1 Introduction

As mentioned in the statement, the focus of this work is to find a classifier using PLS1 regression approach on gene expression monitoring by DNA microarrays, to automatically differentiate between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). Our training data has a total of 7129 observations from 78 attributes and our test set has 7129 observations and 70 attributes. Our targets are extracted from $table\_ALL\_AML\_predic.doc$

2 Partial Least Squares Regression

In order to first perform our PRLS, we needed to do some pre-processing of the data. More concretely, we had to form the matrices X and Xt which contain the gene expression for 38 training samples and 34 test samples respectively. Since our data was presented in transposed form and was not sorted, we had to transpose and sort them. Furthermore, we filtered all that was not a numeric value since it was not pertinent to the problem.

Once we have pre-processed our data, we performed our PLS1 regression on the training data. In order to select the number of variables, we perform cross-validation with LOO and we decide to take a component while the RSEMP decreases or R^2_{cv} increases.

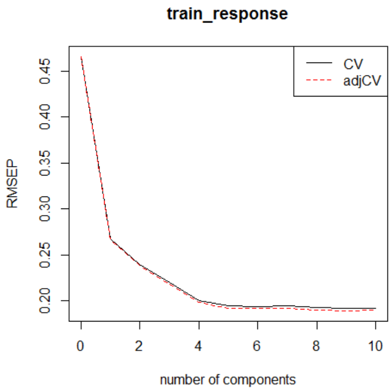


Figure 1: RSEMP

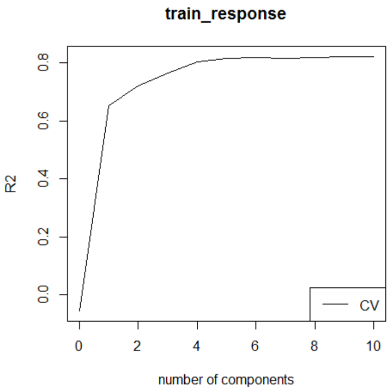
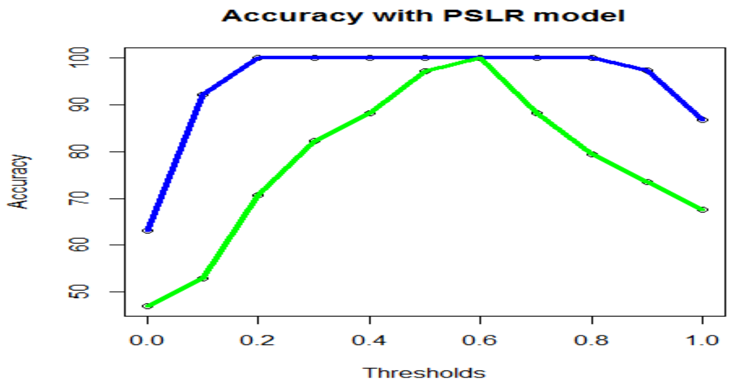


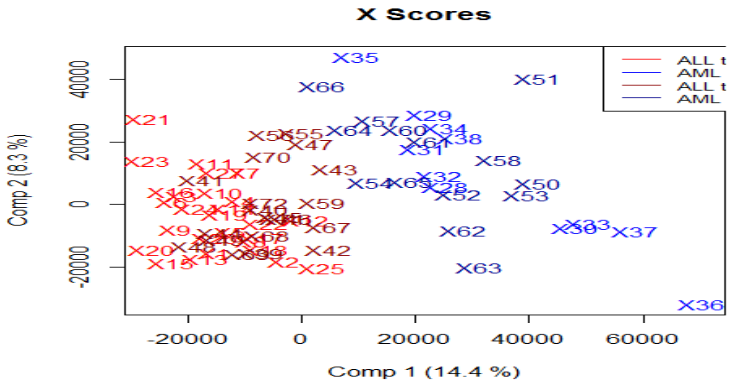
Figure 2: R^2_{cv}

In our particular case, we decided to take 4 components.

Furthermore, we have also analysed how with the number of components chosen, the threshold impacts in our prediction. We can see that we obtain the same performance on the train test from 0.2 - 0.8 whereas we obtain the highest accuracy in 0.6 in the test set. It is pretty estrange that we obtain a 100% of accuracy which could mean that we are somehow over-fitting. Specially if we consider that the test results are not so good.



Next we have projected the test data as supplementary individuals onto the selected PLS1 components and we have centered the data respect to the training set mean. We have performed a joint plot of the train and test individuals in the plane of the two first PLS1 components differentiating AML subjects with ALL subjects. We can see that while on the train set we have the components well separated, we do not have them clearly separated on the test set. They are pretty close to each other.



We have performed a logistic regression in order to predict the response in the training data only using the PLS1 selected components. In the same fashion as when we first calculated our accuracy with the PLS, we have now done the same for logistic regression where we see that our results are way more stable across thresholds with a similar peak results.

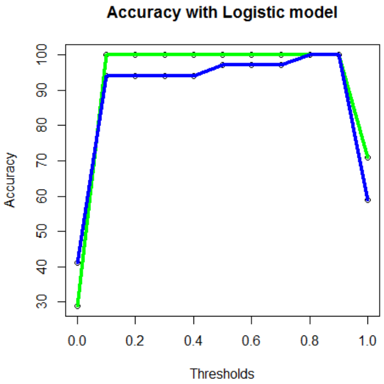


Figure 3: Accuracy logistic regression

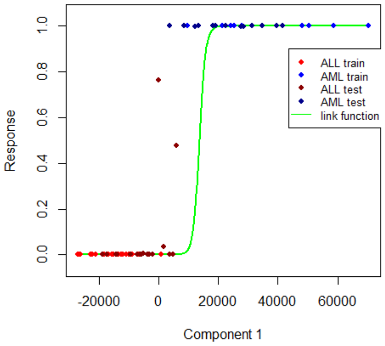


Figure 4: Function its values

Finally we have predicted the probability of AML leukemia in the test sample and we have plotted the results with the training results and with the logit function for the case of threshold=0.5 and 4 components. We can see that even it could previously look like over-fitting, the data itself is distributed in a fashion that allows this precision.