Dimensionality Reduction and Visualization

Md. Abdullah-Al Mamun

Exercise set 15 Solutions

Part I: Optional Problems

Optional Problem I1: Exploration of Your Own Data

The Principal Component Analysis (PCA) reduces the number of variables by finding a new, smaller set of variables.

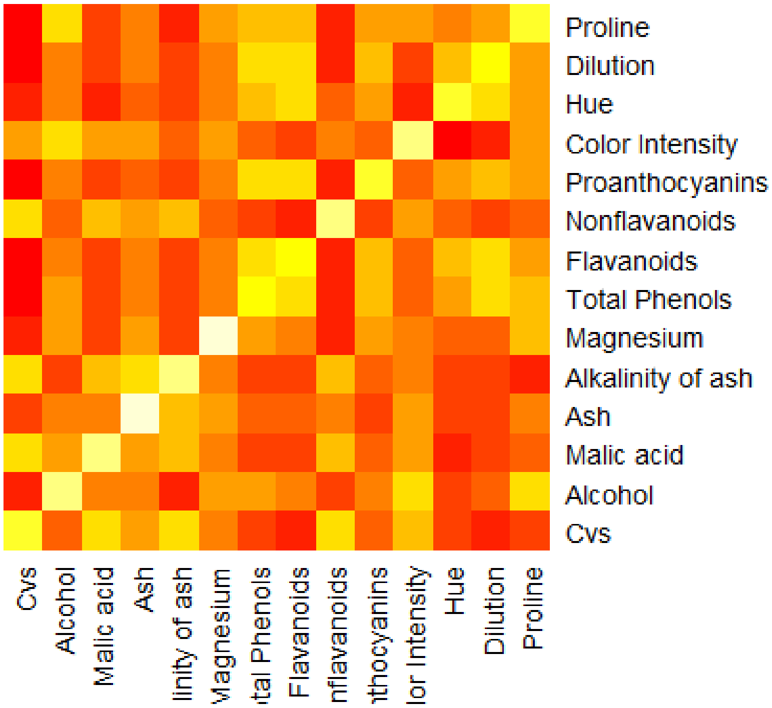
The analysis made use of wine dataset obtained from UC Irvine Machine Learning Repository.

Source: http://ar chive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data

```
str(wine)
'data.frame':  178 obs. of  14 variables:
 $ Cvs               : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Alcohol           : num  14.2 13.2 13.2 14.4 13.2 ...
 $ Malic acid        : num  1.71 1.78 2.36 1.95 2.59 1.76 1.87 2.15 1.64 1.35
...
 $ Ash               : num  2.43 2.14 2.67 2.5 2.87 2.45 2.45 2.61 2.17 2.27 .
..
 $ Alkalinity of ash: num  15.6 11.2 18.6 16.8 21 15.2 14.6 17.6 14 16 ...
 $ Magnesium         : int  127 100 101 113 118 112 96 121 97 98 ...
 $ Total Phenols     : num  2.8 2.65 2.8 3.85 2.8 3.27 2.5 2.6 2.8 2.98 ...
 $ Flavanoids        : num  3.06 2.76 3.24 3.49 2.69 3.39 2.52 2.51 2.98 3.15
...
 $ Nonflavanoids     : num  0.28 0.26 0.3 0.24 0.39 0.34 0.3 0.31 0.29 0.22 ..
.
 $ Proanthocyanins   : num  2.29 1.28 2.81 2.18 1.82 1.97 1.98 1.25 1.98 1.85
...
 $ Color Intensity   : num  5.64 4.38 5.68 7.8 4.32 6.75 5.25 5.05 5.2 7.22 ..
.
 $ Hue               : num  1.04 1.05 1.03 0.86 1.04 1.05 1.02 1.06 1.08 1.01
...
```

```
 $ Dilution         : num  3.92 3.4 3.17 3.45 2.93 2.85 3.58 3.58 2.85 3.55 .
..
 $ Proline          : int  1065 1050 1185 1480 735 1450 1290 1295 1045 1045 .
..
```

The heatmap plot for correlation is as given below:

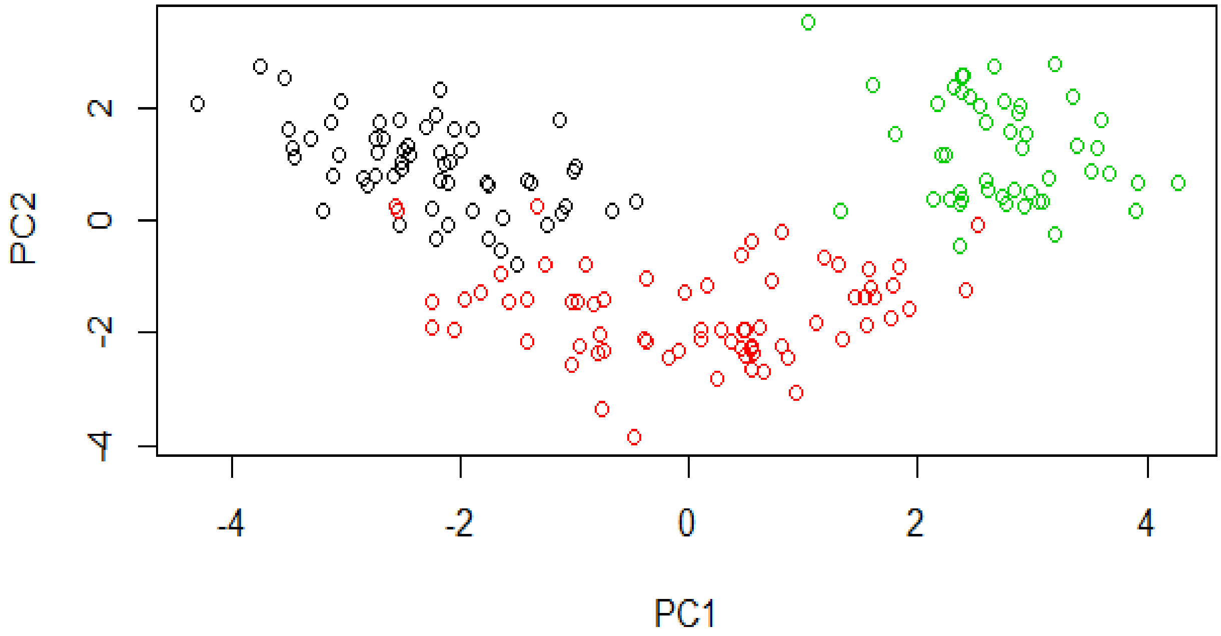


In the heatmap diagram above, dark heatmap signifies strong correlation among variables. Thus, it includes several strong correlations.

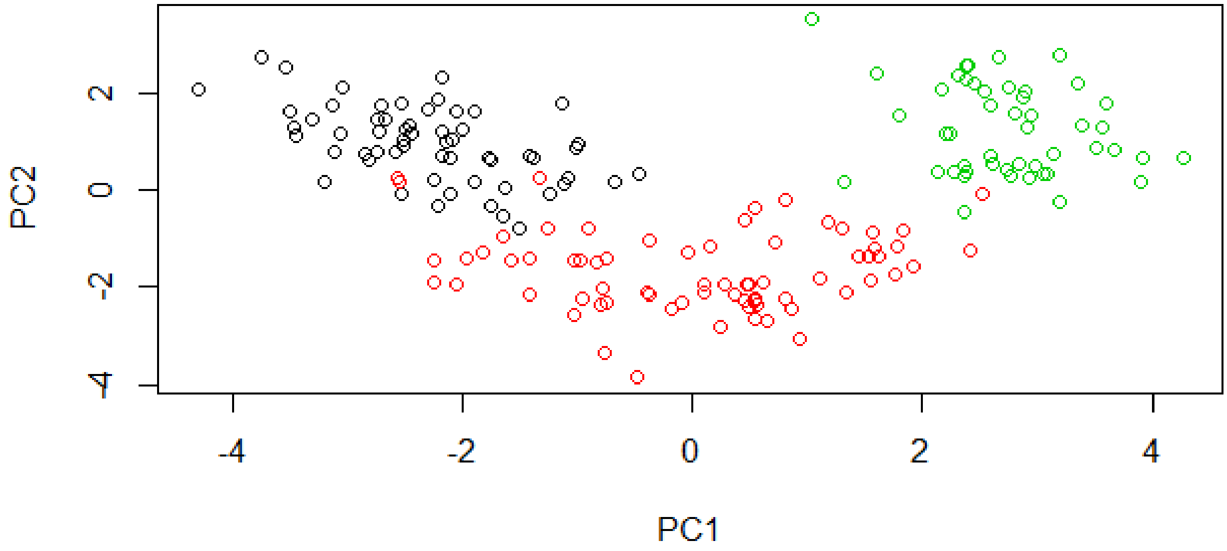
```
summary(winePCA)
Importance of components:
                         PC1    PC2    PC3     PC4     PC5     PC6     PC7
Standard deviation     2.169 1.5802 1.2025 0.95863 0.92370 0.80103 0.74231
Proportion of Variance 0.362 0.1921 0.1112 0.07069 0.06563 0.04936 0.04239
Cumulative Proportion  0.362 0.5541 0.6653 0.73599 0.80162 0.85098 0.89337
                           PC8     PC9   PC10    PC11    PC12    PC13
Standard deviation     0.59034 0.53748 0.5009 0.47517 0.41082 0.32152
Proportion of Variance 0.02681 0.02222 0.0193 0.01737 0.01298 0.00795
Cumulative Proportion  0.92018 0.94240 0.9617 0.97907 0.99205 1.00000
```

It can be concluded that applying the PCA technique reduces the number of variables to 7, i.e. PC1 to PC7 as shown in summary table above.

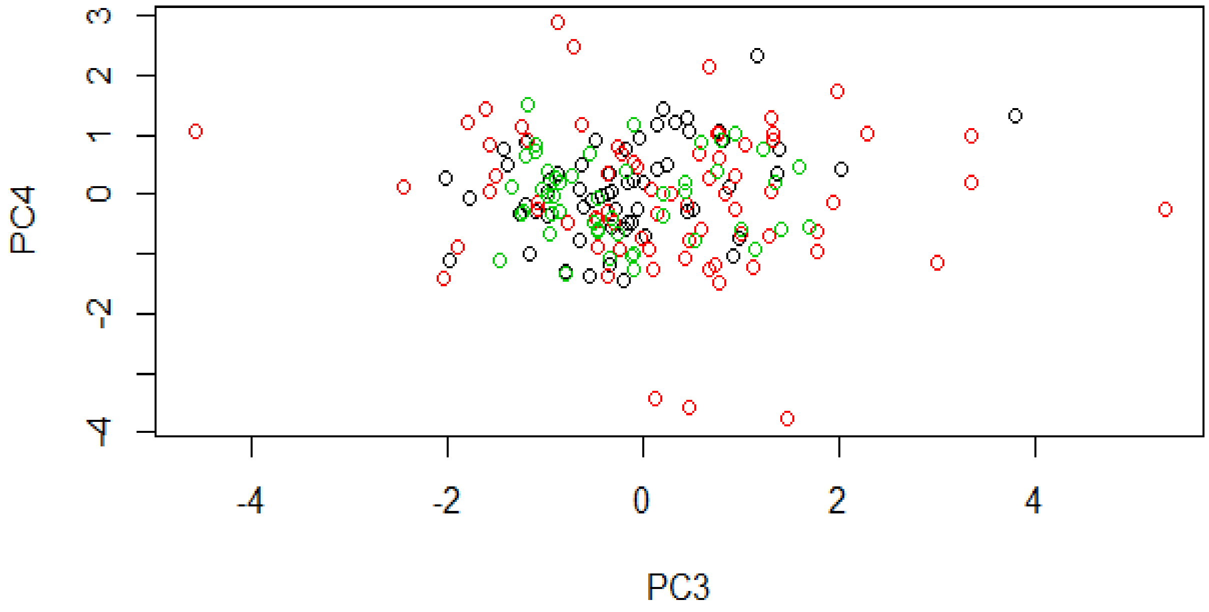
The PCA plot for 2 components, PC1 and PC2 are given below:



The PCA plot for all data is given below:



It can be seen that both plots above looks the same.



PC3 and PC4 look different from both PC1, PC2 and all PCA. In PC3 and PC4, most data points are stacked together especially at the middle of the plot, while for others they are spread with data points with white, black, and green not being overlapping.

Index

Optional I1:

In [ ]:

```r
library(doParallel)
registerDoParallel(cores = detectCores() - 1)
set.seed(10)
library(caret)
library(corrplot)

wine <- read.table("http://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data", sep=",", header=FALSE)
colnames(wine) <- c('Cvs', 'Alcohol', 'Malic acid', 'Ash',
                    'Alkalinity of ash', 'Magnesium', 'Total Phenols',
                    'Flavanoids', 'Nonflavanoids',
                    'Proanthocyanins', 'Color Intensity', 'Hue',
                    'Dilution', 'Proline')
str(wine)
heatmap(cor(wine),Rowv=NA, Colv=NA) #Plot heatmap to show correlation among variables
classes<- factor(wine$Cvs) # To declare classes as Cv1, Cv2, Cv3
winePCA <- prcomp(scale(wine[,-1])) # PCA; Normalize data and exclude the first column
summary(winePCA) # To get the summary statistics on PCA
plot(winePCA$x[,1:2], col= classes) # Plotting PCA1 and PCA2 only; 3 colors
plot(winePCA$x, col= classes) # Plotting PCA of all data with 3 colors
plot(winePCA$x[,3:4], col= classes) # Plotting PCA3 and PCA4 only; 3 colors
```