1 Introduction

We are going to use the ZIP Data of hardwritten digits where we applied MVR, PCR and IBA techniques. In this delivery we are going to use PLSR2 as a component based methodology to predict the digits. As in the previous deliveries, we are going to select a 5% of random sample from the training set and we are going to use all our test set. We are going to use the same sample that we have used before. Furthermore, we will also define a response matrix, a predictor matrix and we are only going to center our predictor matrix, not scale, we are going to perform a one hot enconding for the Y matrix since we are treating this problem as a classification problem. Overall, everything we did in the aforementioned deliveries.

Once we have repeated those steps, we are going to apply PLS2 to our data using cross-validation. We decided to retain 17 components which is the number of components that give us the highest R^2. It gives us an R^2 of $\approx 67\%$ in the training set and a $\approx 48\%$ in the test set.

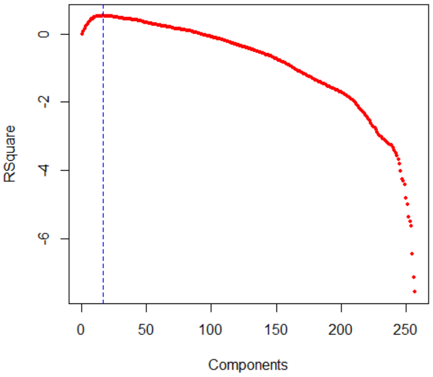

Figure 1: Number of components vs R^2

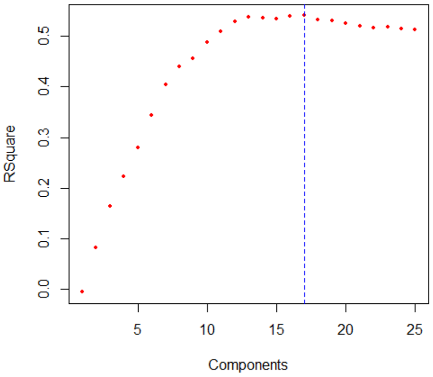

Figure 2: Components [0:25]

If we assign every test individual to the maximum response and compute the error rate we obtain an accuracy of $\approx 82\%$ with an average R^2 of 70%. We can see in the following plots, how the scores from the first two components, show in train vs test set. We can see that in the train scores, we have a way more clear distribution whereas in test set we do not have a clear pattern which is quite normal because the training set is the one we have performed the initial model and it will always perform better and also has a smaller size. Still I find curious that the distribution on the right hand side, is so unclear.

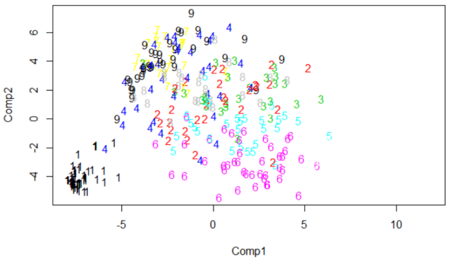

Figure 3: Spread of train set

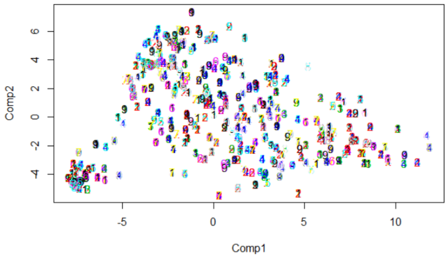

Figure 4: Spread of test set

This results, if compared with the previous methods that we used to classify our digits, we can see that is the has been able to explain our data the best while giving a similar accuracy as PCR. Therefore, we can say that PLSR2 is the one that with the few data that we work with, has been able to explain the model with some consistency while having a decent accuracy which is great, because we have seen that all the other models had a very poor R^2.

Method	**Accuracy**	R^2
MVR	$\approx$ 60%	-0.35
PCR	$\approx$ 83%	0.48
IBA	$\approx$ 77%	0.40
PLS2	$\approx$ 82%	0.70

Table 1: Comparison performances between MVR, PCR, IBA and PLS2