Prediction Comparison with NIR_Dough

Raju Rimal, Trygve Almøy and Solve Sæbø

10 February, 2019

Contents

About Dataset

Dataset Name: NIR_Dough
Data provided by: Kristian Liland

The calibration set has $N = 40$ samples of $p = 700$ wavelengths (1100-2498nm in steps of 2nm). Four response variables yield *percentages of* (a) *fat*, (b) *sucrose*, (c) *flour* and (d) *water* from measurements of biscuit dough. A separate test set contains 32 samples. The two sets were created and measured on different occasions. (A twist to partial least squares regression).

Loading the data

The structure of dataset contains a predictor matrix X, a response matrix Y and a vector specifying which one is the training set and which one are not. In this dataset there are 72 observation in total with 700 predictors and 4 responses. In the following analysis, we will use 15 number of components. We will center the datasets but they are not scaled. Further 40 first observations are considered as training samples.

```
List of 3
 $ X    : 'AsIs' num [1:72, 1:700] 0.25 0.256 0.275 0.244 0.243 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : NULL
  .. ..$ : chr [1:700] "1100" "1102" "1104" "1106" ...
 $ Y    : 'AsIs' num [1:72, 1:4] 21.1 18.4 15.3 21.6 17.9 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : NULL
  .. ..$ : chr [1:4] "fat" "sucrose" "flour" "water"
 $ train: logi [1:72] TRUE TRUE TRUE TRUE TRUE TRUE ...
 - attr(*, "OldName")= chr [1:3] "NIR" "ingredient" "train"
```

Estimating Relevant Components

The relevant predictor components are estimated based on the training samples we have. It is also interesting to see weather the test samples have similar strucutre or not. The estimation of the relevant components are also obtained on the test samples. Figure 1 shows the estimated relevant components for both test and trining datasets. The estimation is done based on following expression.

$$\text{scaled absolute covariance} = \frac{\left|\widehat{\Sigma}_{yz}\right|}{\max\left|\widehat{\Sigma}_{yz}\right|}$$

where, $Z = X \times V$ and V is the eigenvectors corresponding to predictor X.

The bar in the background of the plot are the scaled eigevalues divided by the maximum eigenvalue corresponding to covariance of the predictor matrix.

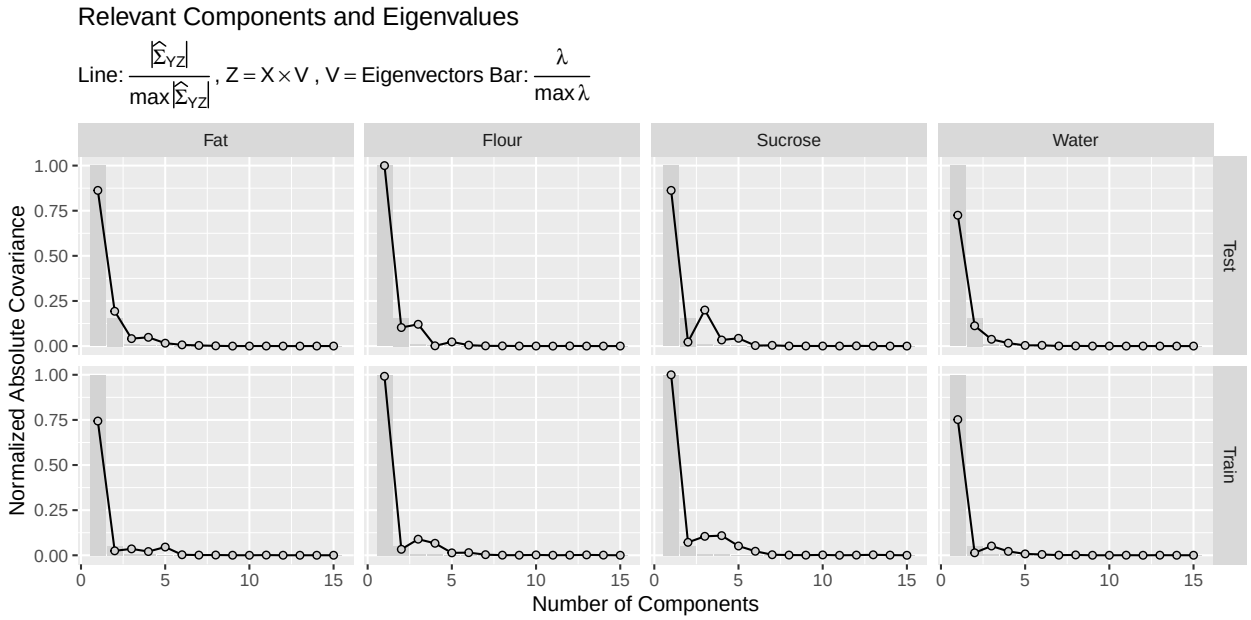

Figure 1: Estimated relevant Components for training and test samples

Eigenvalues of Response Matrix

Since our simulation is based on single informative response component, it is necessary to see the estimated response latent structure. Figure-2 shows the cumulative sum of the variation caputred by each additional components in response and predictors for both test and training set. Here in most of the datasets, the variation is captured by just few number of components in both predictor and response. The eigenvalues are computed based on the covariance of the dataset.

Statistical Modeling

PCR, PLS1, PLS2, Xenv and Senv are used to fit the model using the dataset *NIR_Dough*. Here the models are fitted using 1 to 15 number of components. Since we are testing the performace with each components,

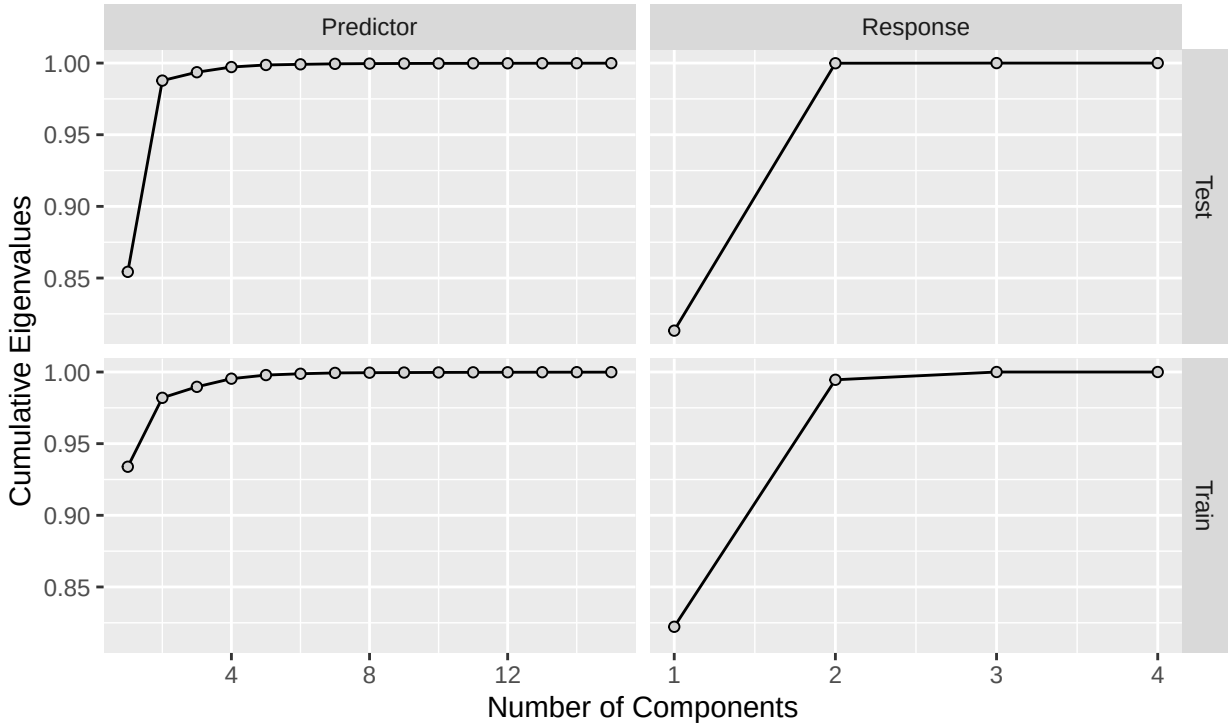

Figure 2: Cumulative sum of eigenvalues

we have not performed any cross-validation. Figure-3 shows both traing and test prediction error for each response variables obtained from different prediction methods. A general observations in most of the datasets are:

a) Test prediction error for all methods fluctute a lot specially when the components has small covariance with response but has large eigenvalues and vice versa.
b) Scaling (divided by standard deviation) removed all useful relationship and impossible to determine and difference. *I could only apply Scaling for PLS and PCR as their function allow for scaling, so to apply for all methods, I need to scale them manually and descale after I get the regression coefficients. (need some suggestion)*

Following table shows the minimum prediction error from each of the method in Figure-3 along with the number of components they have used to get that minimum prediction error. General observation for most of the dataset is that PLS are better in most of the respect but for difficult response, Senv also somewhat robust for the purpose.

Method	fat	flour	sucrose	water
PCR	0.44 (12)	0.86 (5)	**0.93** (8)	0.42 (10)
PLS1	**0.36** (11)	0.91 (5)	1.09 (6)	0.54 (7)
PLS2	0.46 (10)	0.8 (5)	1.08 (6)	**0.39** (9)
Senv	0.49 (9)	**0.79** (4)	1.05 (4)	0.56 (7)
Xenv	0.44 (11)	0.81 (4)	1.17 (6)	0.5 (3)

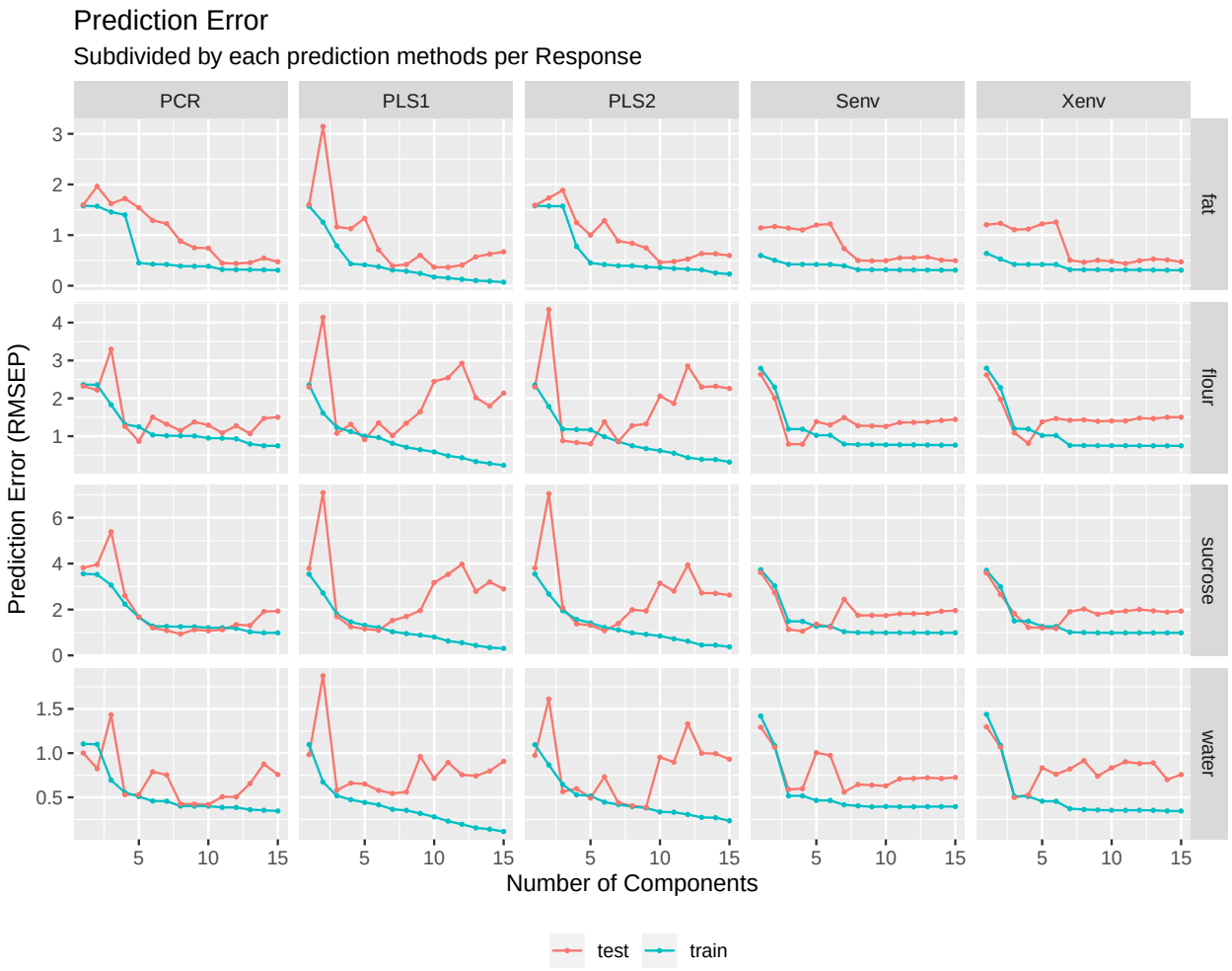

Figure 3: Test and Train Prediction Error for each response using PCR, PLS1, PLS2, Xenv and Senv methods using RMSE as the measuring criteria.