Title: Crop Height

Data Summary

The crop data consists of 36 observations of 5 crops (soybean, corn, sugar beet, clover, and cotton). The height for each crops are recorded at four different times (X1, X2, X3, and X4).

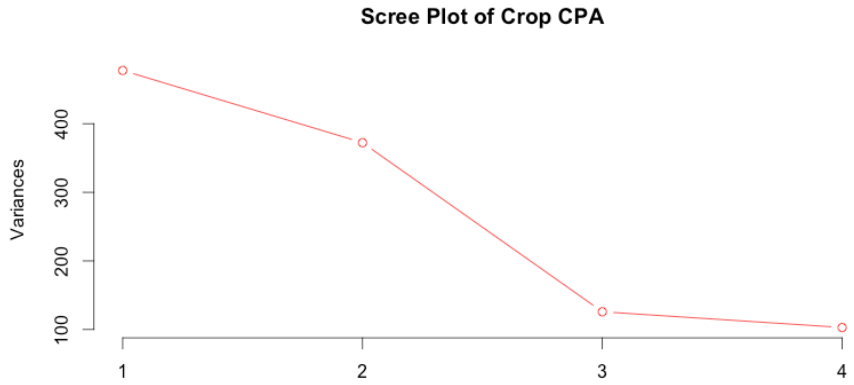
Summary of Height of Crops at Different Times:

```
> summary(crop[-1])
       X1              X2              X3              X4
 Min.   :12.00   Min.   : 8.00   Min.   : 2.00   Min.   :11.00
 1st Qu.:19.50   1st Qu.:23.00   1st Qu.:23.00   1st Qu.:24.00
 Median :26.00   Median :26.50   Median :25.50   Median :32.00
 Mean   :31.56   Mean   :29.69   Mean   :28.86   Mean   :35.86
 3rd Qu.:34.00   3rd Qu.:32.75   3rd Qu.:31.25   3rd Qu.:46.75
 Max.   :96.00   Max.   :58.00   Max.   :75.00   Max.   :78.00
```

From the summary of heights, the mean and the median of height at different times are close. Therefore, there is no reason to believe that the data is skewed. Then, we determine the number of principal components that can explain the variations of the data.

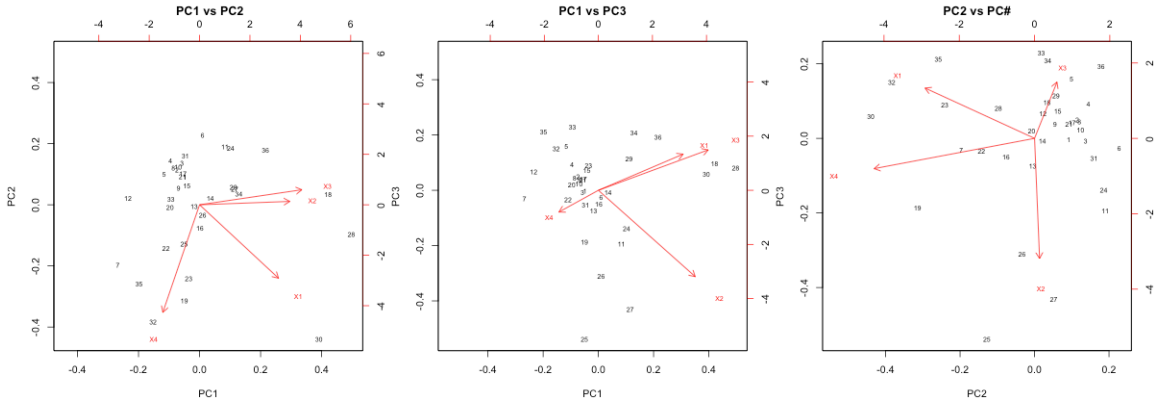
```
> pca = prcomp(crop[,-1])
> summary(pca)
Importance of components:
                          PC1    PC2    PC3    PC4
Standard deviation     1.339 1.0827 0.8020 0.6264
Proportion of Variance 0.448 0.2931 0.1608 0.0981
Cumulative Proportion  0.448 0.7411 0.9019 1.0000

> screeplot(pca, type="lines",col=2, ,main = "Scree Plot of Crop CPA")
```



Based on the cumulative proportion of variance and scree plot, we can see the first three components explained about 90.2% of the variability. The rotations of first three components are the following:

```
> round(pca$rotation[,1:3],2)
     PC1   PC2   PC3
X1  0.49 -0.56  0.35
X2  0.56  0.03 -0.83
X3  0.63  0.11  0.39
X4 -0.23 -0.82 -0.21
```



From the rotation scores, we can see that the first component explained 44.8% of variability in the data. It is mostly measure the weight of heights at time 1 and 3. The second component explained 29.3% of the variability in the data. It mostly weighted heights at time 1 and 4. The third component explained 16.1% of the variability in the data. It mostly weights the heights at time 2. From the biplot, PC1 vs PC2, we can see that X1 and X2 are strongly associated. And from PC1 vs PC3, we see that X1 and X3 are strong associated. All three components accord for 90.2% of the variations in the data. Therefore, three components are enough to explain the variability in the data.

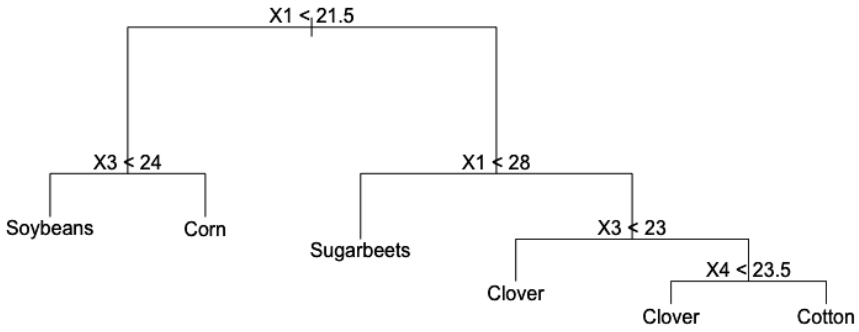
Classification Analysis

```
> round(tapply(X1, CROP, mean),2)
    Clover       Corn     Cotton   Soybeans Sugarbeets
     46.36      15.29      34.50      21.00      31.00
> round(tapply(X2, CROP, mean),2)
    Clover       Corn     Cotton   Soybeans Sugarbeets
     32.64      22.71      32.67      27.00      32.17
> round(tapply(X3, CROP, mean),2)
    Clover       Corn     Cotton   Soybeans Sugarbeets
     34.18      27.43      35.00      23.50      20.00
> round(tapply(X4, CROP, mean),2)
    Clover       Corn     Cotton   Soybeans Sugarbeets
     36.64      33.14      39.17      29.67      40.50
```

From the height of each times, we can observed that the average height of each crops changes over times. Therefore, we want to classify crops according to their height at different times using classification tree.

```
> tree = tree(CROP~X1+X2+X3+X4, CROP,mincut = 2, minsize =10, mindev = 0.01)
> summary(tree)

Classification tree:
tree(formula = CROP ~ X1 + X2 + X3 + X4, data = CROP, mincut = 2, minsize = 10,
mindev = 0.01)
Variables actually used in tree construction:
[1] "X1" "X3" "X4"
Number of terminal nodes:  6
Residual mean deviance:  1.647 = 49.42 / 30
Misclassification error rate: 0.3333 = 12 / 36
```



From the classfication tree, we can see that five crops can be classify based on height measured at time 1, 3, and 4. The rate of misclassificant is relative high(.33), but since we have a small dataset, we considered as acceptable.

Soybeans are classified based on height less than 21.5 at time 1 and less than 24 at time 3.

Corn are classified based on height less than 21.5 at time1 and height greater than 24 at time 3.

Sugar beets are classfied based on the height between 21.5 and 28 at time 1.

Clovers are classified based on the height greater than 28 at time 1 and less than 23.3 at time 4.

Cottons are classified based on the height greater than 28 at time 1 and greater than 23.5 at time 4.

Since, crops can be classified according to their height at different times. We want to find the type of the crop given the measurent of height at different times using linear and quadurtic discrimination analysis. Then we determine which model is more accurate.

Discrimation Analysis

For this analysis, we separate the data into two sets, one train and one test. We preform the discrimation on the train dataset and predict how many classifications are correct, then we use the same linear discrimation model for the test dataset and exam the number of correct classifications.

```
> samp = 1:nrow(crop)
> test = sample(samp, 16)
> train = samp[-test]
>
> lda = lda(crop[train, -1], crop[train,1])
> train.pred = predict(lda, crop[train, -1])
> table(crop[train,1], train.pred$class)

             Clover Corn Cotton Soybeans Sugarbeets
  Clover          3    0      1        1          1
  Corn            0    5      0        0          0
  Cotton          1    0      1        1          0
  Soybeans        0    2      0        2          0
  Sugarbeets      1    0      0        1          0

> test.pred = predict(lda, crop[test,-1])
> table(crop[test,1], test.pred$class)

             Clover Corn Cotton Soybeans Sugarbeets
  Clover          0    0      3        1          1
  Corn            1    1      0        0          0
  Cotton          2    0      0        1          0
  Soybeans        1    1      0        0          0
  Sugarbeets      2    0      0        1          1

> qda = qda(as.matrix(crop[-1]), CROP)
> table(crop[train,1], qtrain.pred$class)

             Clover Corn Cotton Soybeans Sugarbeets
  Clover          4    0      0        0          2
  Corn            0    5      0        0          0
  Cotton          0    0      3        0          0
  Soybeans        0    0      0        4          0
  Sugarbeets      0    0      0        0          2

> qtest.pred = predict(qda, crop[test,-1])
> table(crop[test,1], qtest.pred$class)

             Clover Corn Cotton Soybeans Sugarbeets
  Clover          5    0      0        0          0
  Corn            0    2      0        0          0
  Cotton          0    0      3        0          0
  Soybeans        0    0      0        2          0
  Sugarbeets      0    0      1        1          2
```

For the linear discrimination, the number of misclassification is 8 out 20 for train data and 14 out of 16 for the test data.

For the quadratic discrimination, the number of misclassification is 2 out of 20 for train data and 2 out of 16 for test data.

Conclusion

Based on the misclassification rate, linear discrimination model misclassified 40% of the train data and 87.5% of the test data and quadratic discrimination model misclassified only 10% of the train data and 12.5% of the test data. Therefore, we can conclude that quadratic discrimination analysis is better in classifying crops. This might be that that Quadratic Discrimination function does not assume homogeneity of variance-covariance matrices.