

# Prediction Comparison with majones

*Raju Rimal, Trygve Almøy and Solve Sæbø*

*16 February, 2019*

## Contents

About Dataset	1
Loading the data	1
Estimating Relevant Components	1
Eigenvalues of Response Matrix	2
Statistical Modeling	2

## About Dataset

**Dataset Name:** majones

**Data provided by:** Kristian Liland

## 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 162 observation in total with 351 predictors and 5 responses. In the following analysis, we will use 15 number of components. We will center the datasets but they are not scaled. Further 120 first observations are considered as training samples.

```
List of 3
$ X      : num [1:162, 1:351] 0.246 0.243 0.25 0.273 0.267 ...
..- attr(*, "dimnames")=List of 2
.. ..$ : NULL
.. ..$ : chr [1:351] "x1" "x2" "x3" "x4" ...
$ Y      : num [1:162, 1:5] -1 -1 -1 1 1 1 -1 -1 -1 1 ...
..- attr(*, "dimnames")=List of 2
.. ..$ : NULL
.. ..$ : chr [1:5] "y2" "y3" "y4" "y5" ...
$ train: logi [1:162] TRUE TRUE TRUE TRUE TRUE TRUE ...
- attr(*, "sel1")= num [1, 1:120] 1 2 3 4 5 6 7 8 9 10 ...
- attr(*, "group")= num [1:162] 1 1 1 1 1 1 1 1 1 1 ...
- attr(*, "OldName")= chr [1:3] "X" "Y" "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 training datasets. The estimation is done based on following expression.

$$\text{scaled absolute covariance} = \frac{|\hat{\Sigma}_{yz}|}{\max |\hat{\Sigma}_{yz}|}$$

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 eigenvalues divided by the maximum eigenvalue corresponding to covariance of the predictor matrix.

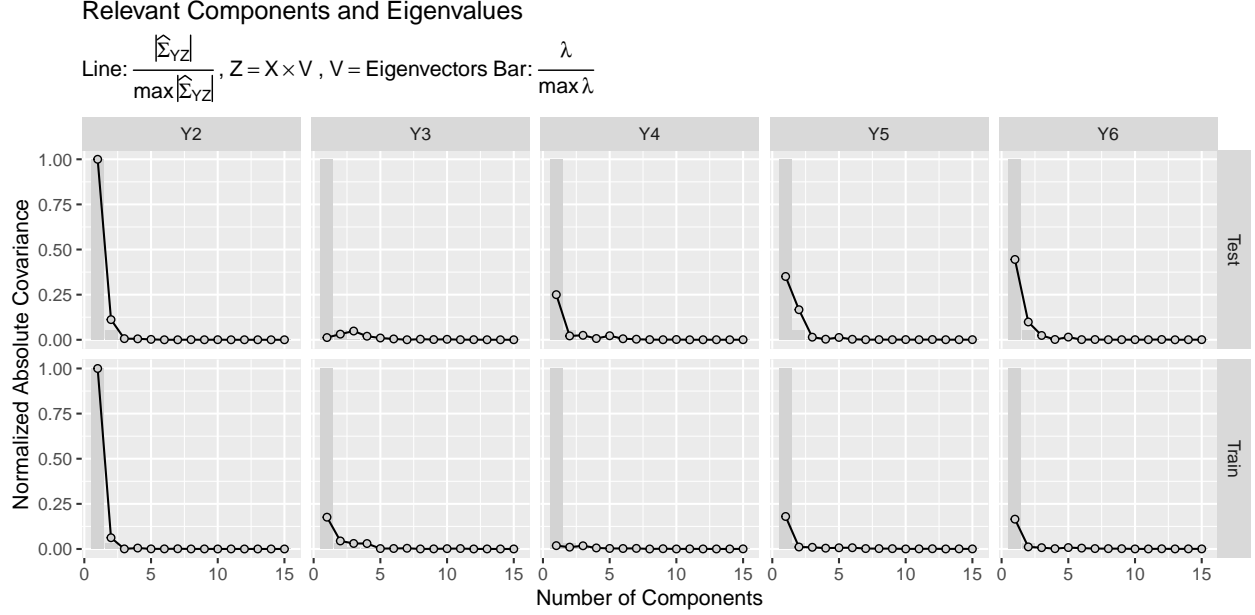


Figure 1: Estimated relevant Components for training and test samples

## 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 captured 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 *majones*. Here the models are fitted using 1 to 15 number of components. Since we are testing the performance with each components, we have not performed any cross-validation. Figure-3 shows both training and test prediction error for each response variables obtained from different prediction methods. A general observations in most of the datasets are:

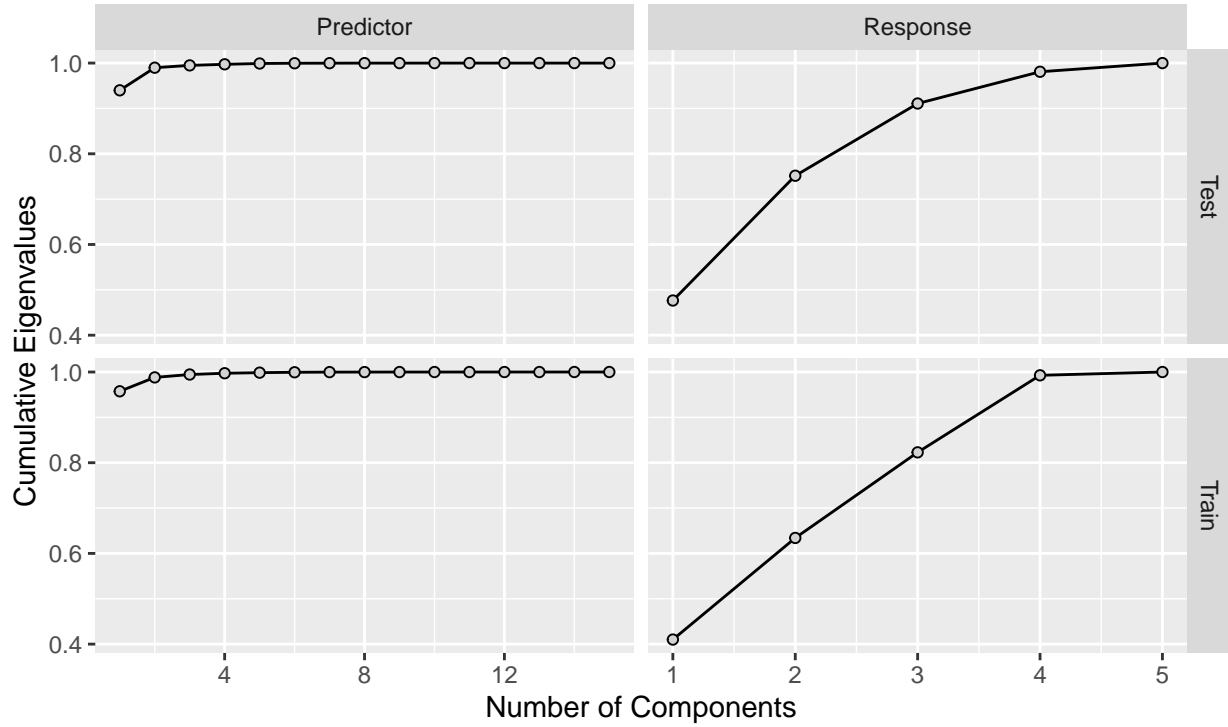


Figure 2: Cumulative sum of eigenvalues

- Test prediction error for all methods fluctuate a lot specially when the components has small covariance with response but has large eigenvalues and vice versa.
- 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	y1	y2	y3	y4	y5
PCR	<b>0</b> (2)	0.36 (12)	0.69 (4)	0.45 (9)	0.55 (5)
PLS1	<b>0</b> (2)	<b>0.24</b> (14)	0.64 (11)	<b>0.36</b> (8)	0.33 (14)
PLS2	<b>0</b> (2)	0.33 (9)	<b>0.62</b> (15)	0.43 (8)	<b>0.29</b> (13)
Senv	<b>0</b> (2)	0.33 (5)	0.69 (4)	0.57 (1)	0.45 (12)
Xenv	<b>0</b> (2)	0.33 (6)	0.69 (4)	0.45 (5)	0.57 (1)

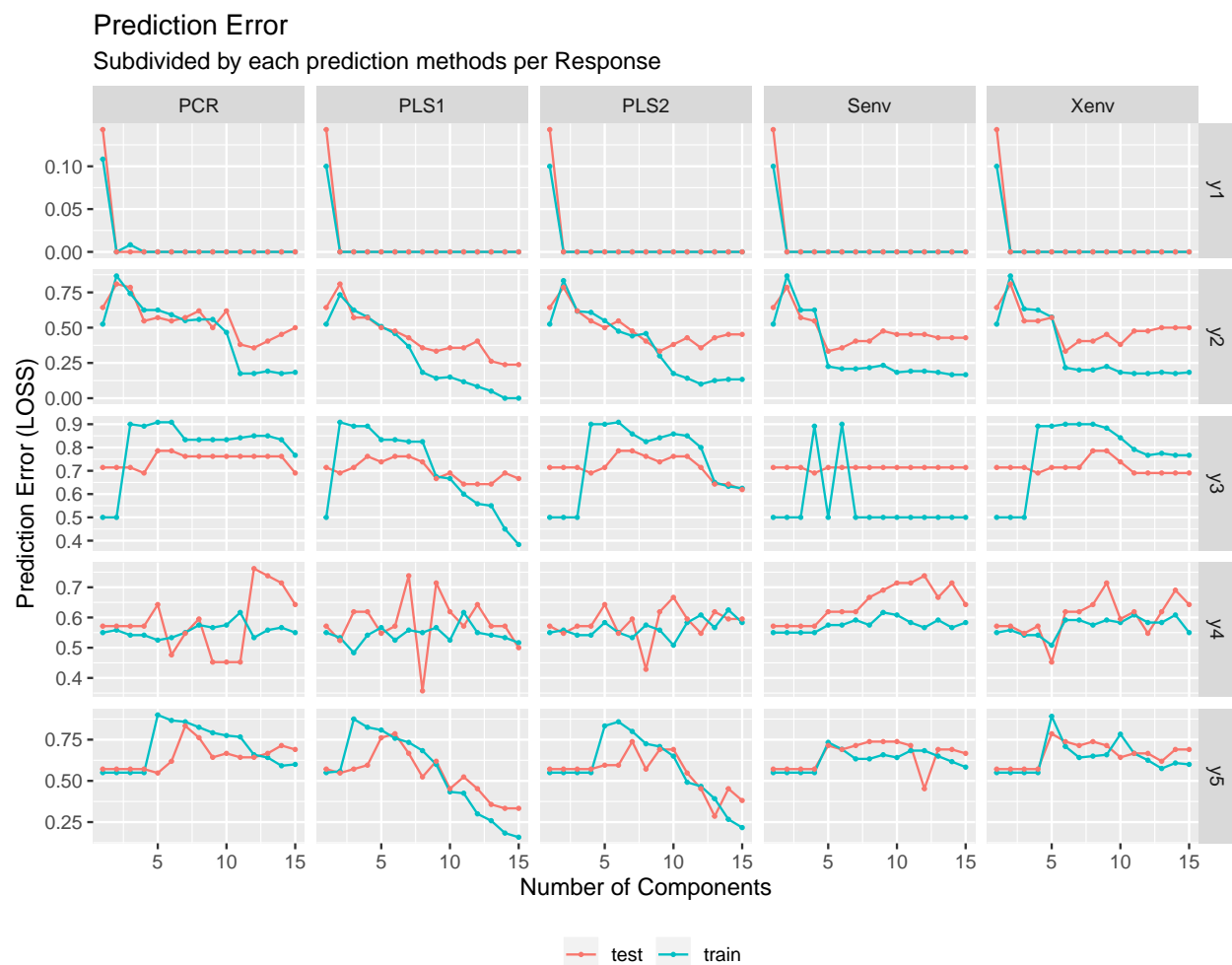


Figure 3: Test and Train Prediction Error for each response using PCR, PLS1, PLS2, Xenv and Senv methods using RMSE as the measuring criteria.