1 Introduction

This work is the continuation of the Zip Practical project we performed before where we performed a multivariate regression and a principal component regression to the Zip data. In this case, we approach this problem in a different way and we are going to perform the Inter Batteries Analysis and decide how many components are we going to retain, finally we are going to compare our R^2 results with the ones obtained in the previous assignment.

In order to be able to compare them with fairness, we are going to be performing the first two steps as we did in the previous assignment. These two steps where mainly selecting a random sample (5%) from train and define two matrices, the response matrix (Y) which corresponds to the first column of our dataset, and the predictor matrix (X) which corresponds to the remaining columns and we are also going to center. Note also that since our problem is going to be treated as a classification problem, we are going to perform a one hot encoding and create a column per possible digit $[0, 9]$ (note that we will do the same to the test set).

1.1 Inter Batteries Analysis

The goal of IBA is to measure the relationship between multivariate vectors by components derived from the original variables. IBA is a compromise between CCA and PCA. This is done by computing the covariance between the targets and predictors, followed by computing the eigen values (aib). A is calculated as :

$$A < -\text{Vxy \%*\% aib\$vectors \%*\% diag(aib\$values}\hat{}\text{(-0.5))}$$

B is not needed for our computation. The inner product of A and the predictor matrix, gives us T and again, we do not need U.

Once implemented the batteries analysis we can see the following R^2 error. We can see that it does not give a good proportion explained by the components therefore, we are going to use all the components for our prediction.

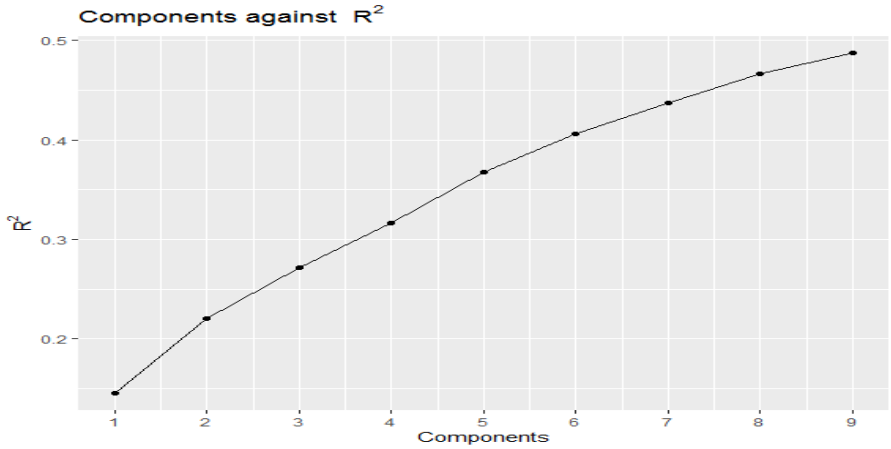

Figure 1: Number of components against R^2

1.1.1 Modelling

Now that we know that we are going to use all 9 components, we are going to train our model with the training data and see the results with train and test sets. In the training set, we see that we have an accuracy of ≈ 87. We also analysed which numbers are those that are confusing our model and we found out that the following accuracies:

nº0	**nº1**	**nº2**	**nº3**	**nº4**	**nº5**	**nº6**	**nº7**	**nº8**	**nº9**
90	100	83.78	96.96	72.72	78.57	79.41	78.78	85.71	87.87

As we can see, our system struggles more with the 4's,5's,6's 7's. However, since there is an imbalance in the data, this results might be biased. What we can see, is that if we do a confusion matrix for instance our system tends to confuse 9's with 4's, 0's with 8's, etc. That if we inspect, we can see that are not this well done numbers.

Regarding to the test, we have a total of $\approx 77.6\%$ accuracy with an R^2 of $\approx 40.5\%$. The individual accuracies in this case, are the following:

nº0	**nº1**	**nº2**	**nº3**	**nº4**	**nº5**	**nº6**	**nº7**	**nº8**	**nº9**
86.35	98.1	67.17	75.90	69.5	60.62	77.64	70.74	71.08	79.09

So we can see that our model is performing worse in the test set and that

again, has a low R^2 as we would expect since with 9 components we were not able to explain more than a 50%.

1.1.2 Comparison with previous models

If we compare the IBA with the previous activity results, we can see that we are actually having that PCR shows the best results among the three methods with an accuracy close to 85%. In this particular case, the IBA analysis does not seem to be the best model, however it might be because we are training with too few data.

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

Table 1: Different roles cost considering project's deviation