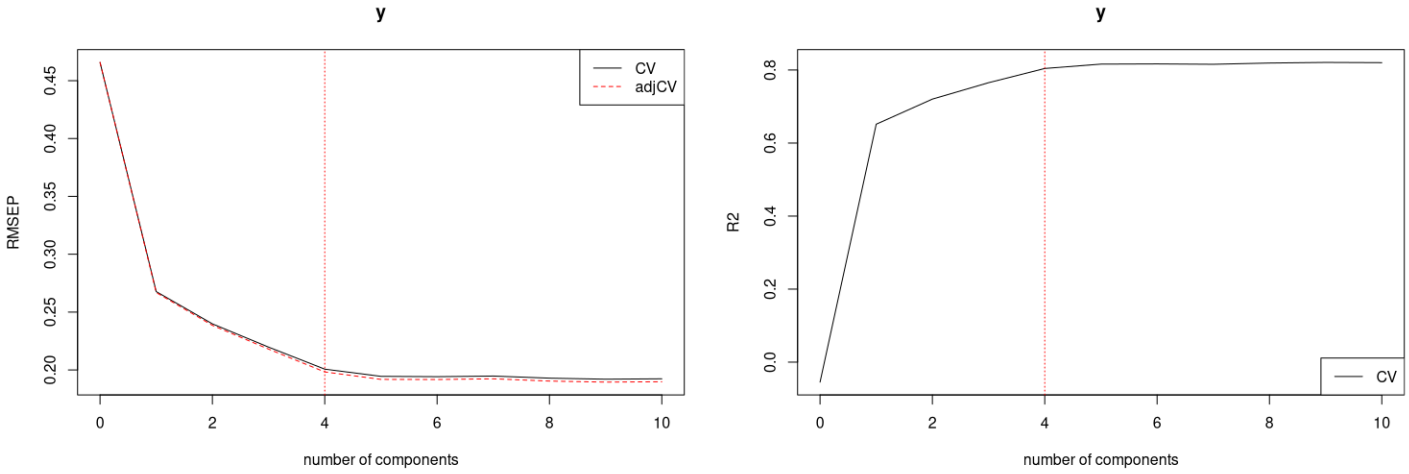
1 Introduction

In this work we apply the PLS1 reggression approach to gene expression monitoring by DNA microarrays to automatically differentiate between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). The dataset consists of 2 parts: training with 38 samples and testing with 34 samples. Each sample contains 7129 attributes. Additionally, for each split we are given true labels which determine the response variable for given sample (ALL or AML). The goal is to analyse this dataset using PLS1 regression (determine the best number of components) and finally predict the probability of AML leukemia in the test sample.

2 Experiments

To perform the Partial Least Squares Regression we use the *plsr* function from R's *pls* package on the training split of the dataset. With the help of *summary*, *RMSEP* and *R2* functions we can analyse the details of the trained model. Based on Figure 1, which shows the values of $RMSE$ and R^2 depending on the number of selected components, we choose to retain only 4 components. Then, we center the test data with respect to the mean of the training data and project the test data as supplementary individuals onto the chosen 4 PLSR components. The projection is presented in the Figure 2 with red color.

Finally, we use the 4 components, obtained by the PLSR model, to predict the types of leukemia for the testing dataset. We use logistic regression as the classification model. The total accuracy of the model in predicting the correct type of leukemia, calculated for the testing set, is 85.29%, and the probability of AML leukemia in the test sample is 55.88%.



(a) Root Mean Squared Error

(b) R^2

Figure 1: In order to select the number of components to retain after Partial Least Squares Regression, the values of $RMSE$ and R^2 metrics are analysed. In both cases it's clearly visible that the improvement stops after the adding 4th component. Hence, for further experiments and prediction only 4 PLSR components will be used.

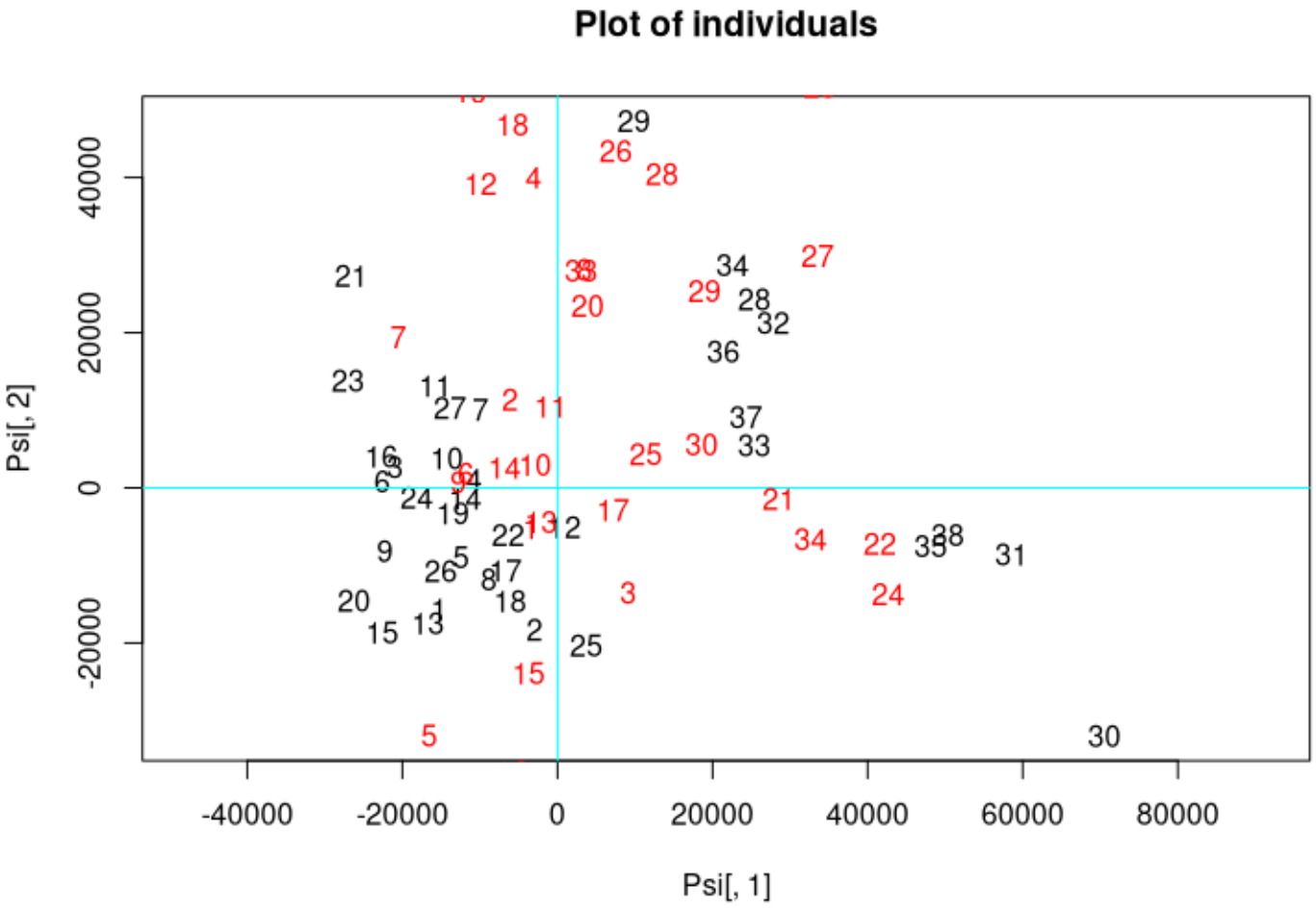


Figure 2: Test data, centered with respect to the mean of the training data, projected onto 4 selected PLSR components. Test data is shown in red color, whereas train data in black.