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

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.
2. 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.

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?
2. Explain why LDA performs better/worse than QDA in this case.
3. 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.