Mikołaj Małkiński
mikolaj.malkinski@est.fib.upc.edu

HOMEWORK PLS 2

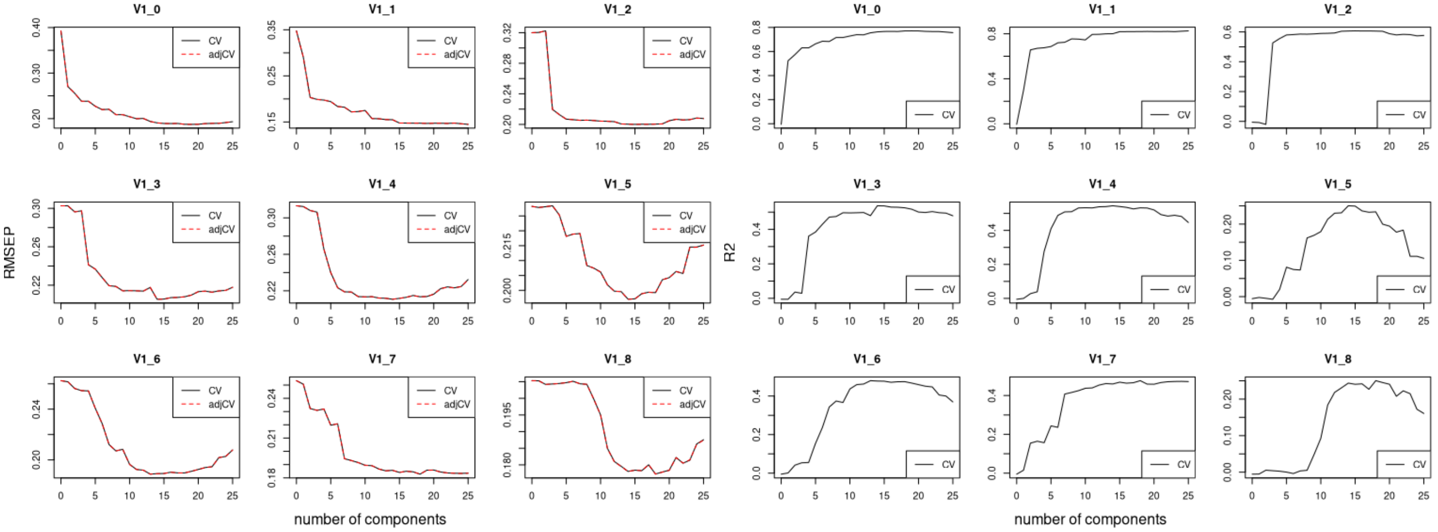
KMLMM
December 15, 2019

1 Introduction

In this work we consider a dataset of normalised handwritten digits, which were automatically scanned from envelopes by the U.S. Postal Service in 16 x 16 grayscale images. The total training set contains 7291 digits, however we will perform the analysis on a 5% subset in order to assess how well tested methods perform without access to large quantities of data. The testing set has 2007 digits. The main method chosen for this work is the Partial Least Squares Regression 2 (PLS2). Similarly as in PCR and IBA assignments, we are dealing with a dataset of size n with $X = (x_1, \dots, x_p)$ and $Y = (y_1, \dots, y_q)$. The goal is to classify images of handwritten digits from the test set, compute the error rate and compare results with previous works.

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. Firstly, we choose to create 25 components and use Leave-One-Out cross-validation method. Then, 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 15 components. Then, we center the test data with respect to the mean of the training data, perform regression using selected 15 PLSR components and finally predict the class based on the max response. The value of R^2_{mean} (R^2 averaged across all classes) is 0.599 and the total error rate of the model in predicting the correct handwritten digit, calculated for the testing set, is 19.38%.



(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 visible that the improvement stops after adding around the 15th component. Hence, for further experiments and prediction only 15 PLSR components will be used.

3 Discussion

In this work an application of Partial Least Squares Regression was presented for the dataset of handwritten digits. When comparing the results obtained by this work to previous assignments, we can see that PLSR2 (error rate of 19.38%) performed better than IBA (error rate of 52.86%) and PCA (error rate of 38.76%). Additionally, PLSR2 required to select more components (15) than in IBA (9).