ZIP Practical Work : Inter-Battery Analysis

Multivariate Modeling

Martin Guy and Hannes Leskela

December 7, 2016

FIB UPC UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Abstract

The goal of this exercise is to recognize the right number that is written, on normalized handwritten digits, automatically scanned from envelopes by the U.S. postal service. For this exercise we will use **Inter-Battery Analysis** to predict these digits. After selecting the right number of components, we present and discuss our results with this method. The results show that we can explain roughly 80% of the data using 10 of our latent factors, but almost the same percent using only 8 factors. The 20% that are unexplained are considered different types of error.

Introduction

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). The goal of this exercise is to recognize the right number that is written. The purpose is to continue the exercise we did for session 1 using **Multivariate Regression** and a **Principal Components Regression**. Now we will try **Inter-Battery Analysis** (IBA) as a component based methodology to predict the digits.

The dataset

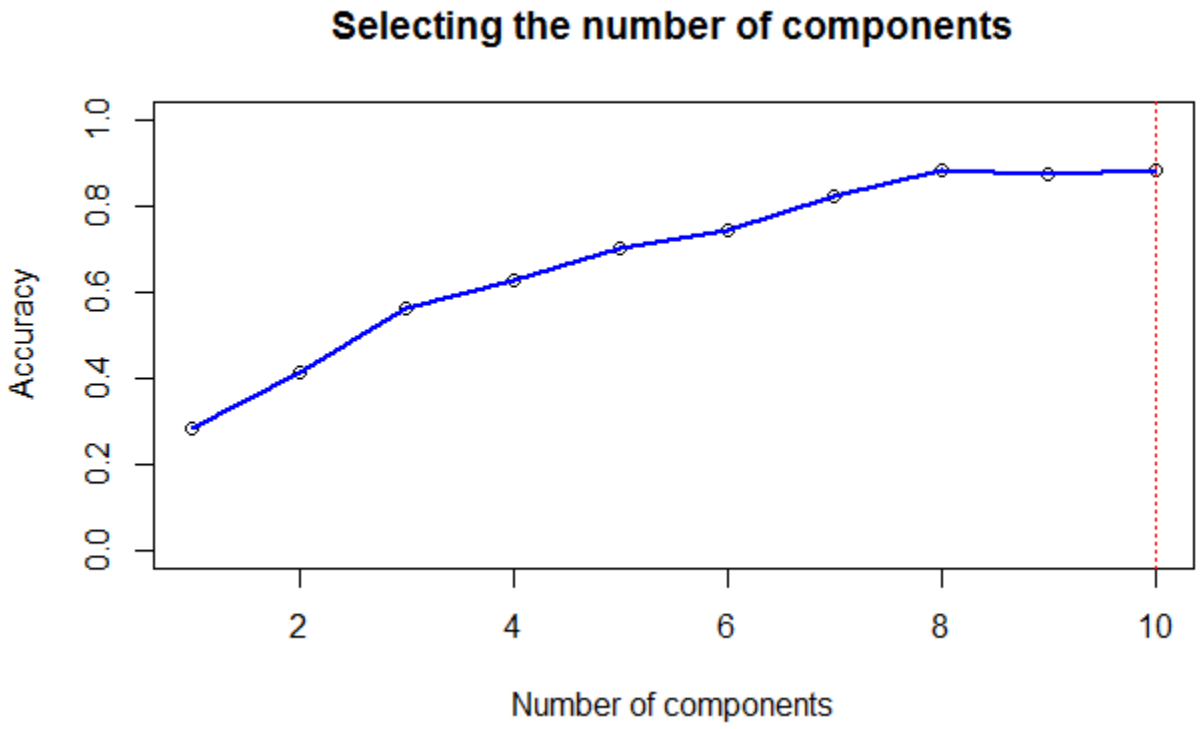
We dispose of a training set of 7291 digits (`zip_train.dat`) and a test set of 2007 digits (`zip_test.dat`). Each line consists of the id (0-9) followed by the 256 grayscale values. In the following we define n as 5% of the training size, so $n = 364$ (rounded down). Based on that, the predictor matrix X will be of size $256 \times n$ (one column for each grayscale value). However, the response matrix Y will not be of size $1 \times n$ but we will arrange it to a matrix of size $10 \times n$ where y_{ij} is the probability of the i-th test to be the number $j - 1$. This allows us to treat the variable as categorical ("attribute") values instead of numerical which is a commonly made assumption for regression methods. The meaning of this is that the model does not make as many assumptions between the response values, like that subtracting a three from a nine would yield a six, when they do not have that kind of relationship.[1]

Inter-Battery Analysis

The inter-battery method of factor analysis was devised to provide information relevant to the stability of factors over different selections of tests.[2] The method builds on the work of Tucker from his paper "An inter-battery method of factor analysis". The method works by creating two batteries of tests, i.e. two sets of correlated presumptions about the data. These two sets have the property that they depend on the same factors, but that they are not parallel which means that the tests are not constructed to have the same μ or σ^2. We then determine the IBA factors by maximizing the sum of covariances between the variables in the two groups. We then rotate the axes for the two batteries independently, and finally take the correlation between corresponding variables from the batteries.

Choosing the number of components

We select the number of components to use in our model by calculating the accuracy we get by taking one more component each time. This works similarly to how we choose the number of components in PCA, where the first components explains the most of the data, resulting in a plot with a decreasing k-value and a maximum of 1, as seen in the plot below.



Then, we just have to pick the model with the highest accuracy, which in this case is the one that uses 10 components. However, as we can see the difference between 8 and 10 components is marginal, but using a model with less components means that the model is less complex. Explaining our data using a model that has many latent components means that we increase the complexity, and thus the bias towards our data, and possibly a higher variance.

Results

Remind that we considered only 5% of our training set, so only 365 samples but test on the full test set which contains 2007 digits. We obtain an **error rate of 0.2013** so **80% of accuracy**.

Conclusion

Even though IBA is a method that is not used extensively compared to for instance CCA or PCR, our experiments show that IBA is at least as good as PCR. One key difference lies in using latent factors vs using principal components to build our model, which both have their respective benefits. IBA has per default a really low amount of latent factors, while PCR gives us more of a choice in choosing the number of components to use to describe our model.

In practice this resulted in longer run-times for PCR compared to IBA, which is important to consider when selecting a model.

Also, since we use only 5% of our training set we might improve the accuracy by taking more samples.

Not all of the error is necessarily from our model, since it could also be a sampling error for our train or test data. This bias could be accidentally introduced during our sampling, for instance by choosing letters from a certain area more prominently than others. There could also be a hidden bias, for instance the possibility that a certain age group send more letters than others.

References

[1] Smita, Skrivanek

`https://www.moresteam.com/whitepapers/download/dummy-variables.pdf`

Retrieved December 7, 2016

[2] Tucker, Ledyard R.

`https://link.springer.com/article/10.1007%2FBF02289009`

Retrieved December 7, 2016