Liya Li

Course: STA 135

Instructor: Xiaodong Li

Project Report

1. Introduction

In this project, the built-in R dataset called *iris* is used. The overall goal of the analysis is to classify iris species along with some statistical testing and analyzing and see how well done it can be. In order to deeper understand the multi-class classification, Principal Component Analysis plays an important role in reducing dimensionality. Other than that, two Machine Learning models, Linear Discriminant Analysis and Decision Tree are used as classifier for further iris species prediction.

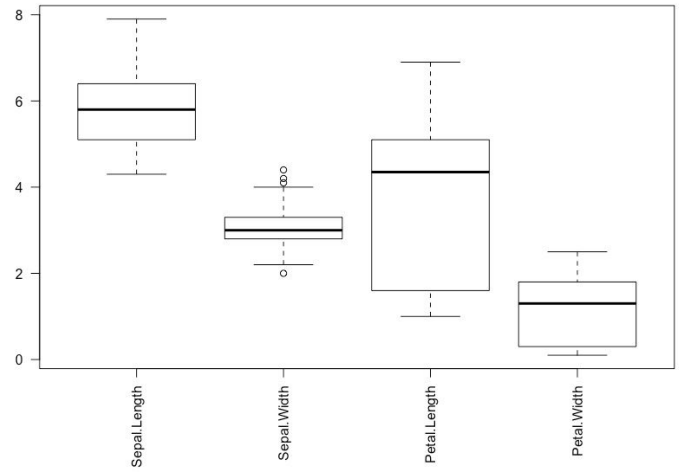
2. Summary

The iris dataset contains 150 observations of 5 variables. There are 3 classes in Species(setosa, versicolor and virginica). Take a first glance at the iris dataset: there are 4 numeric variables(Sepal.Length, Sepal.Width, Petal.Length, Petal.Width) and 1 factor variable(Species).

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa

Obtain the statistical summary and boxplot for the whole dataset in terms of variables:

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
Min. :4.300	Min. :2.000	Min. :1.000	Min. :0.100	setosa :50
1st Qu.:5.100	1st Qu.:2.800	1st Qu.:1.600	1st Qu.:0.300	versicolor:50
Median :5.800	Median :3.000	Median :4.350	Median :1.300	virginica :50
Mean :5.843	Mean :3.057	Mean :3.758	Mean :1.199	
3rd Qu.:6.400	3rd Qu.:3.300	3rd Qu.:5.100	3rd Qu.:1.800	
Max. :7.900	Max. :4.400	Max. :6.900	Max. :2.500	

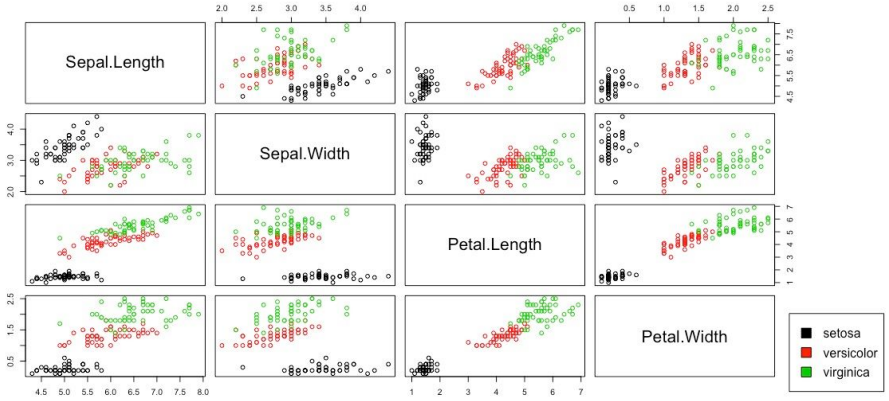


The numeric statistics tells that each species class contains 50 observations, and, the variable sample means are unequal which make sense in real life, yet a statistical testing on population mean differences will be used later. More intuitive comparison on the sample means is shown in the boxplot, where in Sepal.Width, there are outliers detected. The boxplot gives a rough estimate of the distribution of the values for each variable. Petal.Width seems having lower values than the other three iris measurement variables. Let's see the correlations among all four of them:

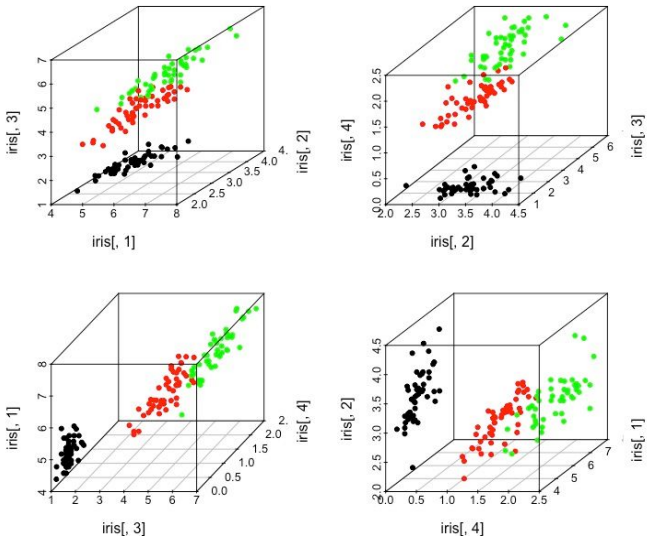
	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Sepal.Length	1.000	-0.118	0.872	0.818
Sepal.Width	-0.118	1.000	-0.428	-0.366
Petal.Length	0.872	-0.428	1.000	0.963
Petal.Width	0.818	-0.366	0.963	1.000

From here, Petal width and length are highly correlated because their correlation coefficient is 0.963 which means 92.16% of the variation is related, while the “inverse correlation” between Sepal width and length shows that they are not so correlated, only 1.3924% of the variation is related.

Now, visualize the dataset with pairs scatterplot and 3d scatterplot:



As shown in the pair scatterplot above, the setosa observations for each pairwise iris measurement visualization are quite separated from the other two species. However, between versicolor and virginica, there seems no obvious separation occurring in any sub-plot, which means that the classification for these two species could be challenging. Try visualize the iris data in 3D scatterplot:



The 3D scatterplot also clearly shows that setosa is well separated to both versicolor and virginica while the latter two species have a blurry boundary in separation.

3. Analysis

- Two-Sample Hotelling's T^2 Test

The null hypothesis is that for versicolor and virginica, the population means of each of the 4 measurement variables are equal at the alpha level 0.05 , and with R, the Two-Sample Hotelling's T-squared test result is as following:

```
        Hotelling's two sample T2-test

data:  Versicolor and Virginica
T.2 = 86.148, df1 = 4, df2 = 95, p-value < 2.2e-16
alternative hypothesis: true location difference is not equal to c(0,0,0,0)
```

Using the HotellingT2 function in R, the F-statistics is shown as T.2 in here because this test uses F transformation of computed T2, so the F-statistics is 86.148. And comparing the

p-value(2.2e-16) with alpha(0.05) which is larger, the null hypothesis is rejected at level of 0.05. Another approach with the traditional way to calculate Hotelling's T-squared statistics with the pooled variance and sample means returns that t2=355.4721 which is compared to the critical value c=10.18166, the null hypothesis is also rejected at level of 0.05.

- Simultaneous Confidence Interval

Compute the simultaneous confidence interval based on Hotelling's T-squared and Bonferroni correction in R and the results are shown respectively below:

```
95% simultaneous confidence interval
> Cis
                    [,1]          [,2]
Sepal.Length -1.0215837 -0.2824162674
Sepal.Width  -0.4070525 -0.0009475142
Petal.Length -1.6190928 -0.9649072114
Petal.Width  -0.8527216 -0.5472783821
```

```
95% Bonferroni simultaneous confidence interval
> Cis.b
                  [,1]        [,2]
Sepal.Length -0.9467342 -0.35726576
Sepal.Width  -0.3659295 -0.04207047
Petal.Length -1.5528487 -1.03115128
Petal.Width  -0.8217919 -0.57820814
```

Notice that for all 95% confidence intervals, 0 is not included in any interval which leads to the conclusion that the sample means are significantly different for each versicolor and virginica components pairs based on both methods.

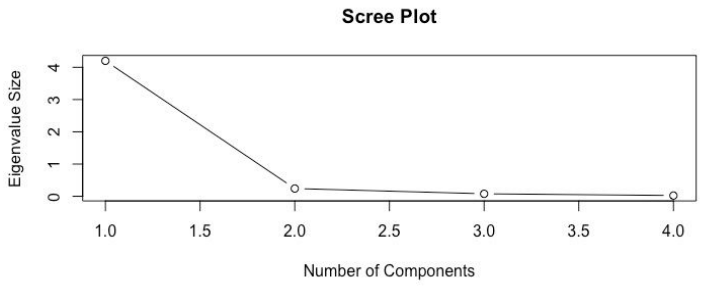
- Principal Component Analysis

The idea of PCA is very simple: reduce the number of variables of a dataset while preserve as much as information as possible. Apply Principal Component Analysis on the whole iris dataset to obtain its underlying structure, the coefficients of components and loadings are computed as the following:

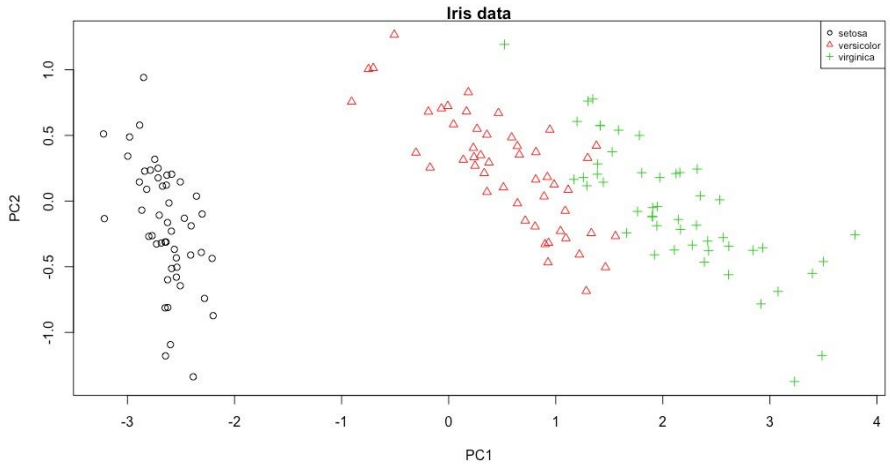
```
Importance of components:
                          Comp.1     Comp.2     Comp.3      Comp.4
Standard deviation     2.0494032 0.49097143 0.27872586 0.153870700
Proportion of Variance 0.9246187 0.05306648 0.01710261 0.005212184
Cumulative Proportion  0.9246187 0.97768521 0.99478782 1.000000000

Loadings:
             Comp.1 Comp.2 Comp.3 Comp.4
Sepal.Length  0.361 -0.657 -0.582  0.315
Sepal.Width         -0.730  0.598 -0.320
Petal.Length  0.857  0.173        -0.480
Petal.Width   0.358         0.546  0.754
```

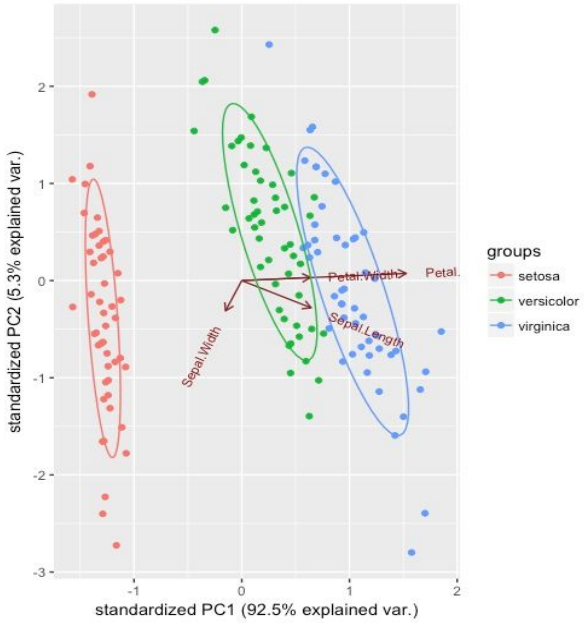
The PC's are constructed as linear combinations of the initial variables, and they are uncorrelated. Informations preserved from PC's are most into the first principal component, as shown here, the proportion of variance for Comp.1 is about 92.5%, while the rest 5.3% goes to Comp.2, and so on. The decision on number of PC to use is made based on the scree plot:



The PCA plot for all data point is as following and it reduced the dimensionality of the original dataset by only using the first two principal component values to represent the data (since the elbow indicates it's reasonable using the first two PC's and ignore the other PC's):



This data plot with PC1(x-axis) and PC2(y-axis) represents the whole iris data, and setosa data points are all away from versicolor and virginica data points is the same as found earlier. Then, with the standardized PC1 and PC2, the biplot can add information about the variables to the plot of the first two PC scores:



In the plot above, the arrows provide a graphical rendition of the loadings of each iris measurement variable on the used PC's. The use of two PC's is shown that Sepal.Width

concentrates on the negative site while Petal.Width, Sepal.length, Petal.length concentrate on the positive site.

- Linear Discriminant Analysis

LDA works similar to PCA, for it attempts to model the differences among classes. Apply LDA to the iris data, the results are:

```
Call:
lda(Species ~ ., data = iris)

Prior probabilities of groups:
    setosa versicolor  virginica
 0.3333333  0.3333333  0.3333333

Group means:
           Sepal.Length Sepal.Width Petal.Length Petal.Width
setosa            5.006       3.428        1.462       0.246
versicolor        5.936       2.770        4.260       1.326
virginica         6.588       2.974        5.552       2.026
           PredictSpeciesversicolor PredictSpeciesvirginica
setosa                         0.00                    0.00
versicolor                     0.96                    0.04
virginica                      0.02                    0.98

Coefficients of linear discriminants:
                                LD1         LD2
Sepal.Length             -0.9458072  0.38946040
Sepal.Width              -1.5216905  0.09771109
Petal.Length              2.3230198 -0.39023316
Petal.Width               3.0276701 -0.14115512
PredictSpeciesversicolor  0.8804986 -3.64698892
PredictSpeciesvirginica  -0.8804986  3.64698892

Proportion of trace:
   LD1    LD2
0.8177 0.1823
```

The output from lda function in R indicates the discriminant function, eg. the first discriminant function which is a linear combination of the variables is:

(-0.9458072) * Sepal.Length + (-1.5216905) * Sepal.Width + 2.3230198 * Petal.Length + 3.0276701 * Petal.Width + 0.8804986 * PredictSpeciesversicolor + (-0.8804986) * PredictSpeciesvirginica

And the proportion of trace is the percentage separation achieved by each discriminant function. So in here for the iris data, 81.77% separation is achieved by the first discriminant

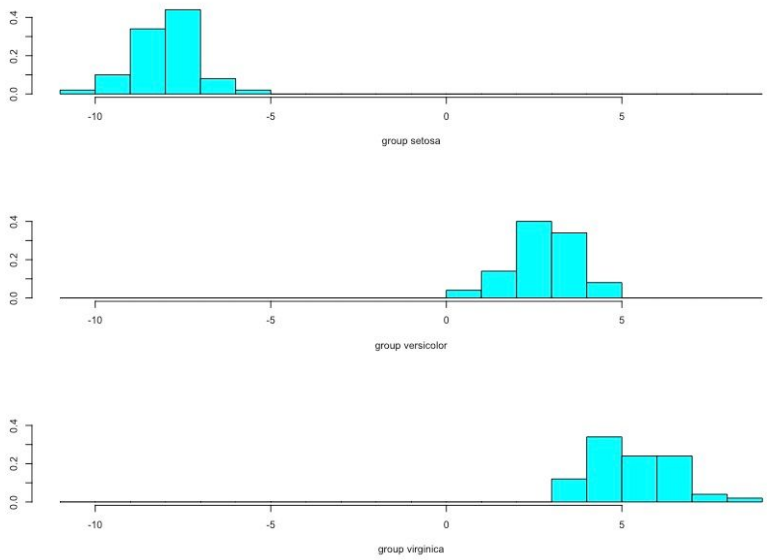
function, while the rest percentages is achieved by the second discriminant function. Using the LDA model to predict species, it's found that the predicted species in the iris data based on this model are not exactly the same as the original species record, where there are three observations getting a different prediction, they are the 71th, 84th, and 134th observation from the original iris dataset:

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species	PredictSpecies
71	5.9	3.2	4.8	1.8	versicolor	virginica
84	6.0	2.7	5.1	1.6	versicolor	virginica
134	6.3	2.8	5.1	1.5	virginica	versicolor

Thus, the LDA model works well to predict species with very high accuracy. And the confusion matrix shows that the predicted virginica and versicolor species counts are changed using LDA model: (original iris data shows 50 observations per species)

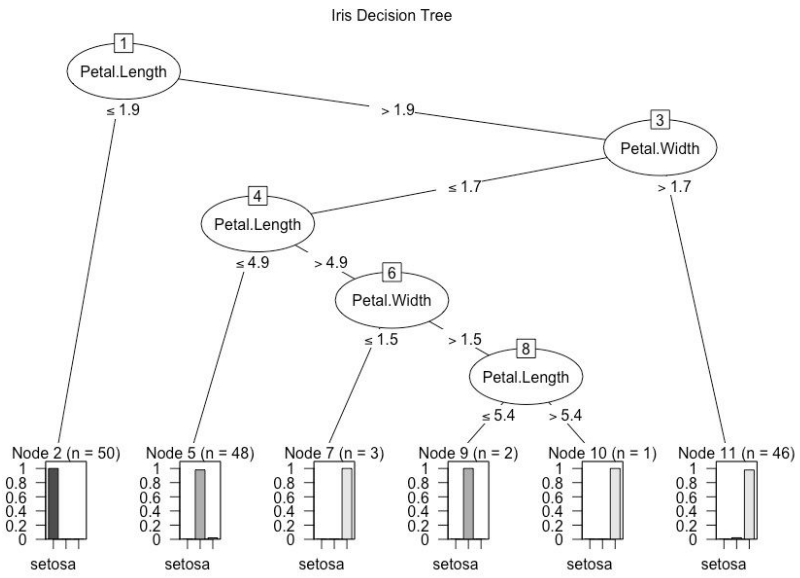
	setosa	versicolor	virginica
setosa	50	0	0
versicolor	0	48	2
virginica	0	1	49

The stacked histogram of the LDA values informs that setosa data are obviously classified from versicolor and virginica data, which is the same conclusion as mentioned before:



- Decision Tree

The decision tree model based on the original iris measurement variables is created for visualization and possible further prediction as well. As shown in the tree:



Each observation is with at least one condition during the classification and this could lead to accurate prediction for new observations. However, there might be overfitting in Decision Tree model because error may occur when the tree tries to perfectly fit all 150 iris data.

4. Interpretation

In the Two-Sample Hotelling's T-squared Test section, while comparing the population mean of respectively 4 measurements for versicolor and virginica at level of 0.05, the null

hypothesis that they are the same is rejected, meaning that the population mean of respectively 4 measurements for versicolor and virginica are significantly different.

In the Simultaneous Confidence Interval section, all 95% confidence intervals doesn't include 0, leading to the conclusion that with both methods, Hotelling's T-squared and Bonferroni correction: the population mean of respectively 4 measurements for versicolor and virginica are significantly different (same as from the earlier section). Also, notice that the boundaries of the confidence intervals are all negative, meaning that virginica has larger measurement population means than versicolor. Detail conclusion could follow as the example that: with Bonferroni correction, the Sepal.Length population mean for virginica is larger from 0.9467342 to 0.35726576 with 95% confidence.

In the Principal Component Analysis section, the PCA plot for all data points shows that the setosa data is easy to be classified from versicolor and virginica data. What's more, in the biplot, the arrow directions indicate that Petal.Width and Petal.Length are highly correlated while Sepal.Width and Sepal.Length are not; also, Sepal.Width is even less correlated to Petal measurements. There are also many small groups of points that are close to each other, representing that they have similar measurement values. The last but not least conclusion is that, the variability of Petal.Width and Petal.length across all species is mainly accounted by PC1, while as well as a large part of variability of Sepal.Length got accounted by PC1.

In the Linear Discriminant Analysis section, the LDA model predicts species with a high accuracy and as shown in the LDA value histograms, three species have different LDA distributions and again, setosa is located on the very left on the same range x-axis as versicolor

and virginica, meaning that setosa and the other species could have a highly separated boundary in between for classification, while it may be hard to create one between versicolor and virginica.

In the Decision Tree section, there shows a tree that conclude the whole iris data within one graph, showing the basic conditions for species prediction.

5. Conclusion

With all the exploration and analysis above, the first conclusion is that the classification for three iris species is not hard since setosa can be easily clearly classified from the other two, and between versicolor and virginica, we are 95% confident that the differences for their each measurement population mean comparison are significant. The second conclusion is that two PC's constructed from the petal and sepal values can successfully identify the three iris species, especially by using PC1, it's enough to classify setosa from versicolor and virginica data clearly. The last conclusion is that the LDA model and Decision model both work well for classifying the three iris species, yet Decision Tree might be overfitted. In order to predict a new observation of iris specie, one can use the LDA model with high accuracy.

Code Appendix

```
## STA135 Project - Liya Li
rm(list = ls())
library(devtools)
#remotes::install_github('vqv/ggbiplot', force = T)
library(ellipse)
library(scatterplot3d)
library(psych)
library(MASS)
library(ggbiplot)
library(ICSNP) # for Hotelling's T2
library(C50)

data(iris)
attach(iris)
head(iris)

######## summarize dataset
str(iris) # there are 5 variables in this dataset, dimension 5*150
table(iris$Species)
levels(iris$Species) # levels of the class attribute
# multi-class classification

# visualize data
# pairs scatterplot
pairs(iris[,1:4], col=iris[,5], oma=c(4,4,6,12))
par(xpd=TRUE)
legend(0.9,0.2, as.vector(unique(iris$Species)), fill=c(1,2,3))
# plot(iris) won't show colors

# 3d scatterplot
par(mfrow = c(2, 2))
mar0 <- c(2.5, 3, 2.5, 3)
scatterplot3d(iris[,1], iris[,2], iris[,3], mar = mar0, color = c("black","red", "green")[iris$Species], pch = 19)
scatterplot3d(iris[,2], iris[,3], iris[,4], mar = mar0, color = c("black","red", "green")[iris$Species], pch = 19)
scatterplot3d(iris[,3], iris[,4], iris[,1], mar = mar0, color = c("black","red", "green")[iris$Species], pch = 19)
scatterplot3d(iris[,4], iris[,1], iris[,2], mar = mar0, color = c("black","red", "green")[iris$Species], pch = 19)

#ggplot(iris, aes(x = Petal.Length, y = Sepal.Length, colour = Species)) +
#  geom_point() +
#  ggtitle('Iris Species by Petal and Sepal Length')
# setosa seems very different than the others in terms of length
# versicolor and virginia seems more familier at some points but still hard to classify
#ggplot(iris, aes(x = Petal.Width, y = Sepal.Width, colour = Species)) +
#  geom_point() +
```

```
#  ggtitle('Iris Species by Petal and Sepal Width')

# the summary statistics for iris dataset and boxplot for it
summary(iris)
par(mfrow = c(1, 1))
par(mar=c(7,4,1,1)) # more space for labeling
boxplot(iris[,-5],las=2) # rough estimate of the distribution of the values for each variable
corr <- cor(iris[,1:4]) # correlation
round(corr,3)
# eg: Petal width, length highly correlated bc correlation coeff is 0.963 -> 92.16% of the variation is
related
# eg: Sepal width, length not so correlated, -0.118, "inverse correlation"

######## analyze data
Setosa <- iris[iris$Species == "setosa",-5]
Versicolor <- iris[iris$Species == "versicolor",-5]
Virginica <- iris[iris$Species == "virginica", -5]

### Two-sample Hotelling's T2 test
### (partial code adopted from discussion_7.R and credited to Weiping Zhang(USTC))
# since versicolor and virginica are not easy to classify
# we are interested in testing whether they are significantly different
# start with testing whether their population mean flower measurements are the same
# that is, whether the average petal and sepal dimensions are the same
# assume they have the same variance matrix
alpha <- 0.05
HotellingsT2(Versicolor, Virginica)
# F-statistics: 86.148 (shown as T.2 because it's an F transformation of computed T2)
# pval < 2.2e-16 < 0.05: so reject null

# same test but look at test statistics
n <- c(nrow(Versicolor), nrow(Virginica))
p <- 4
versimean <- colMeans(Versicolor)
virgimean <- colMeans(Virginica)
d <- versimean - virgimean
S1 <- var(Versicolor)
S2 <- var(Virginica)
Sp <- ((n[1]-1)*S1+(n[2]-1)*S2)/(sum(n)-2)
t2 <- t(d)%*%solve(sum(1/n)*Sp)%*%d
t2   # 355.4721
cval <- (sum(n)-2)*p/(sum(n)-p-1)*qf(1-alpha,p,sum(n)-p-1)
cval # 10.18166
# T2 > critical value: so reject null
# Thus, the population mean flower measurements for versicolor and virginica are all not the same
# and we want to find the simultaneous C.I for them

### Simultaneous confidence intervals based on T2 and Bonferroni correction
```

```
### (partial code adopted from Discussion_7.R and credited to Weiping Zhang(USTC))
# simultaneous confidence intervals
wd <- sqrt(((n[1]+n[2]-2)*p/(n[1]+n[2]-p-1))*qf(1-alpha,p,n[1]+n[2]-p-1))*sqrt(diag(Sp)*sum(1/n))
Cis <- cbind(d-wd, d+wd)
cat("95% simultaneous confidence interval","\n")
Cis

#Bonferroni simultaneous confidence intervals
wd.b <- qt(1-alpha/(2*p),n[1]+n[2]-2) *sqrt(diag(Sp)*sum(1/n))
Cis.b <- cbind(d-wd.b, d+wd.b)
cat("95% Bonferroni simultaneous confidence interval","\n")
Cis.b

### Principal component analysis
### (partial code adopted from Discussion_9.R)
# then we apply PCA to reduce dimentionality of the orthogonal transformation on original data
# to convert the correlated variables into a set of values of linearly uncorrelated variables called PC.
flowers <- iris[,1:4]
classid <- iris[,5]
flowers.pca<- princomp(~., data=flowers)

plot(1:(length(flowers.pca$sdev)),  (flowers.pca$sdev)^2, type='b',
    main="Scree Plot", xlab="Number of Components", ylab="Eigenvalue Size")
summary(flowers.pca, loadings=TRUE)   # coefficient of components

plot(flowers.pca$scores[,1], flowers.pca$scores[,2], xlab="PC1", ylab="PC2", pch=rep(1:3,each=50),
col=classid, main="Iris data")
legend("topright", legend=levels(classid), pch=1:3, col=1:3, cex=0.7)
ggbiplot(flowers.pca, ellipse=TRUE, groups=classid)

### Linear discriminant analysis
### (code adopted from Discussion_10.R)
iris.lda <- lda(Species~.,data=iris)
iris.lda
iris.pred <- predict(iris.lda)
iris$PredictSpecies <- iris.pred$class
iris[iris$PredictSpecies!=iris$Species,] # predicted results that are diff to original

# Confusion matrix
table(iris$Species,iris.pred$class)

# A Stacked Histogram of the LDA Values
ldahist(data = iris.pred$x[,1], g=iris$Species)

### Another model
### Interesting step: create a decision tree to predict species
model <- C5.0(flowers, classid, control = C5.0Control(noGlobalPruning = TRUE, minCases=1))
plot(model, main="Iris Decision Tree") # but there might be overfitting using Decision Tree model
```