

Comparison of Multivariate Prediction Methods

Raju Rimal^{a,*}, Trygve Almøy^a, Solve Sæbø^b

^aFaculty of Chemistry and Bioinformatics, Norwegian University of Life Sciences, Ås, Norway

^bProrector, Norwegian University of Life Sciences, Ås, Norway

Abstract

While Data science is battling to extract information from the enormous explosion of data, many estimators and algorithms are being developed for better prediction. Researchers and data scientists often introduce new methods and evaluate them based on various aspects of data. However, studies on the impact of/on model with multiple response model is limited. This study compares some newly-developed (envelope) and well-established (PLS, PCR) prediction methods based on simulated data specifically designed by varying properties such as multicollinearity, correlation between multiple responses and amount of information content in predictor variables. This study aims to give some insight on these methods and help researcher to understand and use them for further study.

Keywords: model-comparison, multi-response, simrel

1. Introduction

Prediction has been an essential components of modern data science, weather it is statistical analysis or machine learning. Modern technology has facilitated a massive explosion of data, however, such data often contain irrelevant information consequently making prediction difficult. Researchers are devising new methods and algorithms in order to extract information to create robust predictive models. Mostly such models contain

*Corresponding Author

Email addresses: raju.rimal@nmbu.no (Raju Rimal), trygve.almoy@nmbu.no (Trygve Almøy), solve.sabo@nmbu.no (Solve Sæbø)

predictor variables that are directly or indirectly correlated with other predictor variables. In addition studies often constitute of many response variables correlated with each other. These interlinked relationships influence any study, whether it is predictive modeling or inference.

Modern inter-disciplinary research fields such as chemometrics, econometrics and bioinformatics are handling multi-response models extensively. This paper attempts to compare some multivariate prediction methods based on their prediction performance on linear model data with specific properties. The properties includes correlation between response variables, correlation between predictor variables, number of predictor variables and the position of relevant predictor components. These properties are discussed more in the [Experimental Design](#) section. [Sæbø et al. \(2015\)](#) and [Almøy \(1996\)](#) have made a similar comparison in the single response setting. In addition, [Rimal et al. \(2018\)](#) has also made a basic comparison on some prediction methods and their interaction with the data properties of a multi-response model. The main aim of this paper is to present a comprehensive comparison of contemporary prediction methods such as simultaneous envelope estimation (Senv) ([Cook and Zhang, 2015](#)) and envelope estimation in predictor space (Xenv) ([Cook et al., 2010](#)) with customary prediction methods such as Principal Component Regression (PCR), Partial Least Squares Regression (PLS) using simulated dataset with controlled properties. An experimental design and the methods under comparison are discussed further, followed by a brief discussion of the strategy behind the data simulation.

2. Simulation Model

Consider a model where the response vector (\mathbf{y}) with m elements and predictor vector (\mathbf{x}) with p elements follow a multivariate normal distribution as follows,

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{x} \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} \boldsymbol{\mu}_y \\ \boldsymbol{\mu}_x \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{yx} \\ \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_{xx} \end{bmatrix} \right) \quad (1)$$

where, $\boldsymbol{\Sigma}_{yy}$ and $\boldsymbol{\Sigma}_{xx}$ are the variance-covariance matrices of \mathbf{y} and \mathbf{x} , respectively, $\boldsymbol{\Sigma}_{xy}$ is

Relevant space within a model

A concept for reduction of regression models

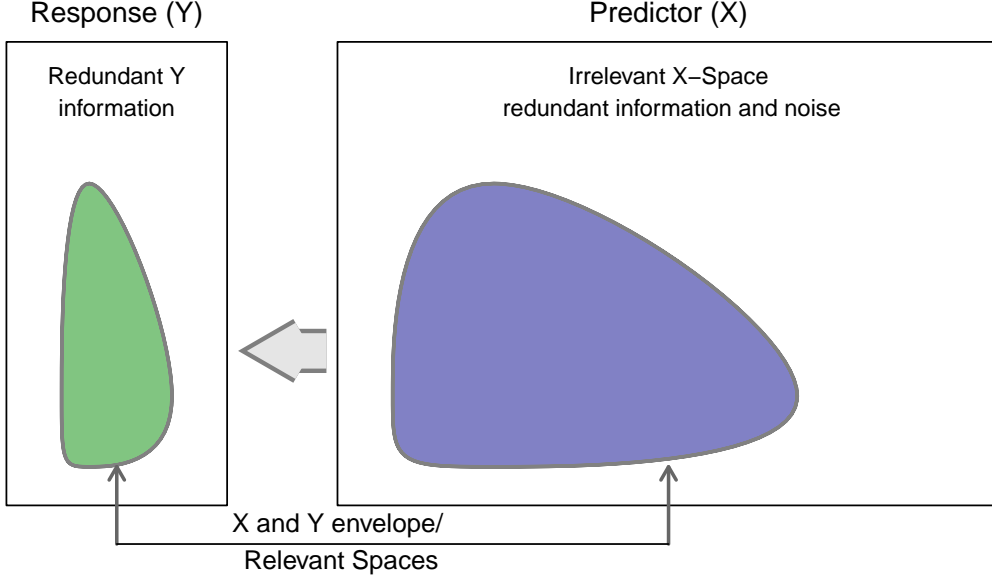


Figure 1: Relevant space in a regression model

the covariance between \mathbf{x} and \mathbf{y} and μ_y and μ_x are mean vectors of \mathbf{x} and \mathbf{y} , respectively. A linear model based on (1) is,

$$\mathbf{y} = \mu_y + \beta^t(\mathbf{x} - \mu_x) + \epsilon \quad (2)$$

where, β^t is a matrix of regression coefficients and ϵ is an error term such that $\epsilon \sim \mathcal{N}(0, \Sigma_{y|x})$

In a model like (2), we assume that the variation in response \mathbf{y} is partly explained by the predictor \mathbf{x} . However, in many situations, only a subspace of the predictor space is relevant for the variation in the response \mathbf{y} . This space can be referred to as the relevant space of \mathbf{x} and the rest as irrelevant space. In the similar manner, we can assume that a subset of the response space contains the information that the predictors can explain for a given model (Figure-1). Cook et al. (2010) and Cook and Zhang (2015) have referred to the relevant space as material space, and the irrelevant space as immaterial space.

With an orthogonal transformation of \mathbf{y} and \mathbf{x} to latent variables \mathbf{w} and \mathbf{z} , respectively, by $\mathbf{w} = \mathbf{Q}\mathbf{y}$ and $\mathbf{z} = \mathbf{R}\mathbf{x}$, where \mathbf{Q} and \mathbf{R} are orthogonal rotation matrices, an equivalent model to (2) in terms of the latent variables can be written as,

$$\mathbf{w} = \boldsymbol{\mu}_w + \boldsymbol{\alpha}^t(\mathbf{z} - \boldsymbol{\mu}_z) + \boldsymbol{\tau} \quad (3)$$

where, $\boldsymbol{\alpha}^t$ is a matrix of regression coefficients and $\boldsymbol{\tau}$ is an error term such that $\boldsymbol{\tau} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{w|z})$. Model (3) follows the distribution,

$$\begin{bmatrix} \mathbf{w} \\ \mathbf{z} \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} \boldsymbol{\mu}_w \\ \boldsymbol{\mu}_z \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{ww} & \boldsymbol{\Sigma}_{wz} \\ \boldsymbol{\Sigma}_{zw} & \boldsymbol{\Sigma}_{zz} \end{bmatrix} \right) \quad (4)$$

where, $\boldsymbol{\Sigma}_{ww}$ and $\boldsymbol{\Sigma}_{zz}$ are the variance-covariance matrices of \mathbf{w} and \mathbf{z} , respectively. $\boldsymbol{\Sigma}_{zw}$ is the covariance between \mathbf{z} and \mathbf{w} . $\boldsymbol{\mu}_w$ and $\boldsymbol{\mu}_z$ are mean vector of \mathbf{z} and \mathbf{w} respectively.

Here, the elements of \mathbf{w} and \mathbf{z} are the principal components of responses and predictors, which will respectively be referred as “response components” and “predictor components”. The column vectors of respective rotation matrices \mathbf{Q} and \mathbf{R} are the eigenvectors corresponding to these principal components.

Following the concept of relevant space, a subset of predictor components can be imagined to span the predictor space. These components can be regarded as relevant predictor components. Naes and Martens (1985) introduced the concept of relevant components which was explored further by Helland (1990), Næs and Helland (1993), Helland and Almøy (1994) and Helland (2000). The corresponding eigenvectors were referred to as relevant eigenvectors. A similar logic is introduced by Cook et al. (2010) and later by Cook et al. (2013) as an envelope which is the space spanned by the relevant eigenvectors (Cook, 2018, pp. 101).

In addition, various simulation studies have been performed with the model based on the concept of relevant subspace. A simulation study by Almøy (1996) has used a single response simulation model based on reduced regression and has compared some contemporary multivariate estimators. In the recent years Helland et al. (2012), Sæbø

et al. (2015), Helland et al. (2018) and Rimal et al. (2018) implemented similar simulation examples as we are discussing in this study. This paper, however, presents an extensive simulation study based on multi-response data simulated with experimental design and compares relatively new methods such as simultaneous envelopes with well established methods such as partial least squares and principal components regression. Rimal et al. (2018) has a detail discussion about the simulation model that we have opted here. The following section presents the estimators under comparison in more detail.

3. Prediction Methods

Partial least squares regression (PLS) and Principal component regression (PCR) has been used in many disciplines such as chemometrics, econometrics, bioinformatics and machine learning, where wide predictor matrices, i.e. p (number of predictors) $> n$ (number of observation) is common. These methods are popular in multivariate analysis, especially for exploratory studies and prediction.

In recent years, a concept of envelope introduced by Cook et al. (2007) based on reduction in regression model has been implemented for the development of envelope estimation in the subsequent papers.

In this study, we will follow estimation methods based on their prediction performance on data simulated with different controlled properties.

Principal Components Regression (PCR): Principal components are the linear combinations of predictor variables such that the transformation makes the new variables uncorrelated and the variation of the original dataset captured by them are ordered. In other words, each successive components captures maximum variation left by the preceding components in predictor variables (Jolliffe, 2002). Principal components regression uses these principal components to explain the variation in the response.

Partial Least Squares (PLS): Two variants of PLS: PLS1 and PLS2 will be used for comparison. The first one considers individual response variables separately, i.e. each response is predicted with a single response model, while the latter considers all

response variables together. In PLS regression the components are determined such as to maximize a covariance between response and predictors (de Jong, 1993).

Envelopes: The envelope, introduced by Cook et al. (2007), was first used as a response envelope (Cook et al., 2010) as a smallest subspace \mathcal{E} in the response space such that the span of regression coefficients lies in that space. Since a multivariate linear regression model contains relevant (material) and irrelevant (immaterial) variation in both response and predictor, the relevant part provides information, while irrelevant part increases the estimative variation. The concept of envelope uses the relevant part for estimation while excluding the irrelevant part consequently increasing the efficiency of the model (Cook and Zhang, 2016).

The concept was later extended to the predictor space, where the predictor envelope was defined (Cook et al., 2013). Further Cook and Zhang (2015) uses envelopes for joint reduction of the responses and predictors and argued to produce efficiency gains greater than using individual envelopes either of the response and predictors. All the variants of envelope estimations are based on maximum likelihood estimation. Here in this study we will also use predictor envelope (Xenv) and simultaneous envelope (Senv) for the comparison.

3.1. Modification in envelope estimation

Since envelope estimators (Xenv and Senv) are based on maximum likelihood estimation (MLE), it fails to estimate in case of wide matrices, i.e. $p > n$. In order to incorporate these methods in our comparison, we have used the principal components (\mathbf{z}) of the predictor variables (\mathbf{x}) as predictors, using the required number of components for capturing 97.5% of the variation in \mathbf{x} . The new set of variables, \mathbf{z} , were used for envelope estimation. The regression coefficients ($\hat{\boldsymbol{\alpha}}$) corresponding to these new variables \mathbf{z} were transformed back to obtain coefficients for each predictor variable as,

$$\hat{\boldsymbol{\beta}} = \mathbf{e}_k \hat{\boldsymbol{\alpha}}_k$$

where, \mathbf{e}_k is the eigenvectors with k number of components.

4. Experimental Design

This study compares prediction methods based on their prediction ability. Data with specific properties are simulated, some of which are easier to predict than others. These data are simulated using the R-package `simrel`, which is discussed in [Sæbø et al. \(2015\)](#) and [Rimal et al. \(2018\)](#). Here we will use four different factors to vary the property of the data: a) Number of predictors (p), b) Multicollinearity in predictor variables (γ), c) Correlation in response variables (η) and d) position of predictor components relevant for the response (relpos). Using two levels of p , γ and relpos and four levels of η , 32 set of distinct properties are designed for the simulation.

Number of predictors: In order to observe the performance of the methods on tall and wide predictor matrices, 20 and 250 predictor variables are simulated. Parameter p controls this properties in the `simrel` function.

Multicollinearity in predictor variables: Highly collinear predictors can be explained completely by few components. The parameter γ (γ) in `simrel` controls decline in the eigenvalues of the predictor variables as (5).

$$\lambda_i = e^{-\gamma(i-1)}, \gamma > 0 \text{ and } i = 1, 2, \dots, p \quad (5)$$

Here, $\lambda_i, i = 1, 2, \dots, p$ are eigenvalues of the predictor variables. Here we have used 0.2 and 0.9 as different levels of γ . The higher the value of γ , the higher will be the correlation between predictors and vice versa.

Correlation in response variables: Correlation among response variables is a less explored area. Here we have tried to explore that part with 4 levels of correlation in the response variables. We have used the η (η) parameter of `simrel` for controlling the decline in eigenvalues corresponding to the response variables as (6).

$$\kappa_i = e^{-\eta(i-1)}, \eta > 0 \text{ and } j = 1, 2, \dots, m \quad (6)$$

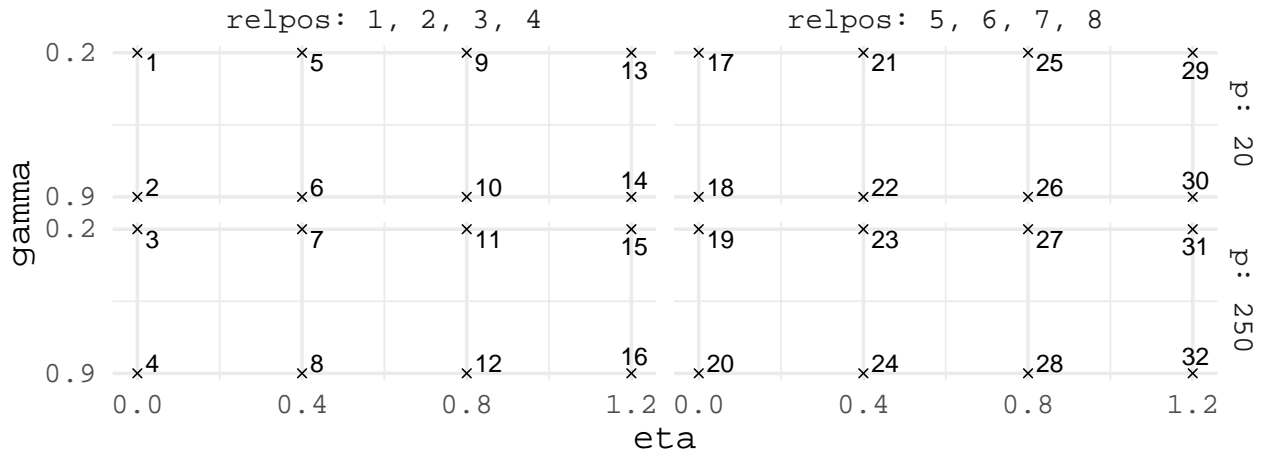


Figure 2: Experimental Design of simulation parameters. Each point represents an unique data property.

Here, $\kappa_i, i = 1, 2, \dots, m$ are the eigenvalues of the response variables and m is the number of response variables. Here we have used 0, 0.4, 0.8 and 1.2 as different levels of η . The larger the value of η , the larger will be the correlation between response variables and vice versa.

Position of predictor components relevant to the response: The principal components of the predictors are ordered. The first principal component captures most of the variation in the predictors. The second captures the most in the rest that is left by the first principal components and so on. In highly collinear predictors, the variation captured by the first few components is relatively high. However, if those components are not relevant for the response, prediction becomes difficult (Helland and Almøy, 1994). Here, two levels of the positions of these relevant components are used: 1, 2, 3, 4 and 5, 6, 7, 8.

Further, a complete factorial design from the levels of the above given parameters gave us 32 designs. Each design is associated with a dataset having unique properties. Figure~2, shows all the designs. For each design and prediction method, 50 datasets were simulated for replication. In total, there were $5 \times 32 \times 50$, i.e. 8000 dataset simulated.

Common parameters: Each dataset was simulated with $n = 100$ number of observation and $m = 4$ response variables. Further, the coefficient of determination correspond-

ing to each response components in all the designs is set to and 0.8. In addition, we have assumed that there is only one informative response component. Hence, the informative response component is rotated orthogonally together with three uninformative response components to generate four response variables. This spread out the information in all simulated response variables. For further details on the simulation tool see (Rimal et al., 2018).

An example of simulation parameters for the first design is as follows:

```
simrel(
  n      = 100,          ## Training samples
  p      = 20,          ## Predictors
  m      = 4,           ## Responses
  q      = 20,          ## Relevant predictors
  relpos = list(c(1, 2, 3, 4)), ## Relevant predictor components index
  eta    = 0,           ## Decay factor of response eigenvalues
  gamma  = 0.2,         ## Decay factor of predictor eigenvalues
  R2     = 0.8,         ## Coefficient of determination
  ypos   = list(c(1, 2, 3, 4)),
  type   = "multivariate"
)
```

Figure 3 shows the covariance structure of the data simulated with this design. The figure shows that the predictor components at position 1, 2, 3 and 4 are relevant for the first response component. After the rotation with orthogonal rotation matrix, all predictors are somewhat relevant for all response variables, fulfilling other desired properties like multicollinearity and coefficient of determination. For this same design, Figure 4(top left) shows that the predictor components 1, 2, 3 and 4 are relevant for the first response component. All other predictor components are irrelevant and all other response components are uninformative. However, due to orthogonal rotation of the informative response component

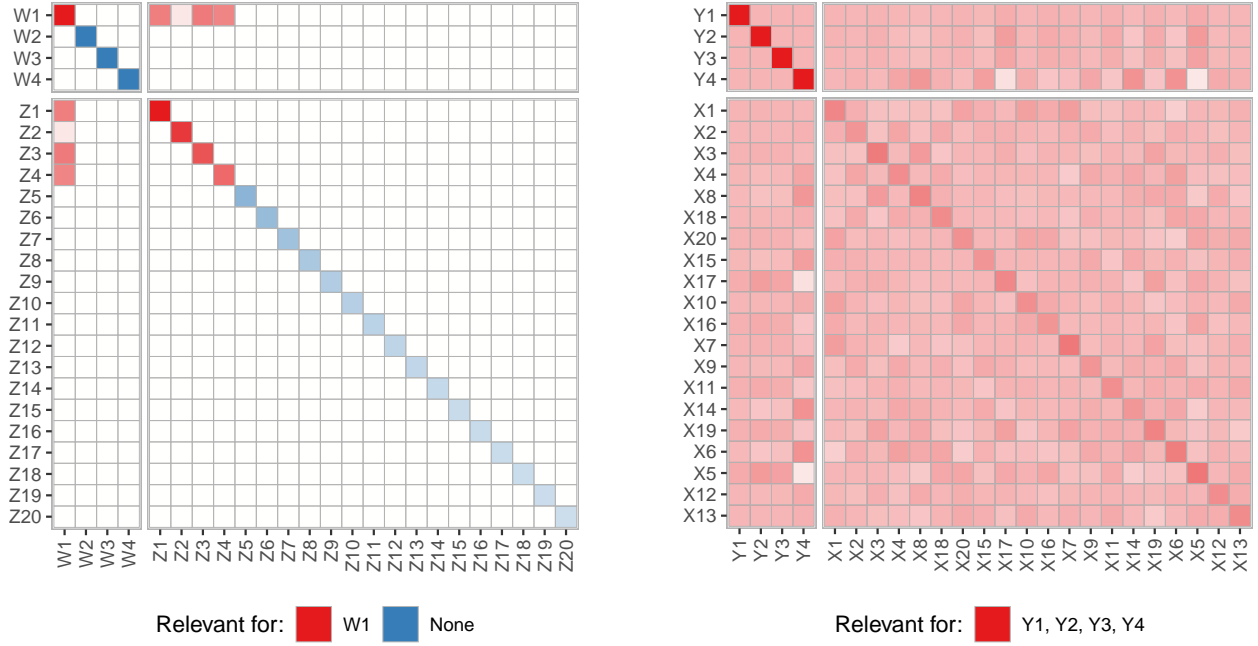


Figure 3: (left) Covariance structure of latent components. (right) Covariance structure of predictor and response

together with uninformative response components, all response variables in the population have similar covariance with the relevant predictor components (Figure 4(top right)). The sample covariances between the predictors components and predictor variables with response variables are in Figure 4 (bottom left) and (bottom right) respectively.

The discussion here is made on the first design. A similar discussion can be made on all 32 designs where each of the design holds the properties of the data they simulate. These data are used by the prediction methods discussed in previous section. Each prediction methods are given independent dataset simulated in order to give them equal opportunity to understand the dynamics in the data.

4.1. Basis of comparison

This study focuses mainly on the prediction performance of the methods and emphasis specifically on the interaction between the properties of the data, controlled by the simulation parameters, and the prediction methods. The prediction performance is measured by the prediction error for each response as in (7). The prediction is the theoretically computed

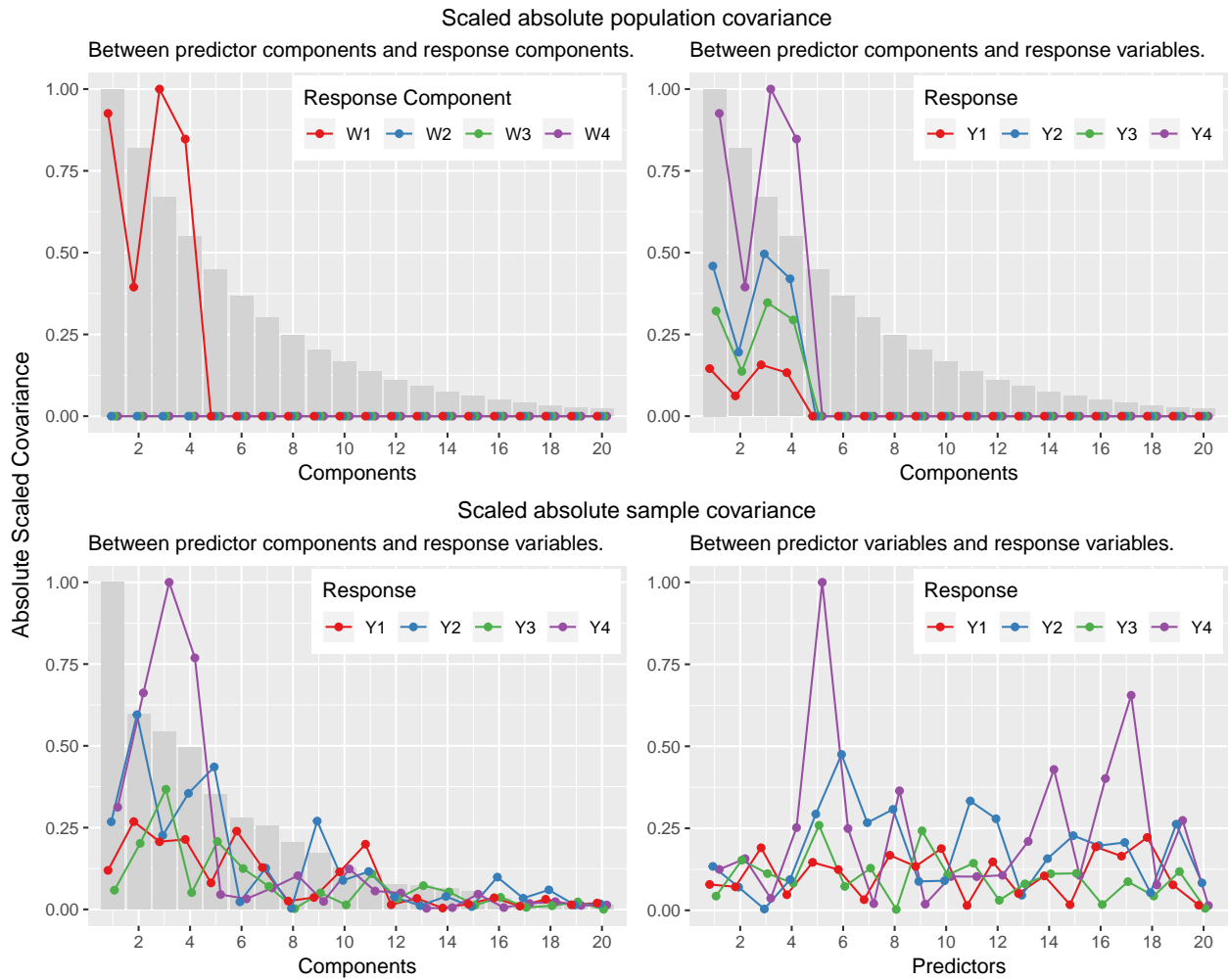


Figure 4: Expected Scaled absolute covariance between predictor components and response components (top left). Expected Scaled absolute covariance between predictor components and response variables (top right). Sample scaled absolute covariance between predictor components and response variables (bottom left). Sample scaled absolute covariance between predictor variables and response variables (bottom right). The bar in the background are eigenvalues corresponding to each components in population (top plots) and in sample (bottom plots).

expected prediction when the model is applied to unseen observations corresponding to each response variable.

$$\text{prediction error}_j = \frac{1}{\sigma_{y_j|x}} \left[\left(\beta_j - \hat{\beta}_j \right)^t \Sigma_{xx} \left(\beta_j - \hat{\beta}_j \right) \right] + 1 \quad (7)$$

where, Σ_{xx} is the true covariance matrix of predictor and $\sigma_{y_j|x}$ is the true model error obtained from simulation of response $j = 1, \dots, m$. Here prediction error is scaled by the true model error to remove its effect on prediction error. (Need better reason.) The prediction error in (7) is computed for all replications of 32 designs.

5. Exploration

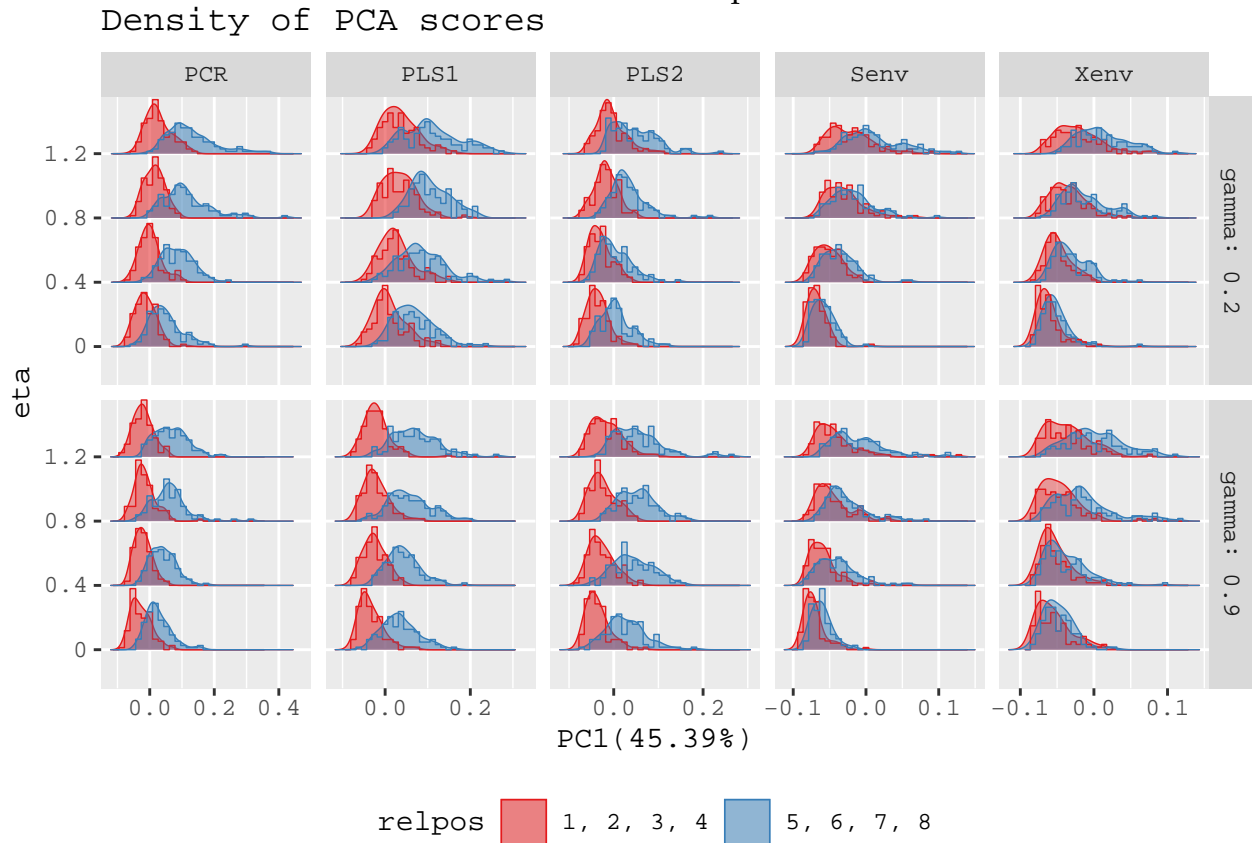
The structure of final data for further analysis contains five factors including the prediction methods and prediction error computed using (7) corresponding to four responses using 0 to 10 predictor components for all 50 replications. Thus there are 32 (design) \times 5 (methods) \times 11 (number of components) \times 50 (replications), i.e. 88000 observations. Here the variables Y1 to Y4 corresponds to prediction error of response variables.

Before performing any statistical analysis, this section tries to explore some observed relationship between the prediction error, the simulation parameters and the prediction methods. In the analysis onward for each factor combination and replication, we will use two datasets:

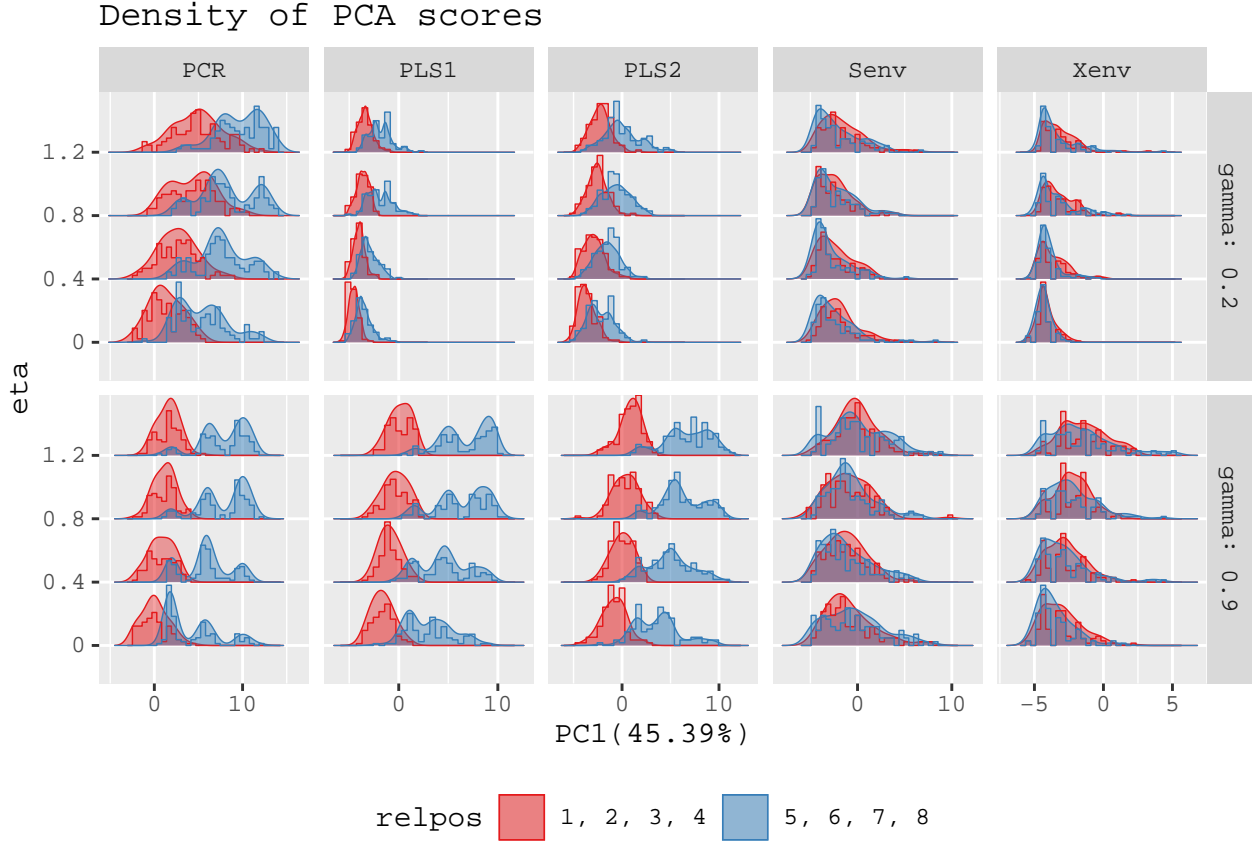
- One with minimum prediction error that a method can give using arbitrary number of components (*Error Dataset*)
- Another with the number of components that the method has used to give that minimum prediction error (*Component Dataset*)

Figure ?? plots first two principal components of *error dataset* using columns from Y1 to Y4. Since higher prediction error results in high scores the plot shows that the PCR, PLS1 and PLS2 methods are influenced by two levels of position of relevant predictor components.

When the position of relevant predictors are at position 5, 6, 7, 8, the eigenvalues corresponding to them becomes smaller making those designs difficult to model. However, the envelope methods have less influence of relpos in this regard. In addition, the effect of γ , the level of multicollinearity, has less effect of all cases. This indicates that the methods are somewhat robust to handle collinear predictors.



- A similar interpretation as the previous plot can be made in the score density. In addition, higher correlation in response (controlled by η parameter) yields in higher variation in the score of prediction error.
- The plot in the right shows that the envelope methods are able to leverage the effect of correlation between the response while in case of others, the effect is similar in low and high correlation between the responses.



6. Statistical Analysis

In order to carry out a proper statistical comparison, a multivariate analysis of variance (MANOVA) is used with minimum prediction error corresponding to each response variables for all design and their replicates. The third order interaction of simulation parameters (p , γ , η and relpos) and Methods is used as independent factors as (8).

$$\mathbf{y}_{abcdef} = \boldsymbol{\mu} + (p_a + \gamma_b + \eta_c + \text{relpos}_d + \text{Methods}_e)^3 + \boldsymbol{\varepsilon}_{abcdef} \quad (8)$$

where, \mathbf{y}_{abcdef} is a vector of prediction error for factors,

- $p_a = 20$ and 250
- $\gamma_b = 0.2$ and 0.9
- $\eta_c = 0, 0.4, 0.8$ and 1.2
- $\text{relpos}_d = 1, 2, 3, 4$ and $5, 6, 7, 8$

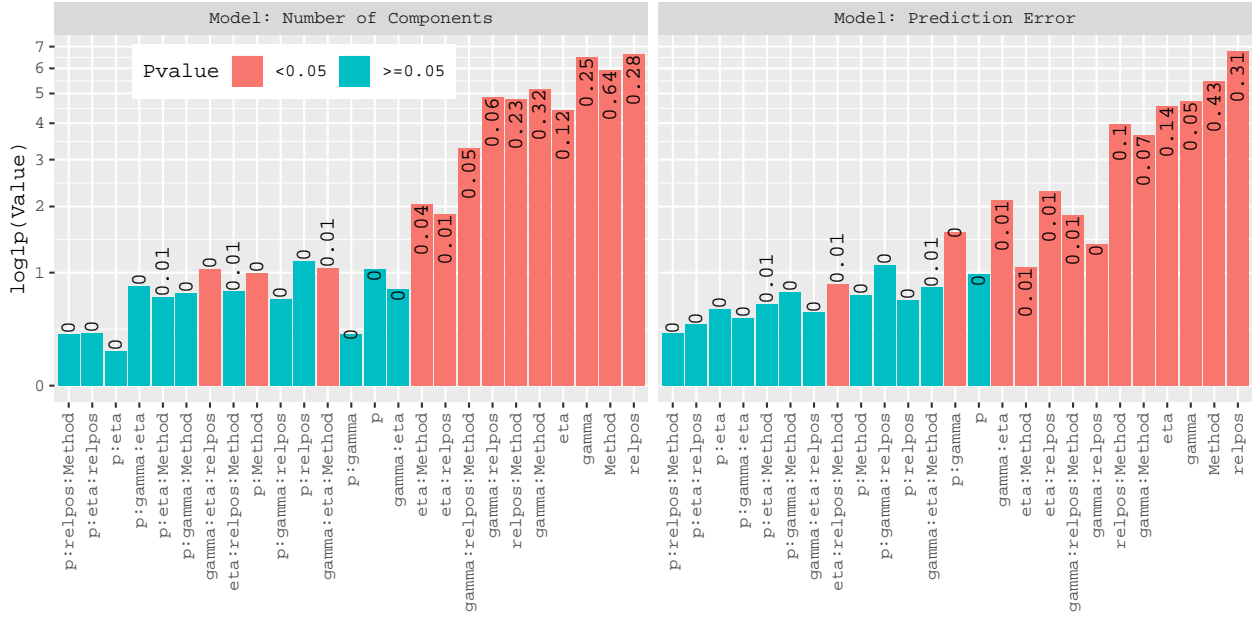


Figure 5: Pillai Statistic and F-value for the MANOVA model. The bar represents the F-value and the text labels are Pillai Statistic for corresponding factor.

- $\text{Methods}_e = \text{PCR}, \text{PLS1}, \text{PLS2}, \text{Xenv} \text{ and } \text{Senv}$

In concise vector form, we can write as (9).

$$\mathbf{y} = \boldsymbol{\mu} + (\mathbf{p} + \text{gamma} + \text{eta} + \text{relpos} + \text{Methods})^3 + \boldsymbol{\varepsilon} \quad (9)$$

where, \mathbf{y} is the vector of prediction error corresponding to response $y_j, j = 1, \dots, 4$.

Prediction methods also varies on number of components they use to get the minimum prediction error. A similar model as (8) is used with \mathbf{y}_{abcdef} as the number of components used to get the minimum prediction error. Here Pellai's trace is used for evaluating these model.

SOME OBSERVATIONS:

- All main effects except \mathbf{p} are significant and has large effect on both minimum number of components and prediction error.
- Position of relevant components have largest effect on prediction error. In case of minimum number of components, multicollineary also have largest effect in addition to the position of relevant components.

- However based on pillai trace statistic, Method has the lastest effect on both of the model.
- The interactions $p:\gamma$ and $\eta:relpos:Method$ is significant in prediction error model but not in minim number of components model. However, all of these interactions have small pillai statistic.
- In case of Number of components model interaction effects $\gamma:\eta:Method$, $p:Method$ and $\gamma:\eta:relpos$ are significant but not in the case of prediciton error model. Similar to previous point, they too have small pillai statistic.

6.1. Effect Analysis

ON PREDICTION ERROR MODEL:

- It would be desirable to observe effect of these interactions. Figure 6 (left) shows clear difference of η for a given $relpos$. The plot also shows a clear difference in effect of methods on prediction error.
- Figure 6 (right) shows effect of γ for a given method and η . It shows that these methods gives low prediction error is high multicollinear situations.

ON MINIMUM NUMBER OF COMPONENTS MODEL:

- Figure 7 (left) shows that $xenv$ has used minimum number of components followed by $senv$ methods than others in order to get their minimum prediction error.
- The same figure also suggest that the minimum number of components used by PLS1, PLS2 and PCR vary and has high effect of η . The $senv$ model which consider both X and Y correlation structure while estimating regression coefficients has smallest variation of number of components used for different η parameters.
- Figure 7 (right) shows that in case of low multicollinearity in the model PLS methods are used less number of components than PCR. This is expected since, PLS methods consider the covariance structure of predictor and response which PCR does not.

ADDITIONAL OBSERVATIONS:

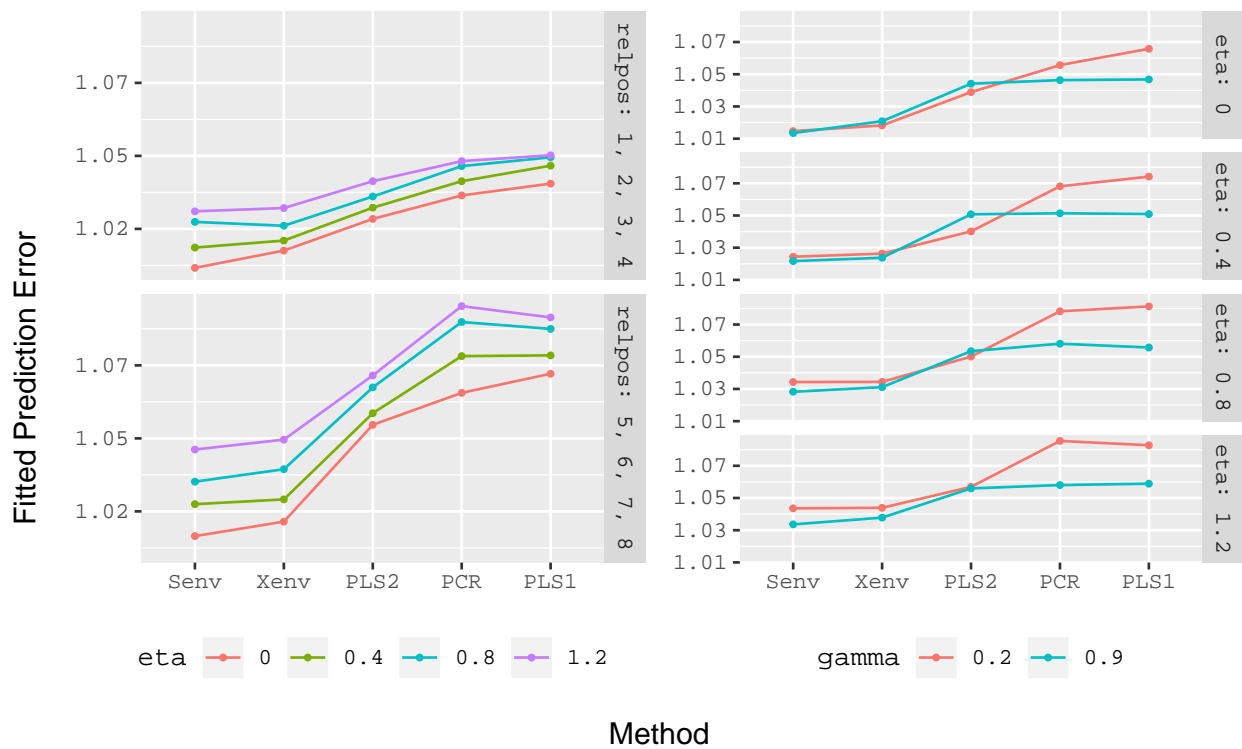


Figure 6: Effect plot of some interactions of the multivariate linear model

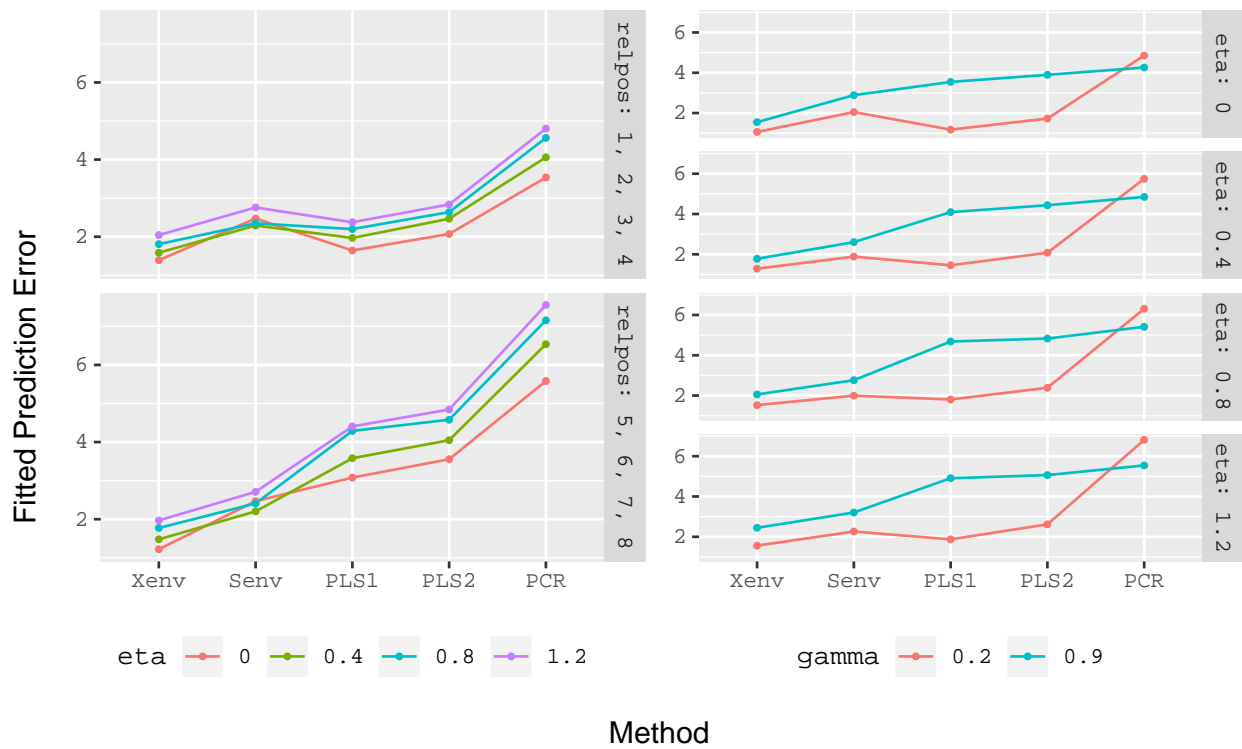


Figure 7: Effect plot of some interactions of the multivariate linear model

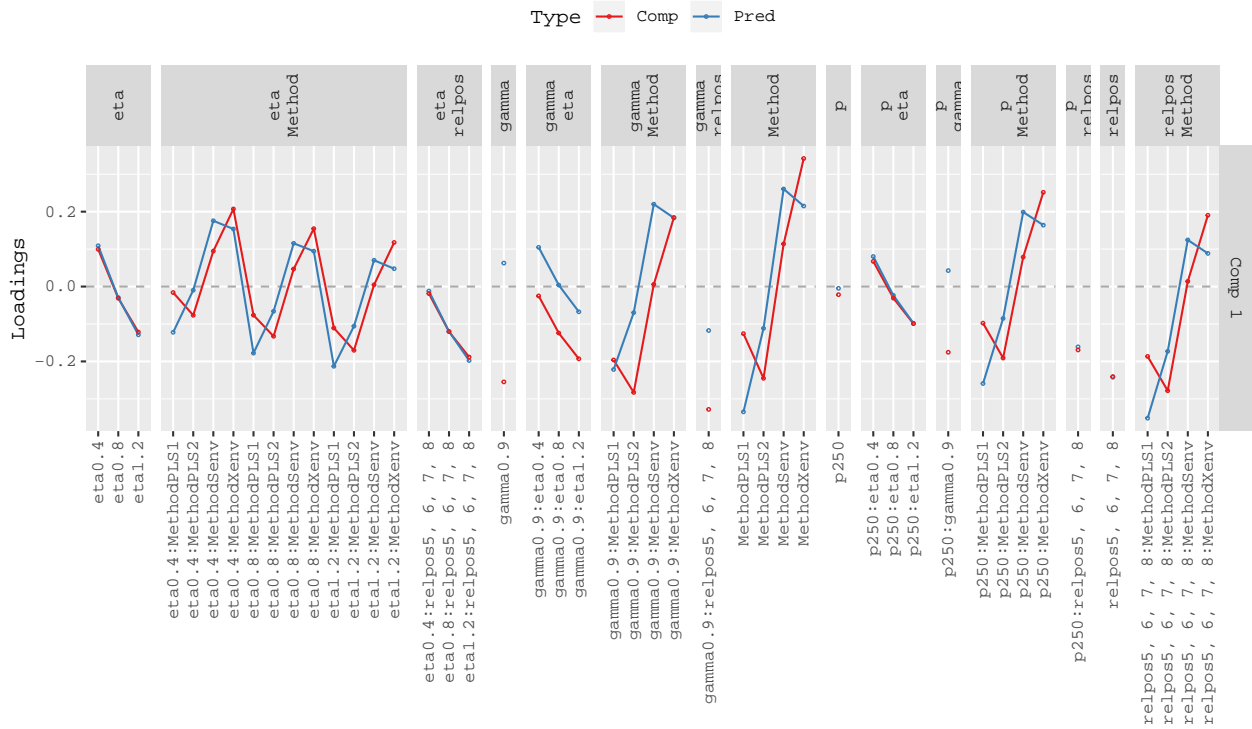


Figure 8: PLS Loadings (Component 1)

- Place for variable selection
- A PLS model is used to cross-validate the effects of these factors. The loading plot for the same model but only with second order interaction is in figure 8.

6.2. A partial least square analysis on the model

SOME OBSERVATIONS:

- The loadings for first components plotted in Figure-8 has clearly separated the envelope models from the rest by giving positive loading for them and negative for the rest.
- This components has only explained 7.969 in prediction error model and 7.697 in minimum components model.

CONFUSION:

- The explained variation by each of these components is not in decending order as

each successive components of PLS model is supposed to explain the maximum covarinace between predictor and response.

- Almøy, T., jan 1996. A simulation study on comparison of prediction methods when only a few components are relevant. *Computational Statistics & Data Analysis* 21 (1), 87–107.
- Cook, R. D., 2018. An introduction to envelopes : dimension reduction for efficient estimation in multivariate statistics, 1st Edition. Hoboken, NJ : John Wiley & Sons, 2018.
- Cook, R. D., Helland, I. S., Su, Z., 2013. Envelopes and partial least squares regression. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 75 (5), 851–877.
- Cook, R. D., Li, B., Chiaromonte, F., aug 2007. Dimension reduction in regression without matrix inversion. *Biometrika* 94 (3), 569–584.
- Cook, R. D., Li, B., Chiaromonte, F., 2010. Envelope Models for Parsimonious and Efficient Multivariate Linear Regression. *Statistica Sinica* 20 (3), 927–1010.
- Cook, R. D., Zhang, X., 2015. Simultaneous envelopes for multivariate linear regression. *Technometrics* 57 (1), 11–25.
- Cook, R. D., Zhang, X., 2016. Algorithms for Envelope Estimation. *Journal of Computational and Graphical Statistics* 25 (1), 284–300.
- de Jong, S., mar 1993. SIMPLS: An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems* 18 (3), 251–263.
- Helland, I. S., 1990. Partial least squares regression and statistical models. *Scandinavian Journal of Statistics* 17 (2), 97–114.
- Helland, I. S., mar 2000. Model Reduction for Prediction in Regression Models. *Scandinavian Journal of Statistics* 27 (1), 1–20.
- Helland, I. S., Almøy, T., 1994. Comparison of prediction methods when only a few components are relevant. *Journal of the American Statistical Association* 89 (426), 583–591.
- Helland, I. S., Sæbø, S., Almøy, T., Rimal, R., Sæbø, S., Almøy, T., Rimal, R., sep 2018. Model and estimators for partial least squares regression. *Journal of Chemometrics* 32 (9), e3044.
- Helland, I. S., Sæbø, S., Tjelmeland, H. K., mar 2012. Near Optimal Prediction from Relevant Components. *Scandinavian Journal of Statistics* 39 (4), 695–713.
- Jolliffe, I. T., 2002. *Principal Component Analysis*, Second Edition.
- Næs, T., Helland, I. S., 1993. Relevant components in regression. *Scandinavian Journal of Statistics* 20 (3), 239–250.
- Naes, T., Martens, H., jan 1985. Comparison of prediction methods for multicollinear data. *Communications in Statistics - Simulation and Computation* 14 (3), 545–576.

- Rimal, R., Almøy, T., Sæbø, S., may 2018. A tool for simulating multi-response linear model data. *Chemometrics and Intelligent Laboratory Systems* 176, 1–10.
- Sæbø, S., Almøy, T., Helland, I. S., 2015. Simrel - A versatile tool for linear model data simulation based on the concept of a relevant subspace and relevant predictors. *Chemometrics and Intelligent Laboratory Systems* 146, 128–135.