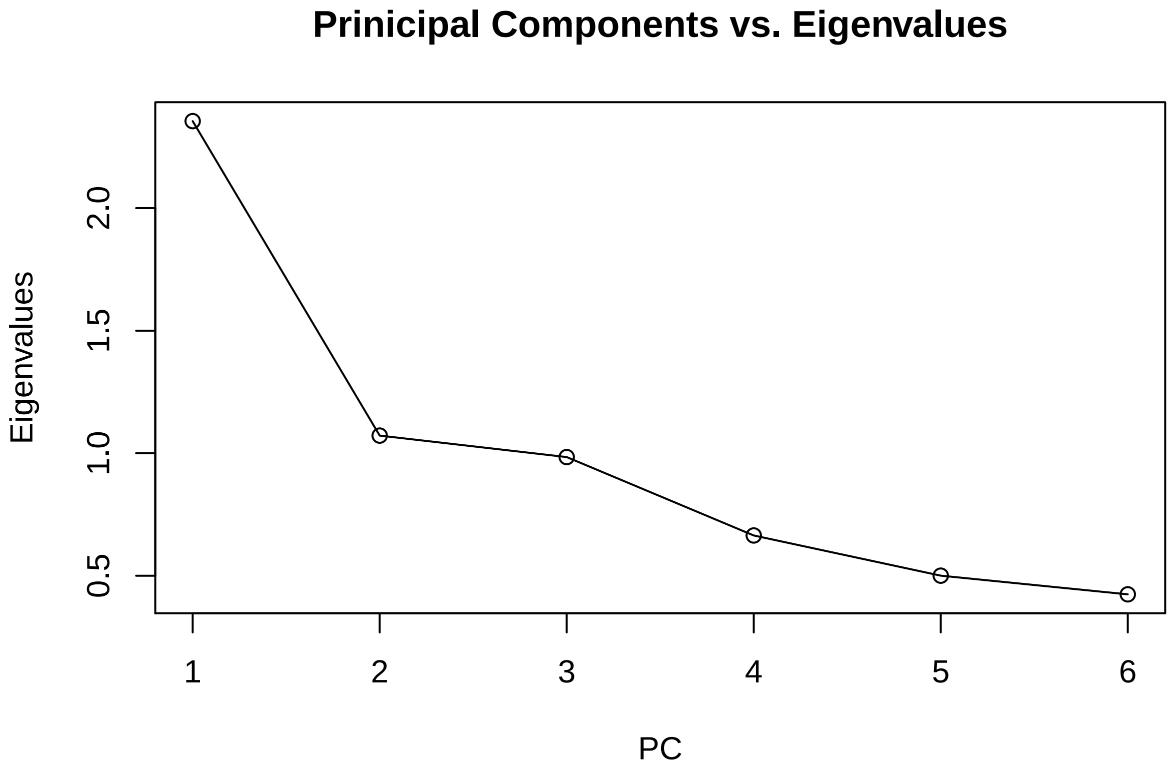
```
eig_decompose <- eigen(R[1:4, 1:4])
eig_decompose$vectors[, 1:2]
```

```
##            [,1]       [,2]
## [1,] -0.5265384 -0.4571582
## [2,] -0.5033080 -0.4119833
## [3,] -0.4427711  0.7584337
## [4,] -0.5228691  0.2146031
```

The first loading vector is simply an average of everything. The second loading vector separates x1 = Bluegill and x2 = Black crappie from x3 = Smallmouth bass and x4 = Largemouth bass.

c. Peform PCA on all six variables and generate a scree plot.

```
eig_decompose <- eigen(R)
plot(eig_decompose$values ~ c(1:6), xlab = 'PC',
     ylab = 'Eigenvalues', main = 'Prinicipal Components vs. Eigenvalues')
lines(eig_decompose$values ~ c(1:6))
```



d. What fraction of variation accounted for by the first two PCs in part c?

```
sum(var_explained_ratio[1:2])
```

```
## [1] 0.019237
```

e. Interpret the first loading vector from part c.

```
eig_decompose$vector[, 1]
```

```
## [1] -0.4753543 -0.4719328 -0.3931540 -0.4963538  0.2563177 -0.2910014
```

The first principal component is contrasting Walleye with all the other fish.

Question 5

a. Compute $X^T X$ and $X X^T$.

```
X <- matrix(c(1:4, 1, 4, 9, 16), nrow = 4)
tXX <- t(X) %*% X
tXX
```

```
##      [,1] [,2]
## [1,]   30  100
## [2,]  100  354
```

```
XXt <- X %*% t(X)
XXt
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    2    6   12   20
## [2,]    6   20   42   72
## [3,]   12   42   90  156
## [4,]   20   72  156  272
```

b., c., d. Compute the eigen value, eigen vectors of the two matrices above and comment on the eigen values.

```
eigen(tXX)
```

```
## $values
## [1] 382.37857   1.62143
##
## $vectors
##           [,1]       [,2]
## [1,] 0.2730054 -0.9620125
## [2,] 0.9620125  0.2730054
```

```
eigen(XXt)
```

```
## $values
## [1] 3.823786e+02 1.621430e+00 1.489561e-14 1.265654e-14
##
## $vectors
##             [,1]       [,2]       [,3]       [,4]
## [1,] -0.06315773 -0.5410964  0.8385856  0.0000000
## [2,] -0.22470839 -0.6533954 -0.4385264 -0.5746958
## [3,] -0.48465200 -0.3368969 -0.2538837  0.7662610
## [4,] -0.84298854  0.4083989  0.2000296 -0.2873479
```

The first two eigenvalues are essentially identical, and the last two eigenvalues of the XX^T matrix are very close to zero. This intuitively makes sense, as there are only two variables and thus two directions in which variation can occur. Therefore, we would only expect two non-zero eigenvalues regardless of whether we compute XX^T or X^TX.