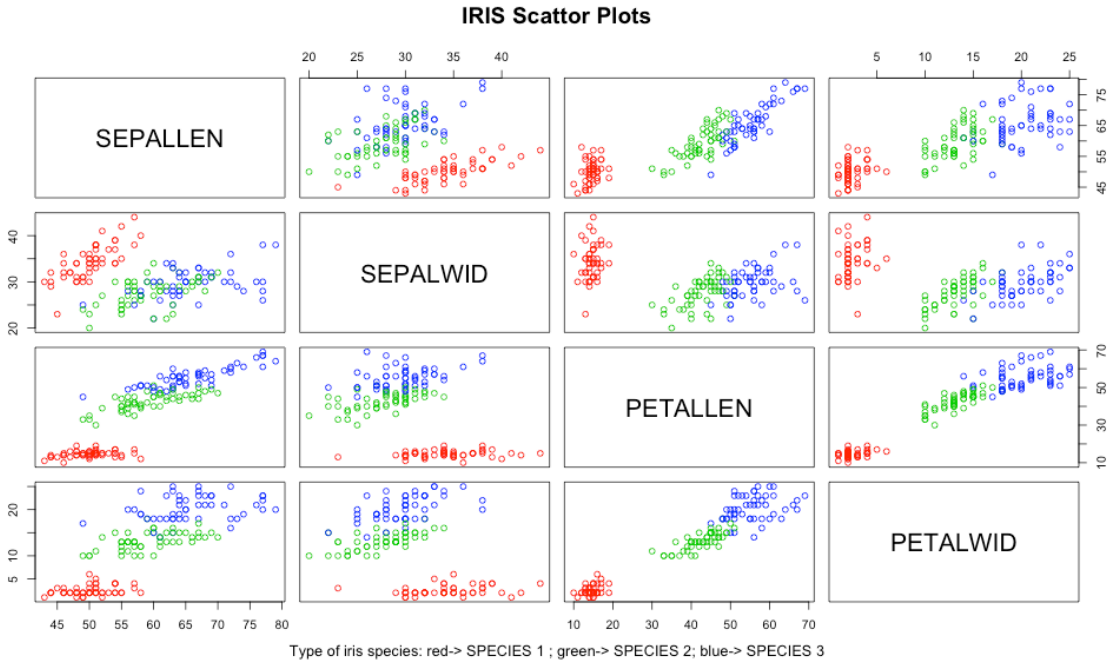
Title: Iris Flowers

Data Summary

The dataset, *IRIS,* contains 150 observations of petal length, petal width, sepal length, and sepal width of three different species of Iris flower.

```
> summary(iris)
    SEPALLEN        SEPALWID        PETALLEN        PETALWID      SPECIES
 Min.   :43.00   Min.   :20.00   Min.   :10.00   Min.   : 1.00   1:50
 1st Qu.:51.00   1st Qu.:28.00   1st Qu.:16.00   1st Qu.: 3.00   2:50
 Median :58.00   Median :30.00   Median :43.50   Median :13.00   3:50
 Mean   :58.43   Mean   :30.57   Mean   :37.58   Mean   :11.99
 3rd Qu.:64.00   3rd Qu.:33.00   3rd Qu.:51.00   3rd Qu.:18.00
 Max.   :79.00   Max.   :44.00   Max.   :69.00   Max.   :25.00

> pairs(iris[1:4], main = "IRIS Scattor Plots", col =
as.numeric(iris$SPECIES)+1)
> mtext("Type of iris species: red-> SPECIES 1 ; green-> SPECIES 2; blue->
SPECIES 3", 1, line=3.7,cex=.8)
```



Based on the summary of the data, we can see that there are 50 observations in each species. The mean and median of the variables are about the same; therefore we can conclude that there are no unusual observations. From the scatter plot, we can see there are linear relationships between variables and different species are clustering to their own group.

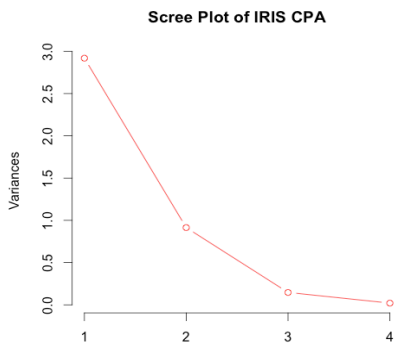
Statement of Problem

For this dataset, we want to find the number of principal component for the data. And we want to classify the species according to their petal and sepal measurements using linear discrimination and quadratic discrimination. Then we determine which method for classification is better based on the number of misclassifications.

Principal Component

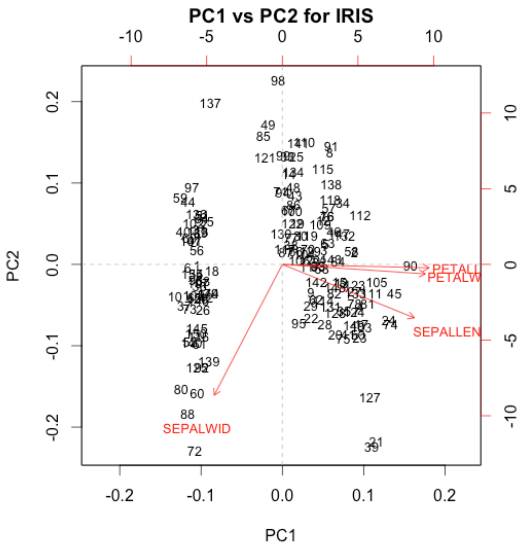
```
> pca = prcomp(iris[-5], scale = T)
> summary(pca)

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
> screeplot(pca, type="lines",col=3, ,main = "Scree Plot of IRIS CPA")
> biplot(pca, cex = 0.8, main = "PC1 vs PC2 for IRIS")
> abline(h=0, v=0, lty=2, col = 8)
```



From the scree plot and the values of the cumulative proportion of variance in the summary we can conclude that retaining 2 components would give us enough information, which account for over 95% of the variation in the original data.

```
> round(pca$rotation[,1:2],2)
             PC1   PC2
SEPALLEN    0.52 -0.38
SEPALWID   -0.27 -0.92
PETALLEN    0.58 -0.02
PETALWID    0.56 -0.07
```



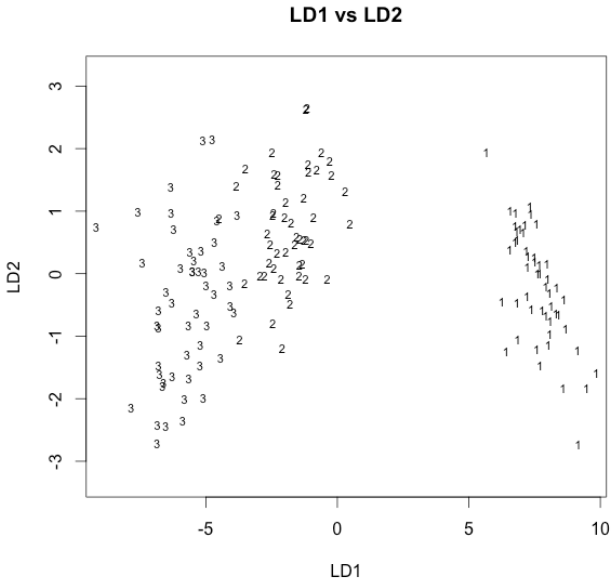
From the rotation of principal analysis that the first two principal components is a linear combination of the variables:

PC1 = 0.521*z1 - 0.269*z2 + 0.580*z3 + 0.565*z4
PC2 = -0.774*z1 - 0.923*z2 - 0.0245*z3 - 0.0669*z4

The weights of the PC1 are similar except the associate to Sepal Width variable is negative. From the biplot, we can see that sepal width is on one side of the have the graph and rest variable on the other side. PC1 accounts for about 72% of the variability in the data. PC2 are all negative weights, which might be considered as an overall size measurement. When the iris has larger measurements than the average, the PC2 will be smaller than average. PC2 explains about 23% of the variability. The combination of PC1 and PC2 explained about 95% of the variability, which is good enough. Therefore, only two components are needed to explain most the Iris data.

Linear Discrimination

```
library(MASS)
lda = lda(as.matrix(iris[-5]), iris$SPECIES)
lda
plot(lda, main = "LD1 vs LD2")
```



From the first linear discrimination vs second linear discrimination, we can see that the species of Iris are much separated. Therefore, we can use the discrimination score to classify species according to their petal and sepal measurements.

Frist, we randomly sample 50 observations from the data and used as train dataset to perform linear/ quadratic discrimination. The rest 100 samples are used as our test dataset. We will predict output of train and test data and compare their accuracy of classification to determine which model is better.

```
samp = sample(1:150, 50)
train = iris[samp,1:4]
train.ref = iris[samp,5]
```

```
test = iris[-samp, 1:4]
test.ref = iris[-samp,5]
lda.t = lda(train, train.ref)
> pre = predict(lda.t, train)$class
> pre.train = predict(lda.t, train)$class
> pre.test = predict(lda.t, test)$class
> table(pre.train, train.ref)
         train.ref
pre.train  1  2  3
        1 19  0  0
        2  0 14  1
        3  0  0 16
> table(pre.test, test.ref)
        test.ref
pre.test  1  2  3
       1 31  0  0
       2  0 33  1
       3  0  3 32
```

From the table, we can see that 98% of classification is correct for the train data and 96% of classification is correct for the test data for linear discrimination classification.

```
qda.t = qda(train, train.ref)
pre.p = predict(qda.t, train)$class
pre.q = predict(qda.t, test)$class
> table(pre.p, train.ref)
     train.ref
pre.p  1  2  3
    1 11  0  0
    2  0 20  0
    3  0  1 18
> table(pre.q, test.ref)
     test.ref
pre.q  1  2  3
    1 39  0  0
    2  0 28  2
    3  0  1 30
```

From the table, there are 98% correction classification for the train data and 97% correction for the test data for quadratic discrimination classification.

Conclusion

Linear discrimination results in 2% misclassification in train data and 4% misclassification in test data. Quadratic discrimination results in 2% misclassification in train data and 3 % misclassification in test data. Therefore, quadratic discrimination is a better method. This might because that Quadratic Discrimination function does not assume homogeneity of variance-covariance matrices.