Question 1

The file project1.csv contains a small dataset of 40 observations, each with nine continuous variables X1,…,X9X1,…,X9 and one binary categorical variable ClassClass. Use the R function prcomp to perform PCA on X1,…,X9X1,…,X9, and answer the following questions:

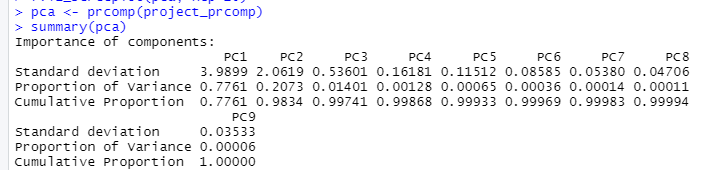
1. You can either use the original variables or use the standardised variables to perform PCA. For this dataset, which way you will choose? Numerically justify your choice.

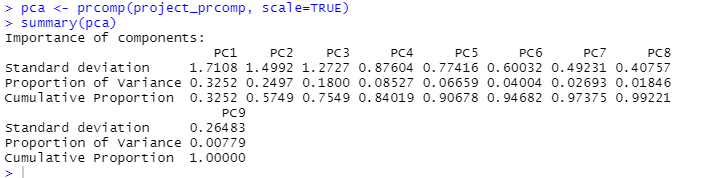
#Here we have used standardized the input data so that it has zero mean and variance one before doing PCA. Scale = True defines it.

pca <- prcomp(project\_prcomp, scale=TRUE)

summary(pca)

Without scale u can see standard deviation has higher weight for the calculation of axis than a variable with a low standard deviation.





How many principal components will you retain if you must reduce the dimension of this dataset? Numerically justify your choice.

#we can use sdev to compute variance explained by each Principal Component

#Visualizing the variance will help us identifying how many principal components are needed to explain the variation data

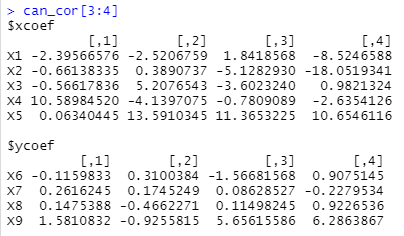
#Looking at the plot we can retain minimum 3 TO 4 and max 7 , as per requirement

Question 2

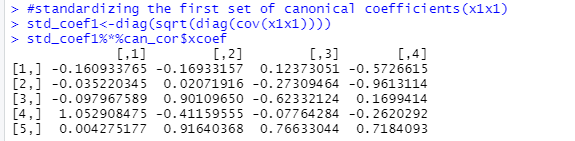
1. Will the canonical correlations change under non-singular linear transformations AX(1)AX(1) and BX(2)BX(2) of two sets of variables X(1)X(1) and X(2)X(2)? Prove your answer.

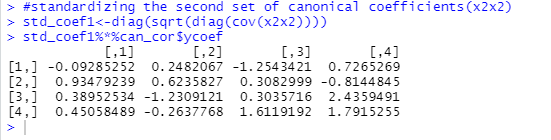
# CCA tries to project both datasets into a new space in a way that conserves their invariant quantities (i.e. those not affected by linear transformations). So it will not change.

1. Consider a set X(1)X(1) including five variables X1,…,X5X1,…,X5 from the dataset in the file project1.csv and another set X(2)X(2) including four other variables X6,…,X9X6,…,X9. Use the R function cc from the R package CCA to perform CCA on X(1)X(1) and X(2)X(2). You can either use the original variables or use the standardised variables to perform CCA. For this dataset, which way you will use? Numerically justify your choice.



#the standard deviations between the variables have a large variance between them i.e x1 to x5 and x6 to x9, we need to standardize





Question 3

Split the 40 observations in the file project1.csv into a training set, which contains the **first** 20 observations, and a test set, which contains the **remaining** 20 observations. Consider the continuous variables X1,…,X9X1,…,X9 as the features and consider the binary categorical variable ClassClass as the class label. Use the R function lda from the R package MASS to perform LDA, and qda for QDA.

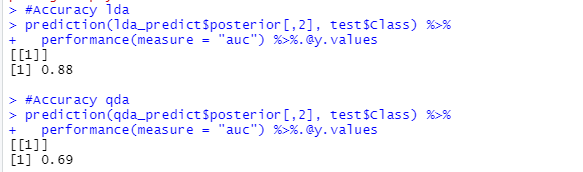
1. Between LDA and QDA, which one performs better on this test set in terms of classification accuracy?

#Prior probabilities show 66.67% of training observations are group with class A and rest with class B for LDA and 52.39% of training observations are group with class A and rest with class B for QDA

#Accuracy for LDA is 88% and QDA is 51% for 50-50 train and test

1. Explain why LDA performs better/worse than QDA in this case.

#Because QDA needs to be trained on less features, LDA can be used for dimensionality reduction but QDA cant



1. Propose a simple extension of LDA (or QDA) to further improve the classification accuracy on this test set. Numerically show the improvement in classification accuracy obtained by your proposed extension.

#Increasing the training dataset and reducing test will help us get more accurate QDA as shown in code