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

HOMEWORK IBA

KMLMM
December 3, 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 Inter Batteries Analysis (IBA). Similarly as in PCR and CCA, 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 measure the relationship between both multivariate vectors by components derived from the original variables $t_h = Xa_h$ and $u_h = Yb_h$.

2 Methods

To perform the Inter Batteries Analysis, one may use the *interbat* function from R's *plsdepot* package. However, it's necessary to observe that in the considered dataset the training Y matrix is not full rank. We can maximally work with $rank(V_{XY})$, where V_{XY} is the covariance matrix between X and Y - for the ZIP dataset it's equal to 9. Hence, the Inter Batteries Analysis has to be computed manually.

We start by creating a matrix N which is a normalised (it's sum equals 1) diagonal matrix of dimensions n by n. Then we define $R_{12} = \frac{X'NY}{n}$. Based on it we can proceed to compute the matrix A by extracting the eigenvectors of $R_{12}R'_{12}$ and B by taking the eigenvectors of $R'_{12}R_{12}$. Next we need to determine how many components we want to retain. For this purpose $\Gamma = A'R_{12}B$ and the eigenvalues of $\Gamma\Gamma'$ are computed. Then we see how many eigenvalues are significant - greater than a given threshold which in this work was set to 10^{-8}. In this way 9 components are selected. Now we update A by selecting its 9 first rows, compute $T = XA$ and update the matrix $B = A(T'NT)T'NY$.

3 Results

To predict the class of observations in X_{test} it's enough to take $argmax$ of $X_{test}B$. Finally the mean values of R^2 and error can be computed. Interestingly, by getting the predictions for the test set none of the images were classified as 9. Obtained values of R^2 for all classes are: $0.31, 0.25, 0.37, 0.14, 0.17, 0.01, 0.10, 0.29, NA, 0.11$ and the $R^2_{mean} = 0.1759$. The error rate for the test dataset equals 52.86%.