Comparison of Multivariate Estimation Methods

Raju Rimal[a,*], Trygve Almøy[a], Solve Sæbø[b]

[a]*Faculty of Chemistry and Bioinformatics, Norwegian University of Life Sciences, Ås, Norway*
[b]*Prorector, Norwegian University of Life Sciences, Ås, Norway*

Abstract

While Data science is battling to extract information from 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, these studies seldom study the impact of/on multiple response model. This study compares some newly-developed (envelope estimation) and well-established (PLS, PCR) prediction methods based on the simulated data specifically designed by varying properties such as multicollinearity, correlation between multiple response and amount of information content in latent variables. This study aims to identify itself as an additional eye for researcher to look at these methods.

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 massive explosion of data however such data often contains irrelevant information integrated within them consequently making the prediction difficult. Researchers are devising new methods and algorithms in order to extract such information to create a robust predictive model. Mostly such models contain predictor variables that are directly or indirectly correlated

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

with other predictor variables. In addition most studies constitute of many response variables correlated with each other. These interlinked relationship influences any study whether it is a predictive modeling or an 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 Experimental Design section. Sæbø et al. (2015) and Almøy (1996) have made similar comparison with 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 strategy behind the data simulation.

2. Simulation Model

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

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

where, $\boldsymbol{\Sigma}_{yy}$ and $\boldsymbol{\Sigma}_{xx}$ are the variance-covariance matrix of $\mathbf{y}$ and $\mathbf{x}$ respectively. $\boldsymbol{\Sigma}_{xy}$ is the covariance between $\mathbf{x}$ and $\mathbf{y}$. $\boldsymbol{\mu}_y$ and $\boldsymbol{\mu}_x$ are mean vector of $\mathbf{x}$ and $\mathbf{y}$ respectively. A linear

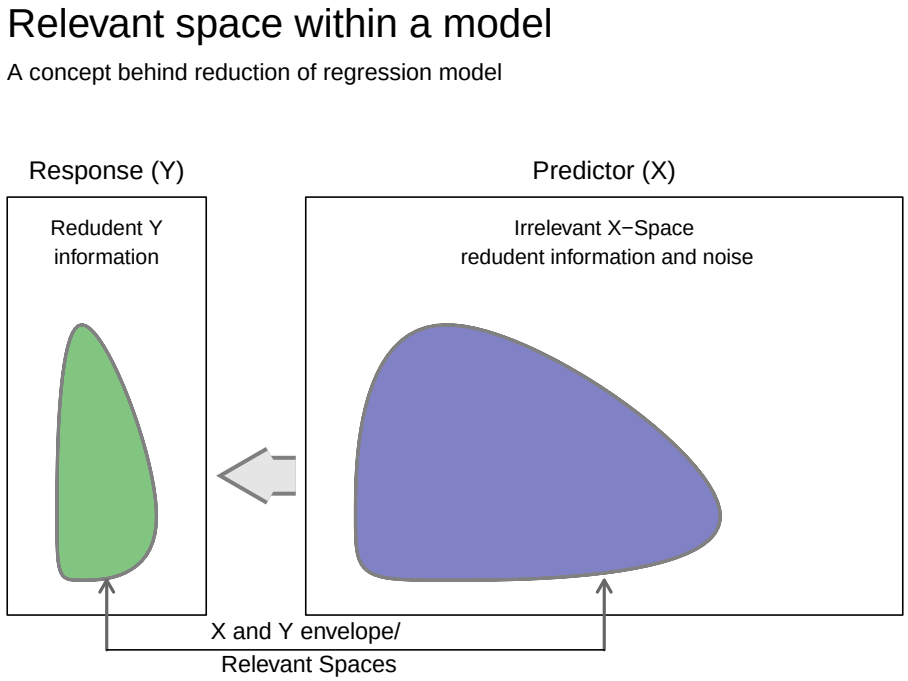

Figure 1: Relevant space in a regression model

model based on (1) is,

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

where, $\underset{m\times p}{\boldsymbol{\beta}^t}$ is a matrix of regression coefficients and ϵ is an error term such that $\epsilon \sim N(0, \Sigma_{y|x})$

In a casual relationship model like (2), we assume that the variation in response $\mathbf{y}$ is caused 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 as relevant space of $\mathbf{x}$ and rest as irrelevant or immaterial space. In the similar manner, we can assume that a subset of response space contains the information that the predictors can explain for a given model (Figure-1). Some literature such as Cook et al. (2010) and Cook and Zhang (2015) has also referred the relevant space as material space and irrelevant space as immaterial space.

With an orthogonal transformation of $\mathbf{y}$ and $\mathbf{x}$ to its latent variables $\mathbf{w}$ and $\mathbf{z}$ respectively

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

$$\mathbf{w} = \mu_w + \boldsymbol{\alpha}^t(\mathbf{z} - \mu_z) + \tau \tag{3}$$

where, $\underset{m\times p}{\boldsymbol{\alpha}^t}$ is a matrix of regression coefficients and τ is an error term such that $\epsilon \sim N(0, \Sigma_{w|z})$. Model (3) follows the distribution,

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

where, $\boldsymbol{\Sigma}_{ww}$ and $\boldsymbol{\Sigma}_{zz}$ are the variance-covariance matrix 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 response 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 then explored further by Helland (1990), Næs and Helland (1993), Helland and Almøy (1994) and Helland (2000). The corresponding eigenvectors were referred 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 span of 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 ex-

amples 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. Following section discusses about estimators under comparison in details.

3. Prediction Methods

Partial least squares regression (PLS) and Principal component regression (PCR) has been used in many discipline such as chemometrics, echonometrics, bioinformatrics and machine learning where wide predictor matrices, i.e. p (number or predictors) > n (number of observation) is common. These methods are popular in multivariate analysis specially for exploratory study and prediction. In the 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 following estimation methods based on their prediction performance in different nature of data simulated with controlled properties.

Principal Components Regression (PCR): Principal components are the linear combination 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 variant of PLS: PLS1 and PLS2 will be used for comparison. The first one consider individual response variables separately, i.e. each response are predicted with a single response model while the later consider all response variables together. In PLS regression the components are determined such as to maximize a covariance between response and predictor (de Jong, 1993).

Envelopes: Envelopes, introduced by Cook et al. (2007), was first used as response envelope (Cook et al., 2010) as a smallest subspace $\mathcal{E}$ in response space such that span of regression coefficients lies in that space. Since a multivariate linear regression model contains both 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 gaining the efficiency of the model (Cook and Zhang, 2016).

The concept was later extended to 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 envelops either of the response and predictors. 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 on wide matrices, i.e. $p > n$. In order to incorporate these method in our comparison, we have used the principal components $(\mathbf{z})$ of predictor variables $(\mathbf{x})$ using required number of components for capturing 97.5% of the variation in it. 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 variables $(\hat{\boldsymbol{\beta}})$ 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

Comparing prediction methods requires measurement of their prediction ability. Data with specific nature are simulated some of which are easier to predict than others. These data are simulated using the R-package `simrel` which are discussed in Sæbø et al. (2015) and Rimal et al. (2018). Here we will use four different factors: a) Number of predictors (`p`), b) Multicollinearity in predictor variables (`gamma`), c) Correlation in response variables (`eta`) and d) position of predictor components relevant for the response (`relpos`). Using two levels of `p`, `gamma` and `relpos` and four levels of `eta`, 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 `simrel` function.

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

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

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

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

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

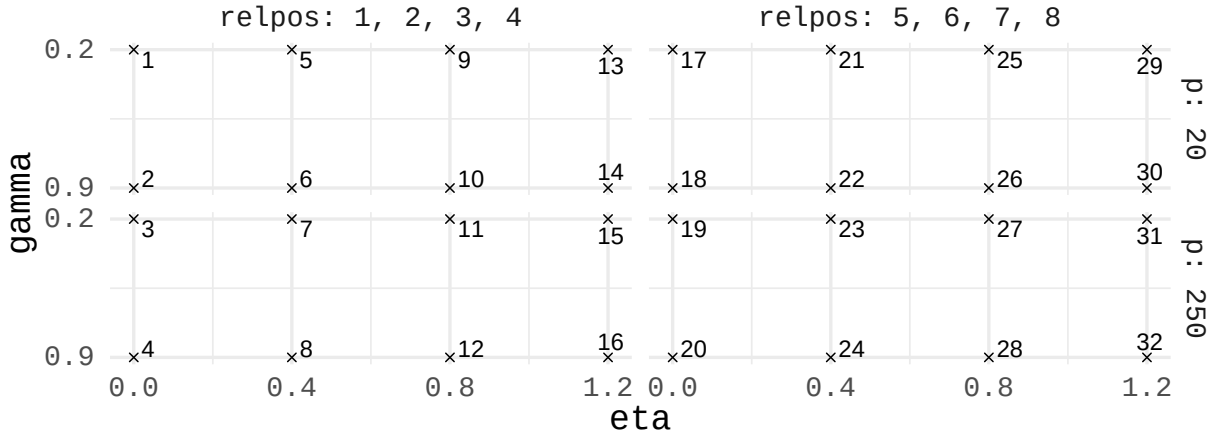

Figure 2: Experimental Design of simulation parameters. Each point represents an unique data property.

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

Position of predictor components relevant to the response: Here principal components of predictors are ordered. In other words, the first principal components captures most of the variation in predictors and second captures the most in the rest and so on. In highly collinear predictors, the variation captured by first few components are relatively high. However, if those components are not relevant for the response, prediction becomes difficult. Here, two levels of position of these relevant components are used as 1, 2, 3, 4 and 5, 6, 7, 8.

Further, a complete factorial design from the levels of above parameters gave us 32 designs. Each design is associated with a dataset having unique properties. Figure~2, shows all the design obtained from above factors. 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 corresponding to each response components in all the designs is set to and 0.8. In addition, we have assumed that there is only 1 number of informative response component. So, that the first informative response component is rotated orthogonally together with 3 uninformative response components. This spread out the information in all simulated response variables. For further details on the simulation tool see: (**?**).

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

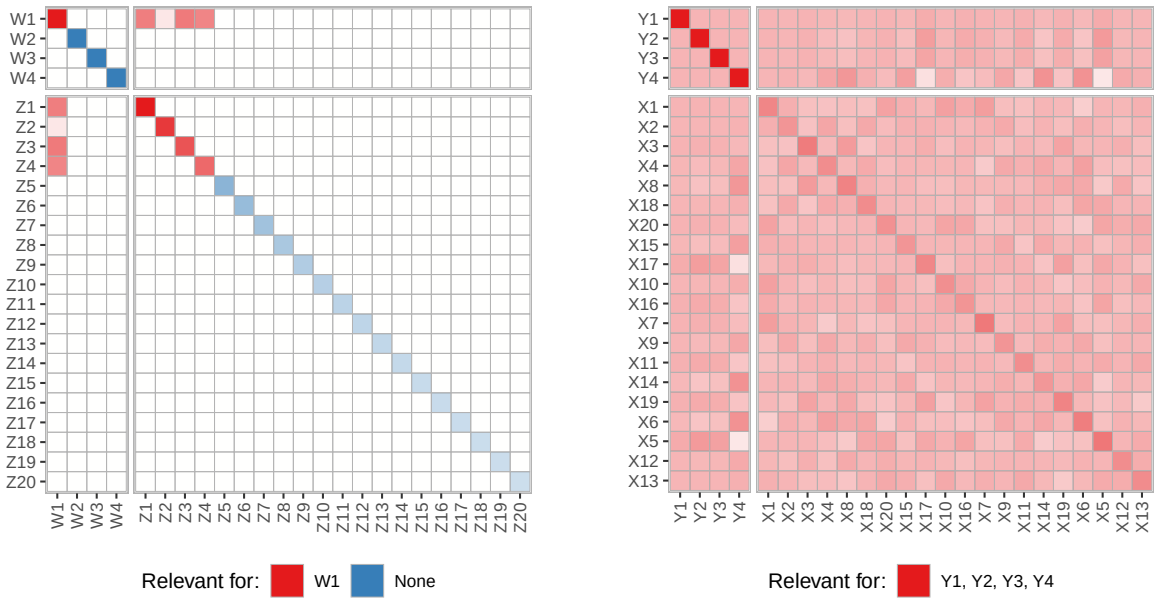

Figure 3: (left) Covariance structure of latent components. (right) Covariance structure of predictor and response

together with uninformative response components, all response variables in 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 expected prediction when the model is applied to unseen observations corre-

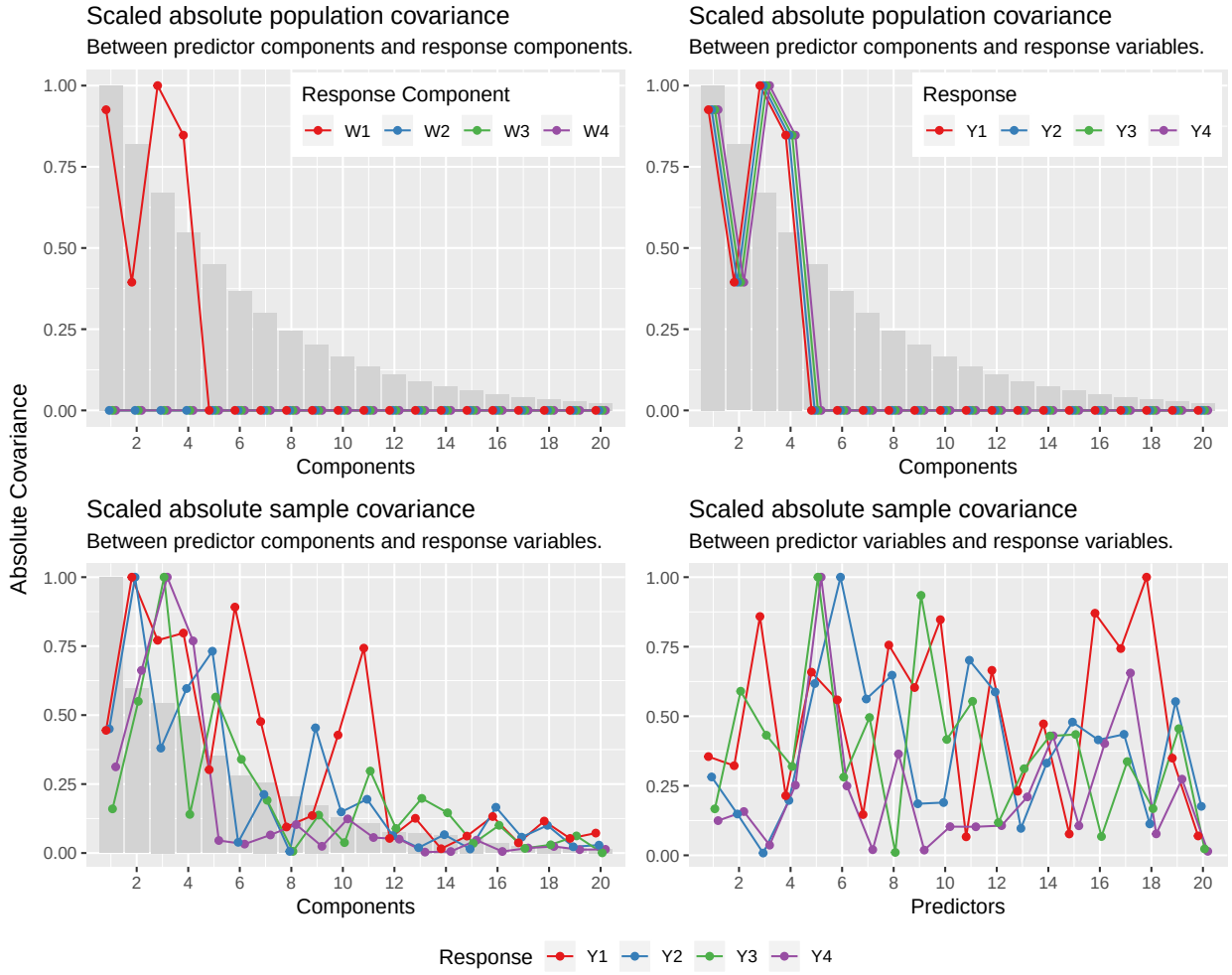

Figure 4: Expected Scaled absolute covariance between predictor components and response components (top left). Expected Scaled absolute covariance between predictor components and response variables (top right). Sample scaled absolute covariance between predictor components and response variables (bottom 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).

sponding to each response variable.

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

where, Σ_{xx} is the true covariance matrix of predictor and $\Sigma_{y_j|x}$ is the true model error both obtained from simulation for response $j = 1, \ldots m$. Here prediction error in (7) is computed for all replications of 32 designs.

5. Exploration

Before performing any statistical analysis, this section tries to explore some observed relationship between prediction error, simulation parameters and the prediction methods. Lets us start with visualizing the principal components of prediction errors together with some of these factors. Figure 5 plots first and second principal components of minimum prediciton prediction error for every replicates of all design.
HIGHLIGHTS:

5.1. PLOT: pca-scatter (Figure 5)

- Clear indication of effect of position of relevant components on the methods. The effect seems more in case of low multicollinearity than in the case of high multicollinerity.
- Envelope methods (`senv` and `xenv`) are less affected by the `relpos` factor.

5.2. PLOT: pca-density (Figure 6)

- A similar interpretation as the previous plot can be made in the score density. In addition, higher correlation in response (controlled by `eta` 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.

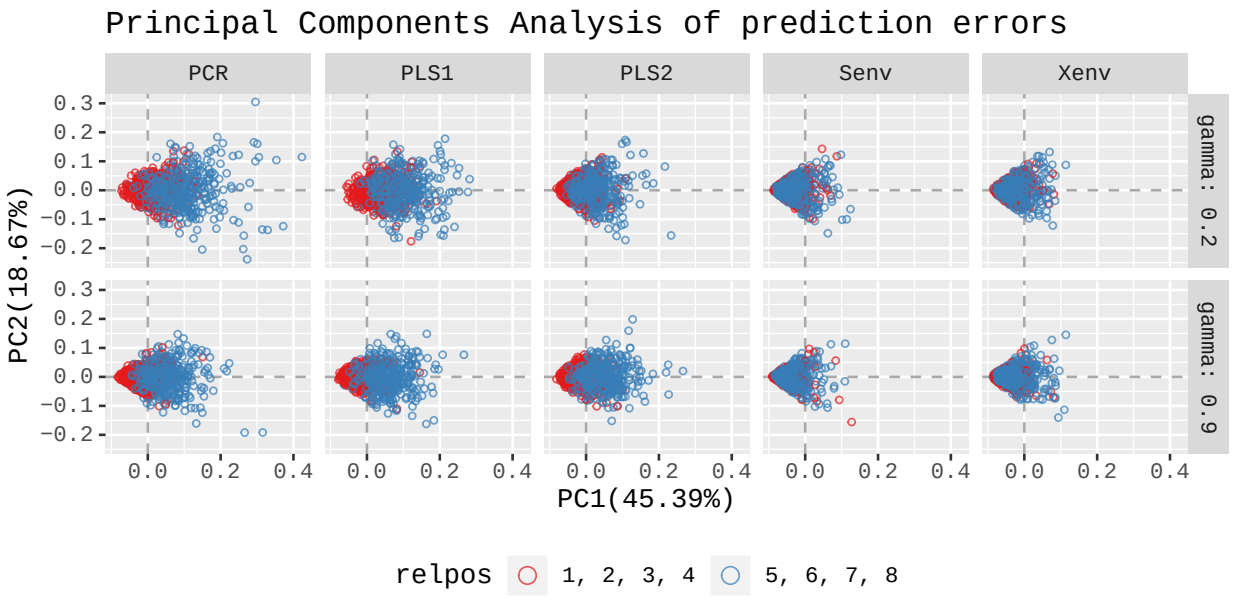

Figure 5: Exploration of Principal Components of Prediction Errors.

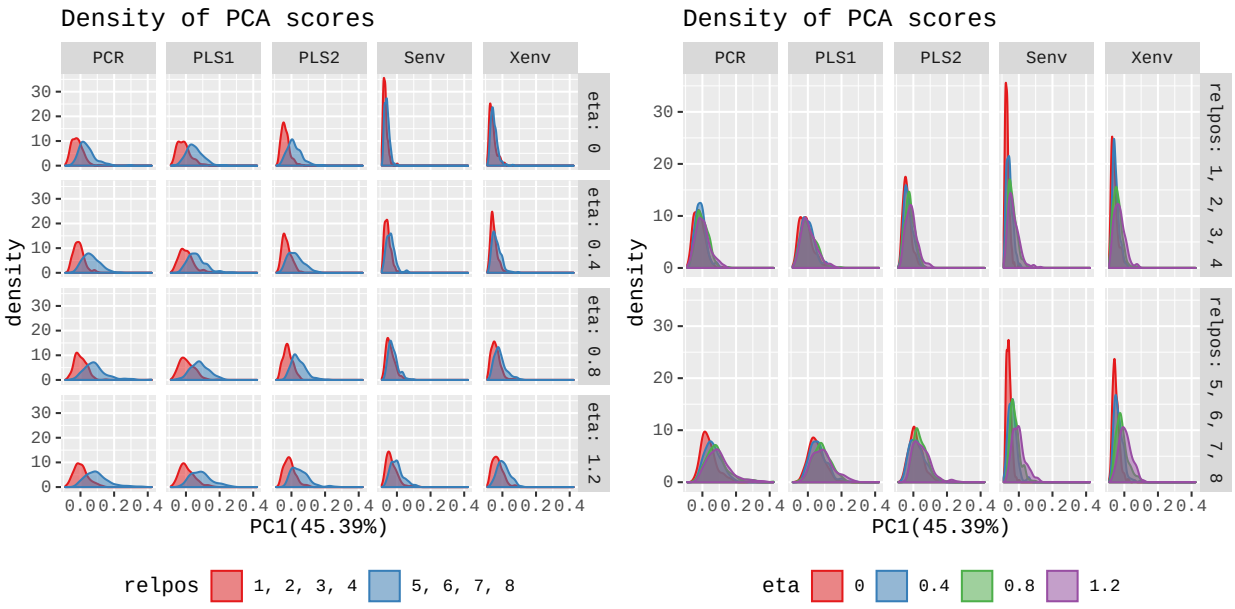

Figure 6: Density of Principal Components of Prediction Errors.

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, gamma, eta and relpos) and Methods is used as independent factors as (8).

$$\mathbf{y}_{abcdef} = \boldsymbol{\mu} + (\mathtt{p}_a + \mathtt{gamma}_b + \mathtt{eta}_c + \mathtt{relpos}_d + \mathtt{Methods}_e)^3 + \boldsymbol{\varepsilon}_{abcdef} \tag{8}$$

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

- $\mathtt{p}_a = 20$ and 250
- $\mathtt{gamma}_b = 0.2$ and 0.9
- $\mathtt{eta}_c = 0$, 0.4, 0.8 and 1.2
- $\mathtt{relpos}_d = 1, 2, 3, 4$ and $5, 6, 7, 8$
- $\mathtt{Methods}_e =$ PCR, PLS1, PLS2, Xenv and Senv

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

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

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

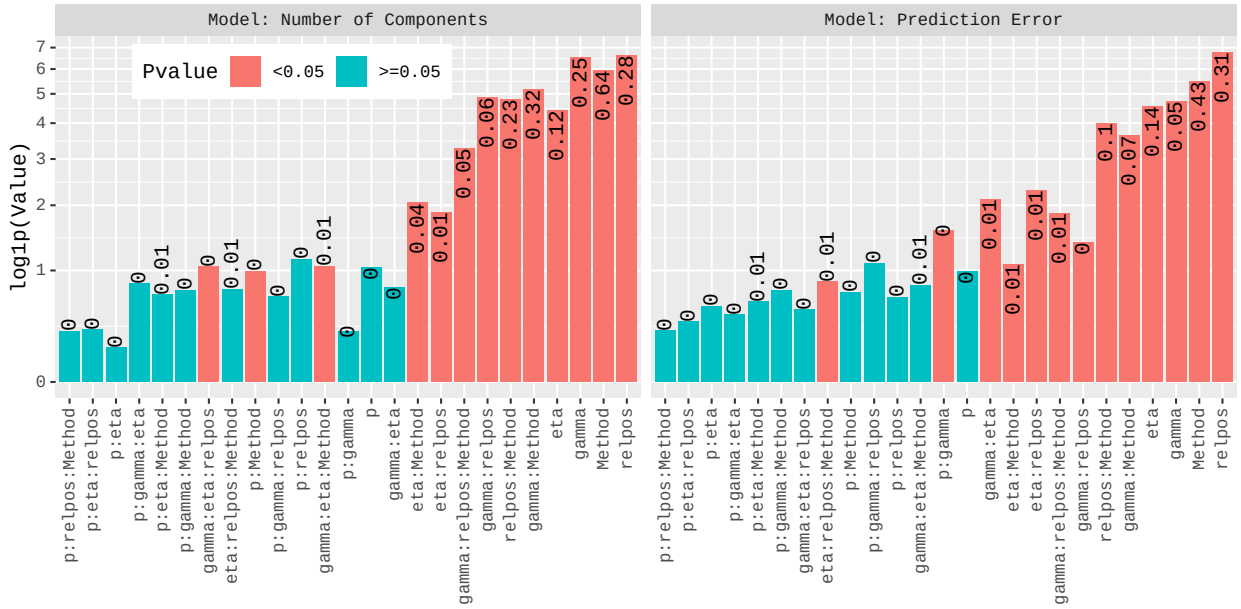

Figure 7: Pillai Statistic and F-value for the MANOVA model. The bar represents the F-value and the text labels are Pillai Statistic for corresponding factor.

- 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 8 (left) shows clear difference of `eta` for a given `relpos`. The plot also shows a clear difference in effect of methods on prediction error.
- Figure 8 (right) shows effect of `gamma` for a given `method` and `eta`. It shows that these methods gives low prediction error is high multicollinear situations.

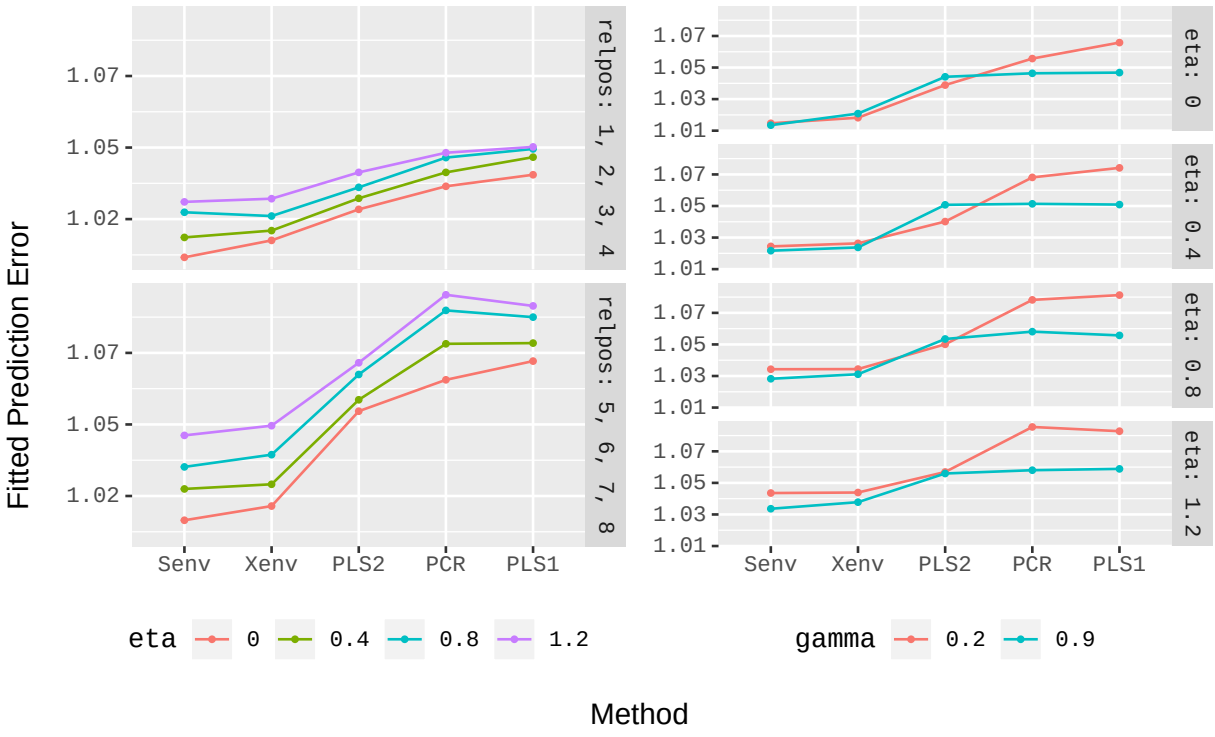

Figure 8: Effect plot of some interactions of the multivariate linear model

ON MINIMUM NUMBER OF COMPONENTS MODEL:

- Figure 9 (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 `eta`. 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 `eta` parameters.
- Figure 9 (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:

- Place for variable selection

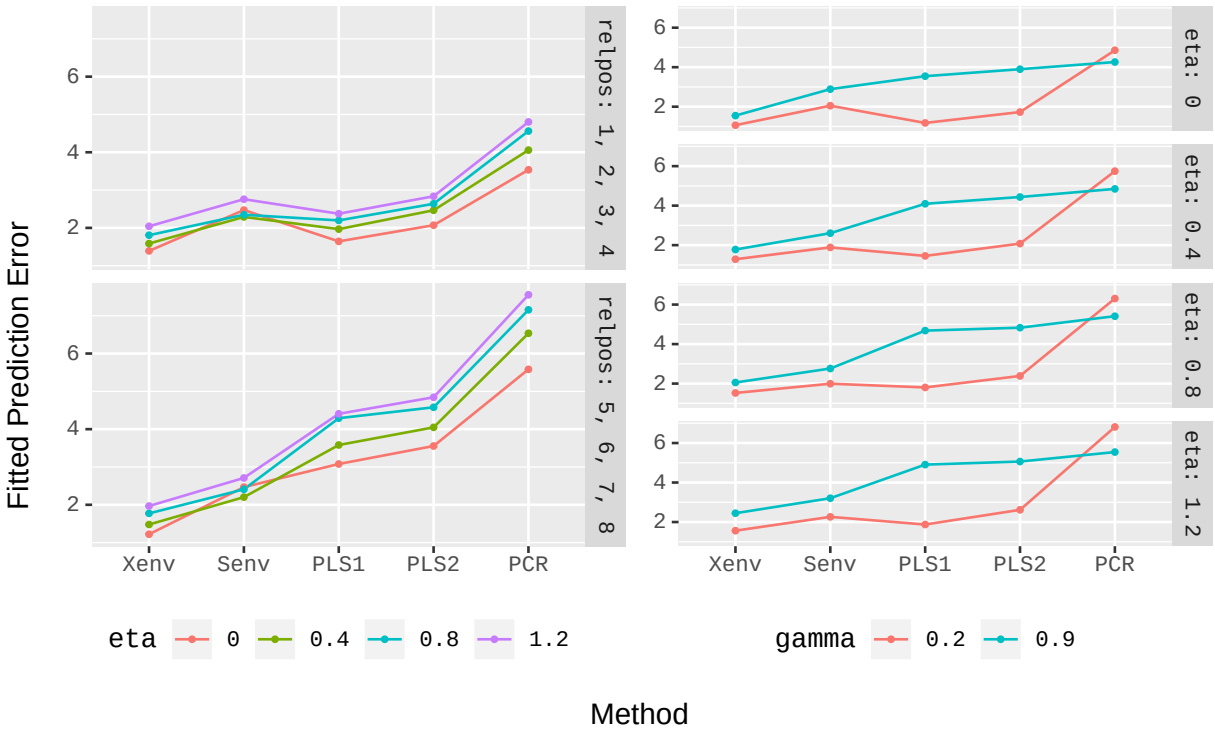

Figure 9: Effect plot of some interactions of the multivariate linear model

- 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 10.

6.2. A partial least square analysis on the model

SOME OBSERVATIONS:

- The loadings for first components plotted in Figure-10 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.

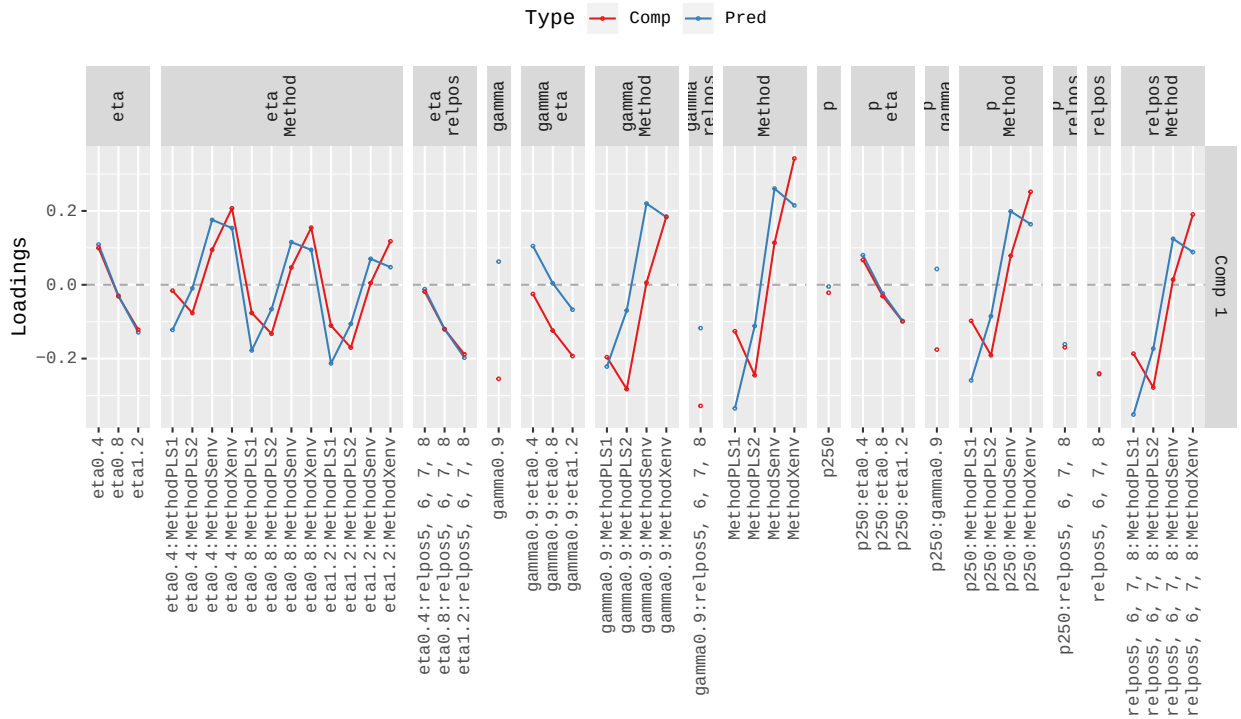

Figure 10: PLS Loadings (Component 1)

References

Almøy, T., 1996. A simulation study on comparison of prediction methods when only a few components are relevant. Computational Statistics & Data Analysis 21, 87–107. doi:doi:10.1016/0167-9473(95)00006-2.

Cook, R.D., 2018. An introduction to envelopes : dimension reduction for efficient estimation in multivariate statistics. 1 ed., 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, 851–877. doi:doi:10.1111/rssb.12018.

Cook, R.D., Li, B., Chiaromonte, F., 2007. Dimension reduction in regression without matrix inversion. Biometrika 94, 569–584. doi:doi:10.1093/biomet/asm038.

Cook, R.D., Li, B., Chiaromonte, F., 2010. Envelope Models for Parsimonious and Efficient Multivariate Linear Regression. Statistica Sinica 20, 927–1010.

Cook, R.D., Zhang, X., 2015. Simultaneous envelopes for multivariate linear regression. Technometrics 57, 11–25. doi:doi:10.1080/00401706.2013.872700.

Cook, R.D., Zhang, X., 2016. Algorithms for Envelope Estimation. Journal of Computational and Graphical Statistics 25, 284–300. doi:doi:10.1080/10618600.2015.1029577, arXiv:1403.4138.

Helland, I.S., 1990. Partial least squares regression and statistical models. Scandinavian Journal of Statistics

17, 97–114. doi:doi:10.2307/4616159.

Helland, I.S., 2000. Model Reduction for Prediction in Regression Models. Scandinavian Journal of Statistics 27, 1–20. doi:doi:10.1111/1467-9469.00174.

Helland, I.S., Almøy, T., 1994. Comparison of prediction methods when only a few components are relevant. Journal of the American Statistical Association 89, 583–591. doi:doi:10.1080/01621459.1994.10476783.

Helland, I.S., Saebø, S., Almøy, T., Rimal, R., Sæbø, S., Almøy, T., Rimal, R., 2018. Model and estimators for partial least squares regression. Journal of Chemometrics 32, e3044. doi:doi:10.1002/cem.3044.

Helland, I.S., Saebø, S., Tjelmeland, H.K., 2012. Near Optimal Prediction from Relevant Components. Scandinavian Journal of Statistics 39, 695–713. doi:doi:10.1111/j.1467-9469.2011.00770.x.

Jolliffe, I.T., 2002. Principal Component Analysis, Second Edition. doi:doi:10.2307/1270093, arXiv:arXiv:1011.1669v3.

de Jong, S., 1993. SIMPLS: An alternative approach to partial least squares regression. Chemometrics and Intelligent Laboratory Systems 18, 251–263. doi:doi:10.1016/0169-7439(93)85002-X.

Næs, T., Helland, I.S., 1993. Relevant components in regression. Scandinavian Journal of Statistics 20, 239–250.

Naes, T., Martens, H., 1985. Comparison of prediction methods for multicollinear data. Communications in Statistics - Simulation and Computation 14, 545–576. doi:doi:10.1080/03610918508812458.

Rimal, R., Almøy, T., Sæbø, S., 2018. A tool for simulating multi-response linear model data. Chemometrics and Intelligent Laboratory Systems 176, 1–10. doi:doi:10.1016/j.chemolab.2018.02.009.

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. doi:doi:10.1016/j.chemolab.2015.05.012.