Principal Component Analysis in R

Arpan

January 4, 2017

Importing the library MASS for iris dataset

```
library(MASS,quietly = TRUE)
```

Storing the data set named "iris" into DataFrame named "DataFrame"

```
DataFrame <- iris
```

Type help("iris") to know about the data set

```
help("iris")
```

```
## starting httpd help server ...
```

```
##  done
```

Lets check out the structure of the data

```
str(DataFrame)
```

```
## 'data.frame':    150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

Check the dimension of this data frame

```
dim(DataFrame)
```

```
## [1] 150   5
```

Check first 3 rows

```
head(DataFrame,3)
```

```
##   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
```

Check the summary of data

```
summary(DataFrame)
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300
##  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
##        Species
##  setosa    :50
##  versicolor:50
##  virginica :50
##
##
##
```

Check the number of unique values

```
apply(DataFrame,2,function(x) length(unique(x)))
```

```
## Sepal.Length  Sepal.Width Petal.Length  Petal.Width      Species
##           35           23           43           22            3
```

Lets check the data set again

```
str(DataFrame)
```

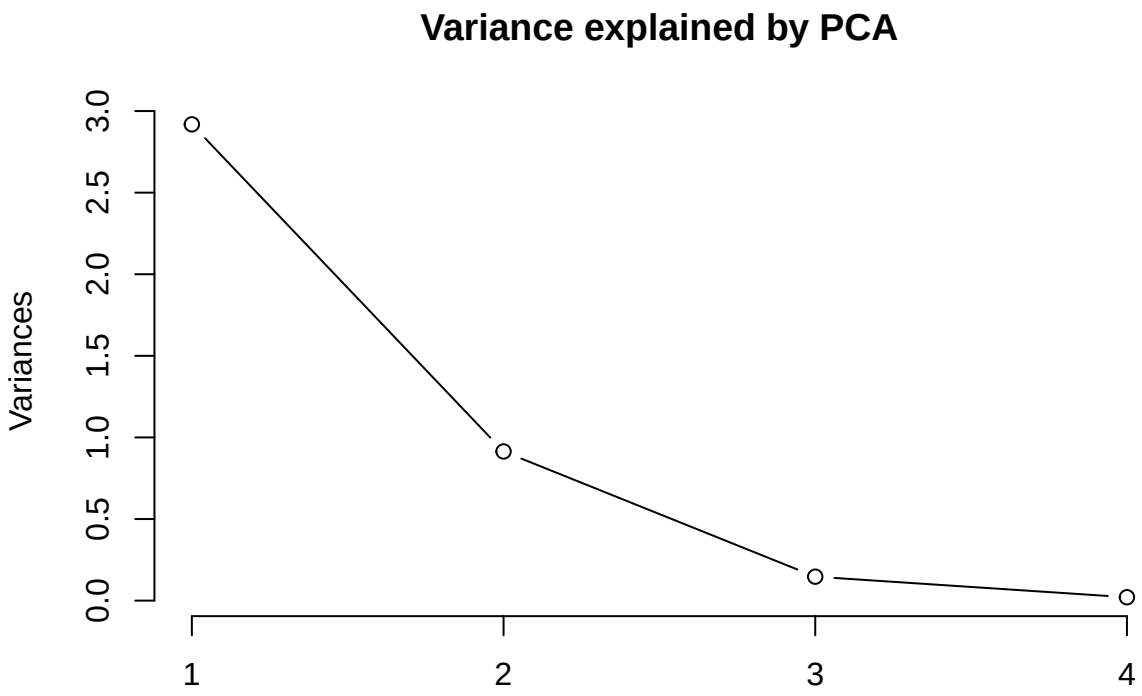
```
## 'data.frame':    150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

Let's do the principal companent analysis.Center=TRUE and scale=TRUE means we are scaling and centering the data before PCA.

```
modelPCA<-prcomp(x=DataFrame[,1:4],
                 center = TRUE,
                 scale. = TRUE)
```

Plot the variance explained by principal components

```
plot(modelPCA,type = "l",
     main="Variance explained by PCA"
     )
```



Let's find the variance explained by the first two Principal components(PC's)

```
sum(modelPCA$sdev[1:2]^2)/sum(modelPCA$sdev^2)
```

```
## [1] 0.9581321
```

We can see from above that only first two principal components alone can explain 95.8 % variance in the data

Let's check the complete summary of PCA.It shows the standard deviance and variance explained by each of the PCA components.Cumulative proportion of PC3 is 0.994 which means if we use first three components together in the data then these three components alltogether explains 99.4 % variablility in the data set.

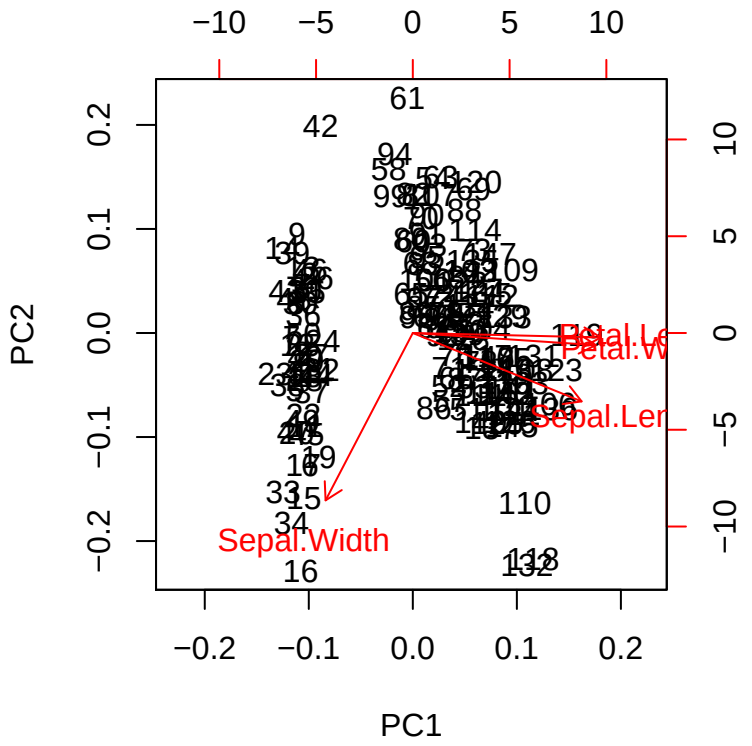
```
summary(modelPCA)
```

```
## 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
```

We can see from above that sum of proportion of variance explained by first two principal components is 95.8 % (0.7296+0.2285).

Let's plot the biplot showing first two PC's and the original feature vectors in this 2D space i.e original feature vectors as linear combination of first two PC's

```
biplot(modelPCA)
```

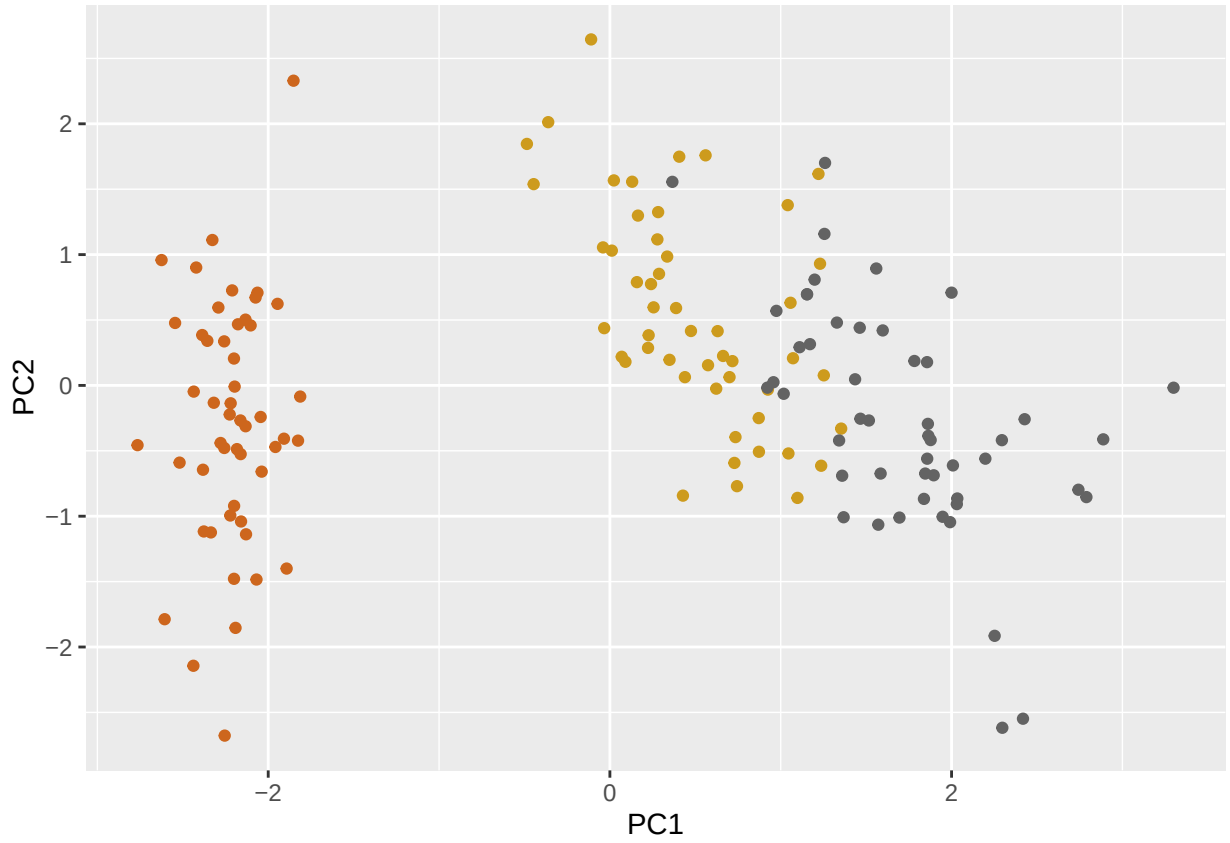


Let's try to do data visualization .We will use the principal component feature vectors instead of actual feature vectors like sepal-width,petal-width,etc.We will then color the data points based on Species variable.It is very easy to see that our PCA has worked!! Just based on two principal components we can see three clusters of setosa,versicolor,virginica in the data which are clearly separate.

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.2.5
```

```
ggplot(as.data.frame(modelPCA$x[,1:2]))+geom_point(aes(x=PC1,y=PC2),
```


2
1
0
−1
−2
PC2
−2
0
2
PC1