Tutorial 2: Principle Components Analysis

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

In this tutorial, we will perform principle components analysis on the `ZIPcode` data set.

Exploratory Analysis

First, set the working directory

```
#setwd("/Users/linxiliu/Dropbox/Teaching/Statistical_Machine_Learning_Spring2019/Tutorials/Week_2_pca")
```

Read in the data

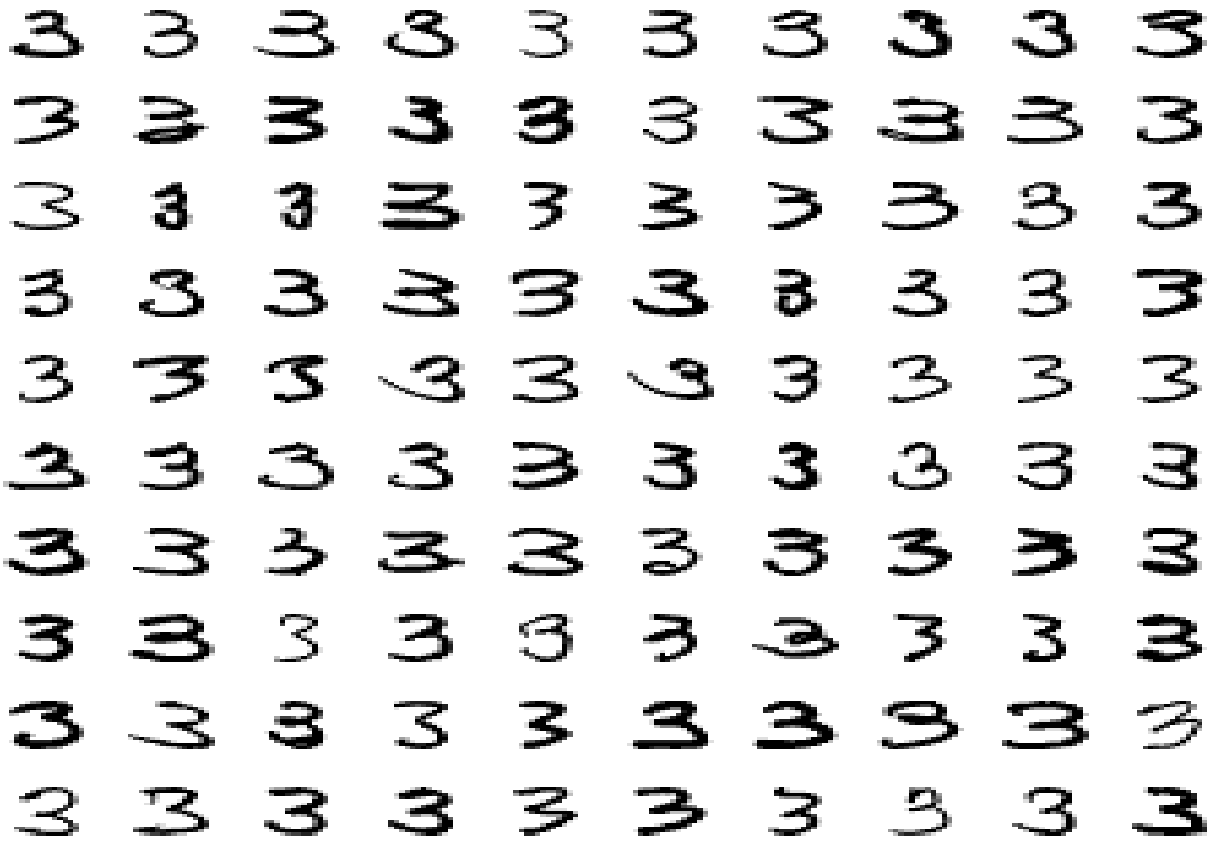
```
### all images corresponding to digit "3"
zip.3<-read.table("train.3.txt", header=FALSE, sep=",")
zip.3<-as.matrix(zip.3)
### all images corresponding to digit "5"
zip.5<-read.table("train.5.txt", header=FALSE, sep=",")
zip.5<-as.matrix(zip.5)
### n.3 and n.5 are the total number of "3"s and "5"s, respectively.
n.3<-length(zip.3[,1])
n.5<-length(zip.5[,1])
### combine two data sets together
data<-rbind(zip.3, zip.5)
```

We write a function to visualize the data. The input is a vector of length 256.

```
output.image<-function(vector) {
    digit<-matrix(vector, nrow=16, ncol=16)
    #index= seq(from=1, to =16, by=1)
    index= seq(from=16, to =1, by=-1)
    sym_digit = digit[,index]
    image(sym_digit, col= gray((8:0)/8), axes=FALSE)
}
```

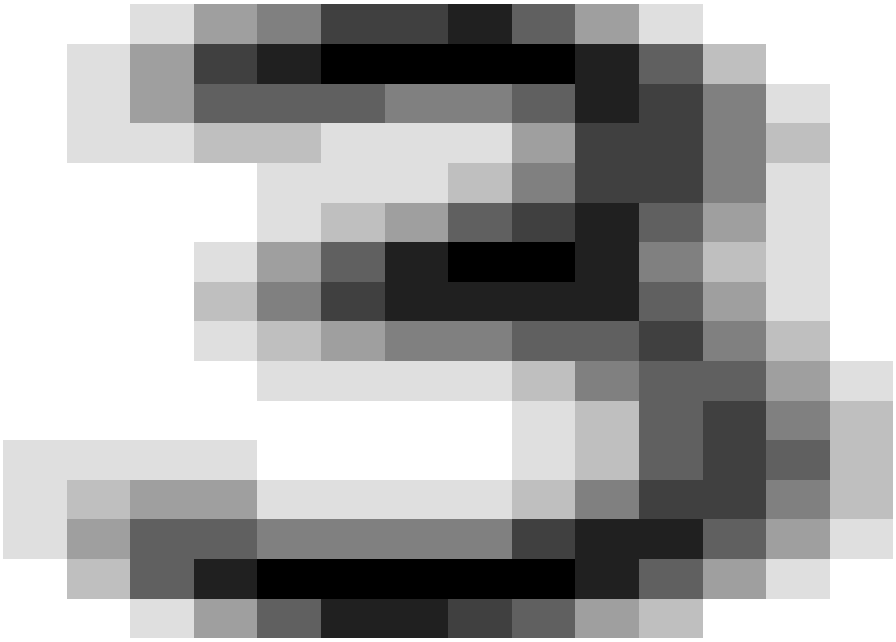
Visualize the first 100 images.

```
par(mfrow=c(10,10),mai=c(0.1,0.1,0.1,0.1))
#par(mfrow=c(1,1),mai=c(0.1,0.1,0.1,0.1))
for(i in 1:100) {
  #output.image(zip.5[i,])
    output.image(zip.3[i,])
}
```



Visualize the mean/center of the images

```
par(mfrow=c(1,1),mai=c(0.6,0.6,0.6,0.6))
mean.3<- apply(zip.3, 2, mean)
### visualize the mean ###
output.image(mean.3)
```



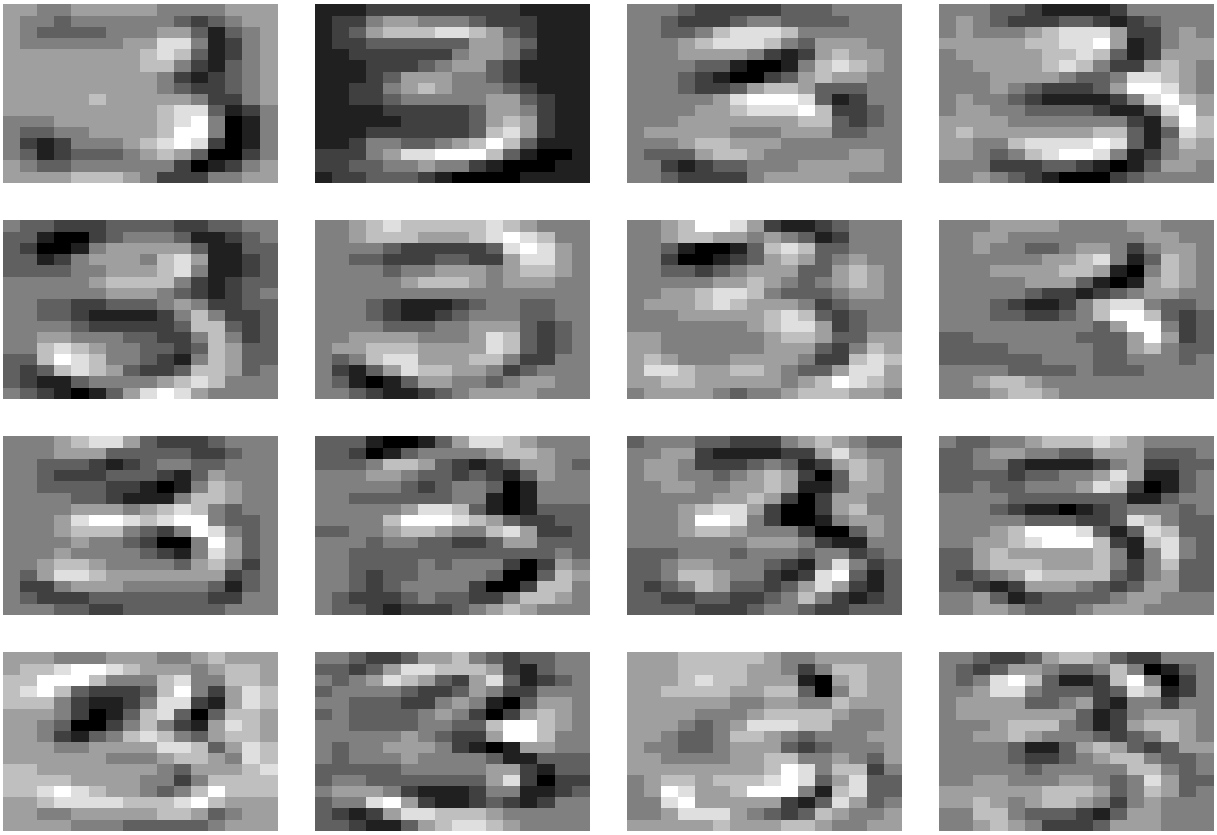
Principle Components

Before performing principle components analysis, we need to center our data. In some sense, the intensities have already been scaled, so we do not need to further scale our data in this example.

```
scaled.3<-scale(zip.3,center=TRUE, scale=FALSE)
```

We apply PCA to the image data corresponding to the digit "3", and visualize the principle components. What type of information we can obtain from the following plot?

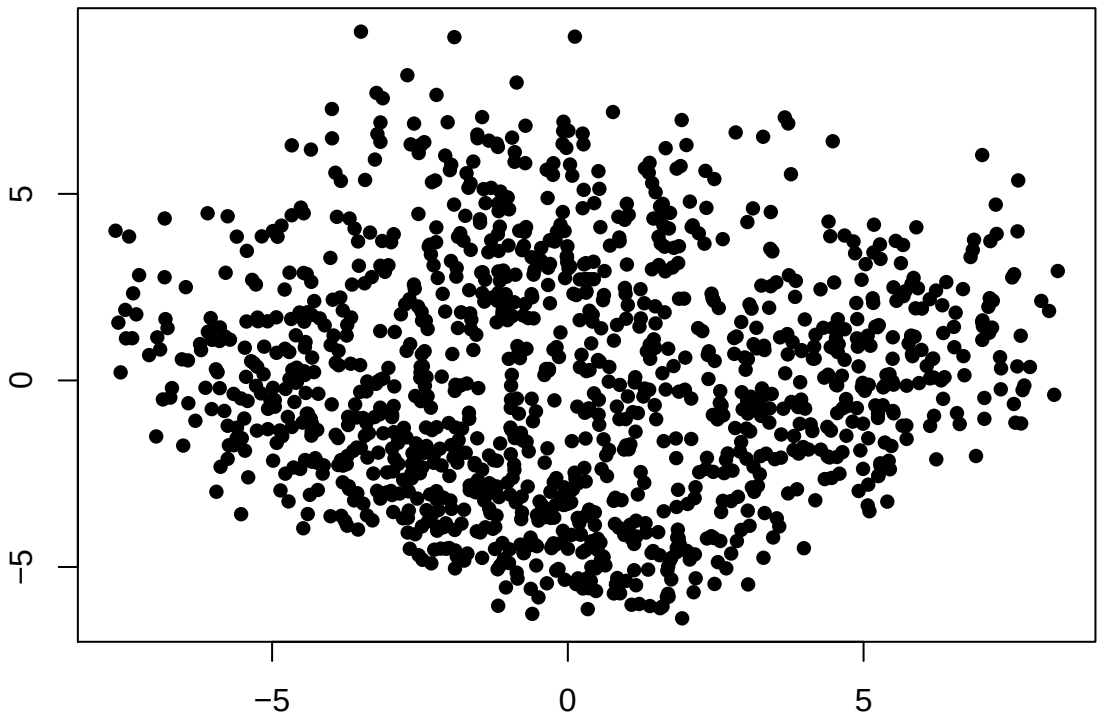
```
pca<-svd(scaled.3)
par(mfrow=c(4,4), mai=c(0.1,0.1, 0.1, 0.1))
for(j in 1:16) {
  output.image(pca$v[,j])
}
```



Projections

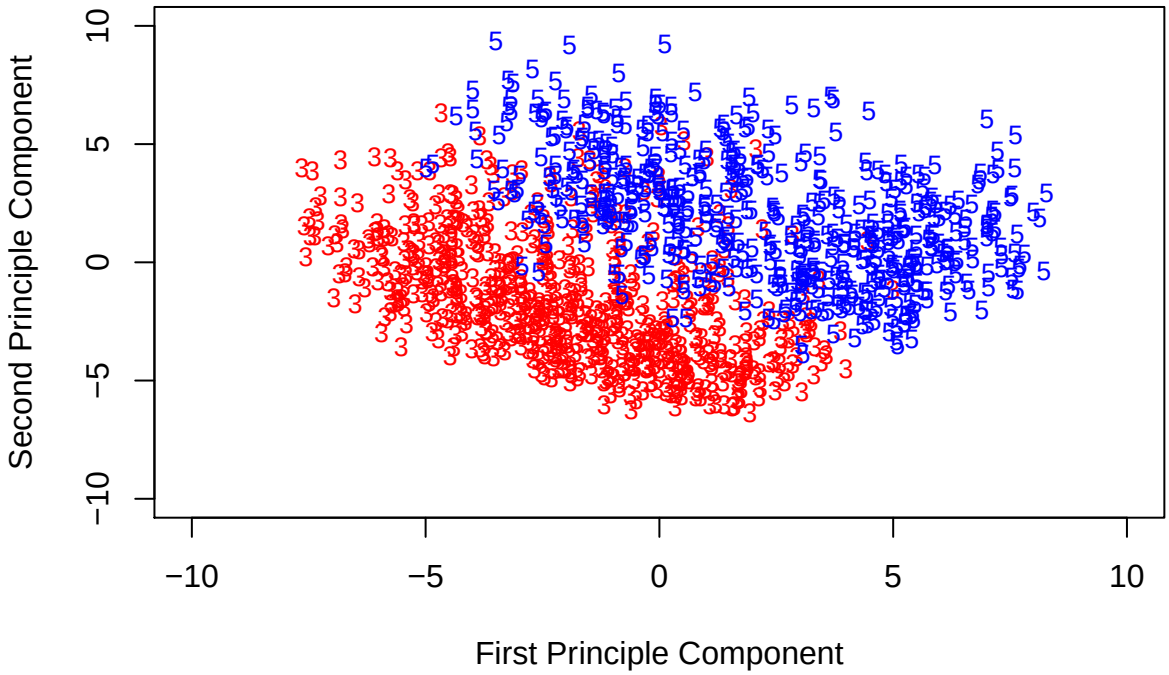
We perform principle component analysis on combined data set, and project the images onto the subspace spanned by the first two principle components.

```
scaled.data<-scale(data, center=TRUE, scale=FALSE)
pca<-svd(scaled.data)
par(mfrow=c(1,1), mai=c(0.6, 0.6, 0.6, 0.6))
plot(pca$d[1]* pca$u[,1], pca$d[2]* pca$u[, 2],pch=16, xlab="First Principle Component", ylab="Second Pr
```



We will color the points by their labels.

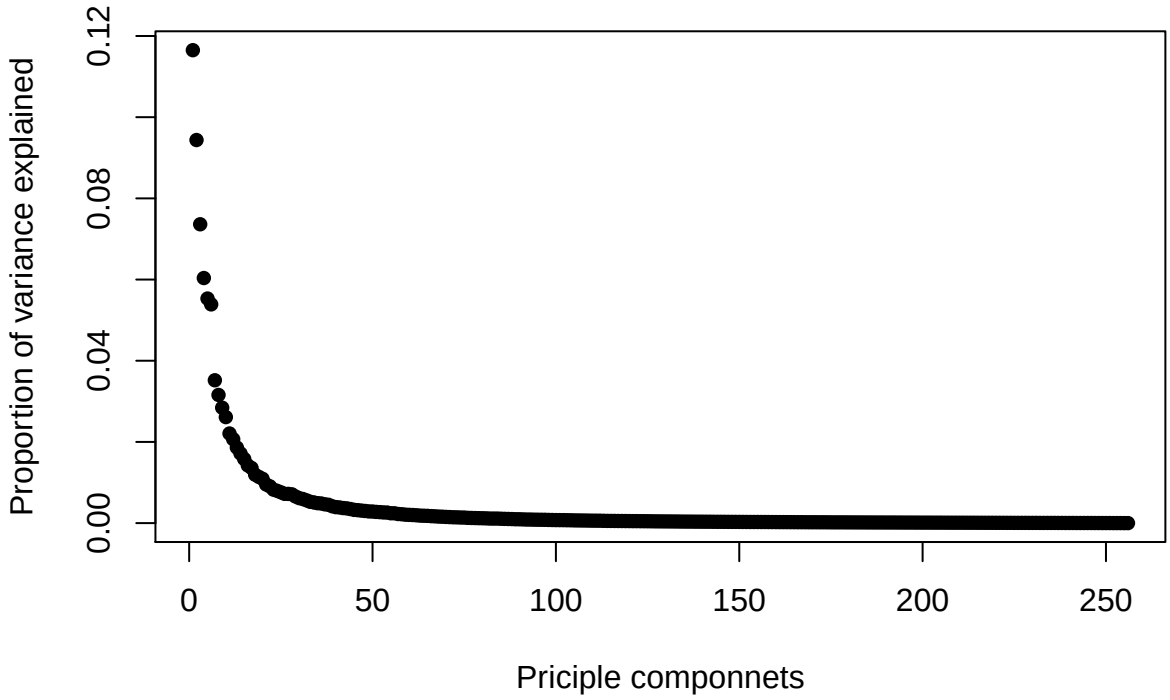
```
plot(pca$d[1]*pca$u[1:n.3, 1], pca$d[2]*pca$u[1:n.3, 2], pch="3", col="red", cex=0.8,xlim=c(-10, 10), yl
points(pca$d[1]*pca$u[(n.3+1):(n.3+n.5), 1], pca$d[2]* pca$u[(n.3+1):(n.3+n.5), 2], cex=0.8,pch="5", col
```



From the plot, we see that though the two types of images are not well seperated, most of them can be seperated based on the first two principle components.

We can also generate the scree plot

```r
plot(seq(from=1,to=256, by=1), (pca$d)^2/sum((pca$d)^2), xlab="Priciple componnets", ylab="Proportion o
```



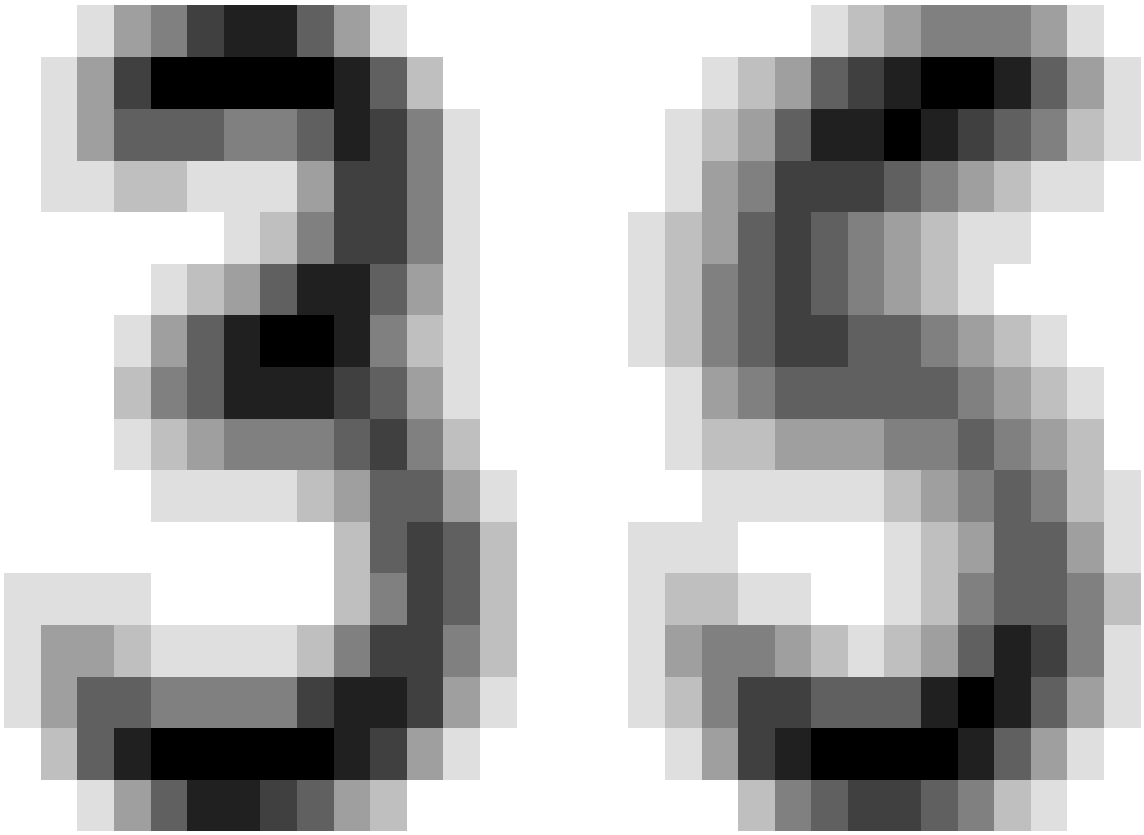
k-means clustering

Here, we pretend that we do not know the labels of our data, and try to apply k-means clustering to identify the meaning subgroups in our data. We cheat a little bit, and set $k = 2$ in the clustering algorithm.

```r
km.out<- kmeans(data, 2, nstart=50)
```

Visualize the centroid of each cluster.

```r
### Visualize the centroids of each cluster ###
digit_centers<-km.out$centers
par(mfrow=c(1,2), mai=c(0.1,0.1,0.1,0.1))
for(i in 1:2) {
    output.image(digit_centers[i,])
}
```



If we change the value of k, the number of clusters, what is the clustering result and how to interpret it?