KMLMM course.

Exercise 4: ZIP Practical work –PLSR2

2019-2020 course
Prof. Tomàs Aluja

We have normalized handwritten digits, automatically scanned from envelopes by the U.S. Postal Service in 16 x 16 grayscale images (from -1 to 1). Each line consists of the id (0-9) followed by the 256 grayscale values. We dispose of a training set of 7291 digits and a test set of 2007 digits. (files "zip_train.dat" and "zip_test.dat" respectively).

The purpose is to continue the exercise 1 using Multivariate Regression and a Principal Components Regression and exercise 2 using IBA. Now we will use PLSR2 as a component based methodology to predict the digits.

Steps for conducting the practice

1. Read the "zip_train.dat" and "zip_test.dat" files provided. Select a 5% random sample (without replacement) of the train data. Use the same sample as previous exercises as your training data, and the complete test data for testing.
2. Define the response matrix (Y) and the predictor matrix (X). Center the predictor matrix.
3. Perform a PLSR2 using "CV" or "LOO" for validation. Decide how many components you retain for prediction?.
4. Predict the responses in the test data, be aware of the appropriate centering. Compute the average R2 in the test data.
5. Assign every test individual to the maximum response and compute the error rate.
6. Compare the obtained results with the previous ones with MVR, PCR and IBA.