Comparison of Multi-response Prediction 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 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 the 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 predictor variables that are directly or indirectly correlated with other predictor variables. In

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

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. Among others 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. In the case of PLS, we have used PLS1 which fits individual response separately and PLS2 which fits all the responses together. 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,

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

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

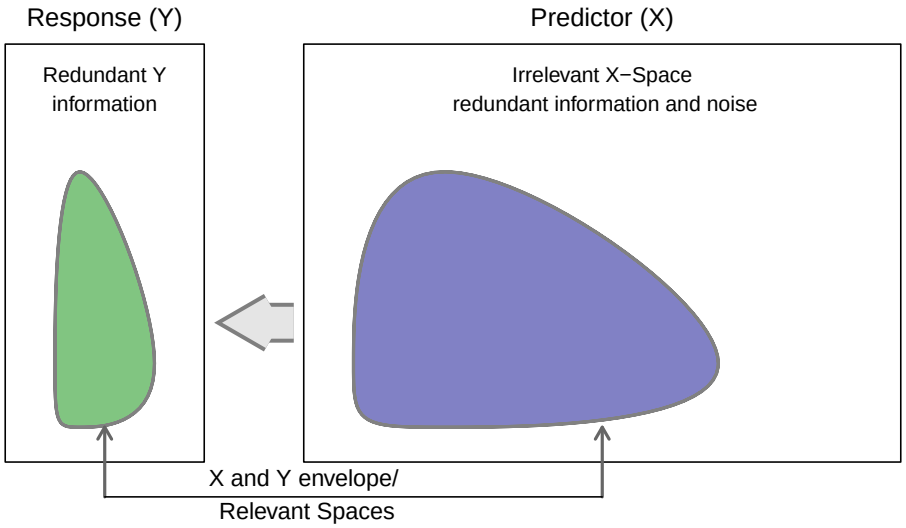

Figure 1: Relevant space in a regression model

the covariance between $\mathbf{x}$ and $\mathbf{y}$ and $\boldsymbol{\mu}_y$ and $\boldsymbol{\mu}_x$ are mean vectors of $\mathbf{x}$ and $\mathbf{y}$, respectively. A linear model based on (1) is,

$$\mathbf{y} = \boldsymbol{\mu}_y + \boldsymbol{\beta}^t(\mathbf{x} - \boldsymbol{\mu}_x) + \boldsymbol{\epsilon} \tag{2}$$

where, $\underset{m\times p}{\boldsymbol{\beta}^t}$ is a matrix of regression coefficients and $\boldsymbol{\epsilon}$ is an error term such that $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{y|x})$. Here, $\boldsymbol{\beta}^t = \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}$ and $\boldsymbol{\Sigma}_{y|x} = \boldsymbol{\Sigma}_{yy} - \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}$

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 way, for a certain model, we can assume that a subspace in the response space exists which contains the information that the relevant space in predictor can explain (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 (1) in terms of the latent variables can be written as,

$$\begin{bmatrix} \mathbf{w} \\ \mathbf{z} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_w \\ \boldsymbol{\mu}_z \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{ww} & \boldsymbol{\Sigma}_{wz} \\ \boldsymbol{\Sigma}_{zw} & \boldsymbol{\Sigma}_{zz} \end{bmatrix} \right) \tag{3}$$

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. We can write a linear model based on (3) as,

$$\mathbf{w} = \boldsymbol{\mu}_w + \boldsymbol{\alpha}^t(\mathbf{z} - \boldsymbol{\mu}_z) + \boldsymbol{\tau} \tag{4}$$

where, $\underset{m\times p}{\boldsymbol{\alpha}^t}$ is a matrix of regression coefficients and $\boldsymbol{\tau}$ is an error term such that $\boldsymbol{\tau} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{w|z})$.
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 con-

temporary 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 elaborate comparision of the prediction using multi-response simulated linear model data. The properties of the simulated data are varied through different levels of simulation parameter based on an experimental design. 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 or 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 different estimators. 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. In addition the variation of the original dataset captured by the new variables are sorted in descending order. 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 as a new predictors 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). R-package `pls` (Mevik et al., 2018) is used for both PCR and PLS methods.

Envelopes: The envelope, introduced by Cook et al. (2007), was first used to define response envelope (Cook et al., 2010) as a smallest subspace 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 envelops 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. R-package `Renvlp` (Lee and Su, 2018) is used for both Xenv and Senv methods.

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 a matrix of eigenvectors with first 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 (`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 the `simrel` function.

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

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

Here, $\lambda_i, i = 1, 2, \ldots p$ are eigenvalues of the predictor variables. Here we have used 0.2 and 0.9 as different levels of `gamma`. The higher the value of gamma, the higher will be the multicollinearity 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 `eta` (η) parameter of `simrel` for controlling the decline in eigenvalues corresponding to the response variables as (6).

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

Here, $\kappa_j, i = 1, 2, \ldots m$ 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 `eta`. The larger the value of eta, 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 as replicates. 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 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

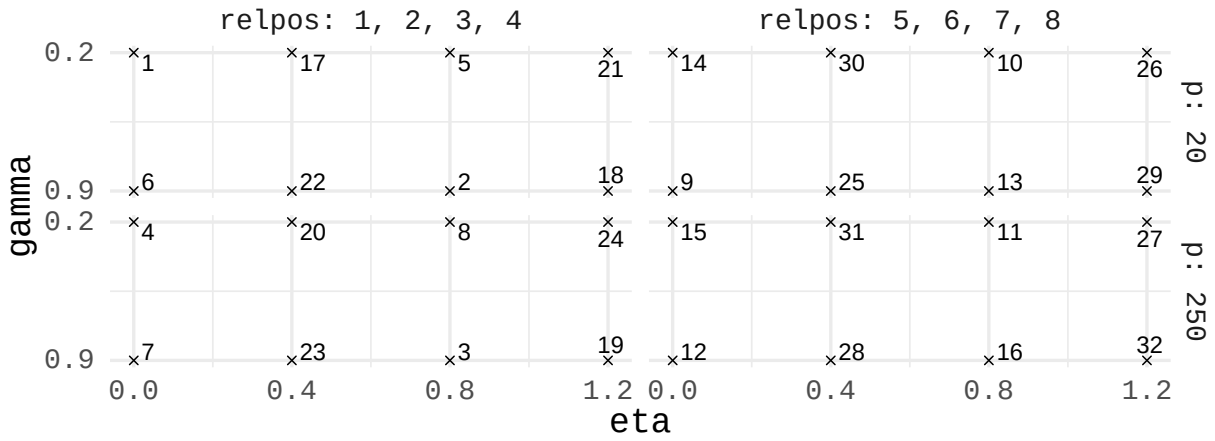

Figure 2: Experimental Design of simulation parameters. Each point represents an unique data property.

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"
)
```

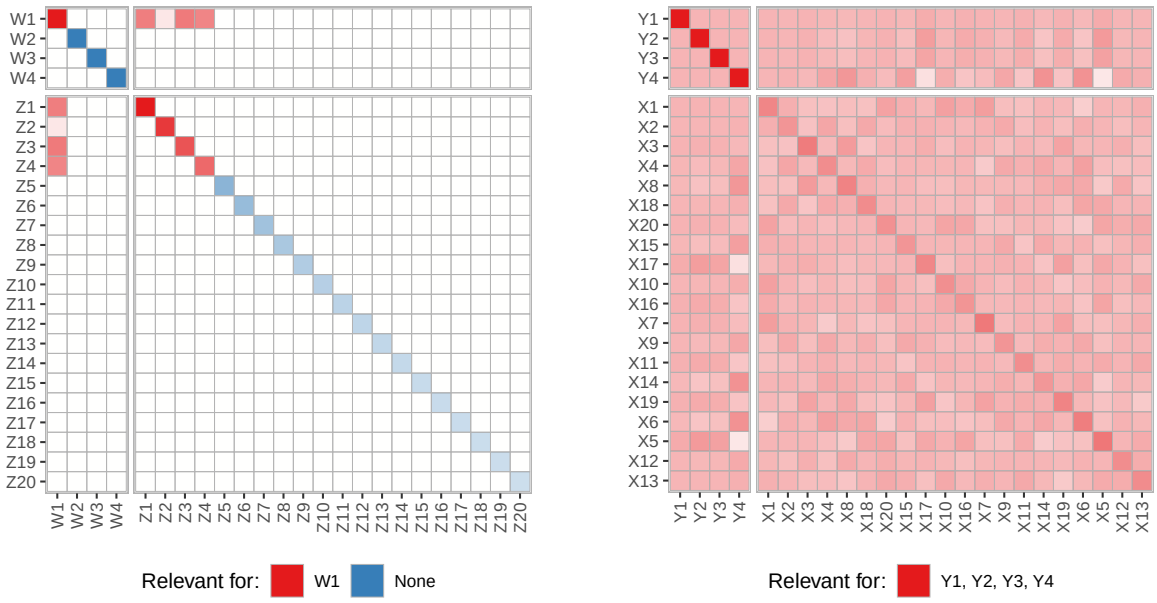

Figure 3: (left) Covariance structure of latent components. (right) Covariance structure of predictor and response

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 the 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 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. 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 method is given independent datasets

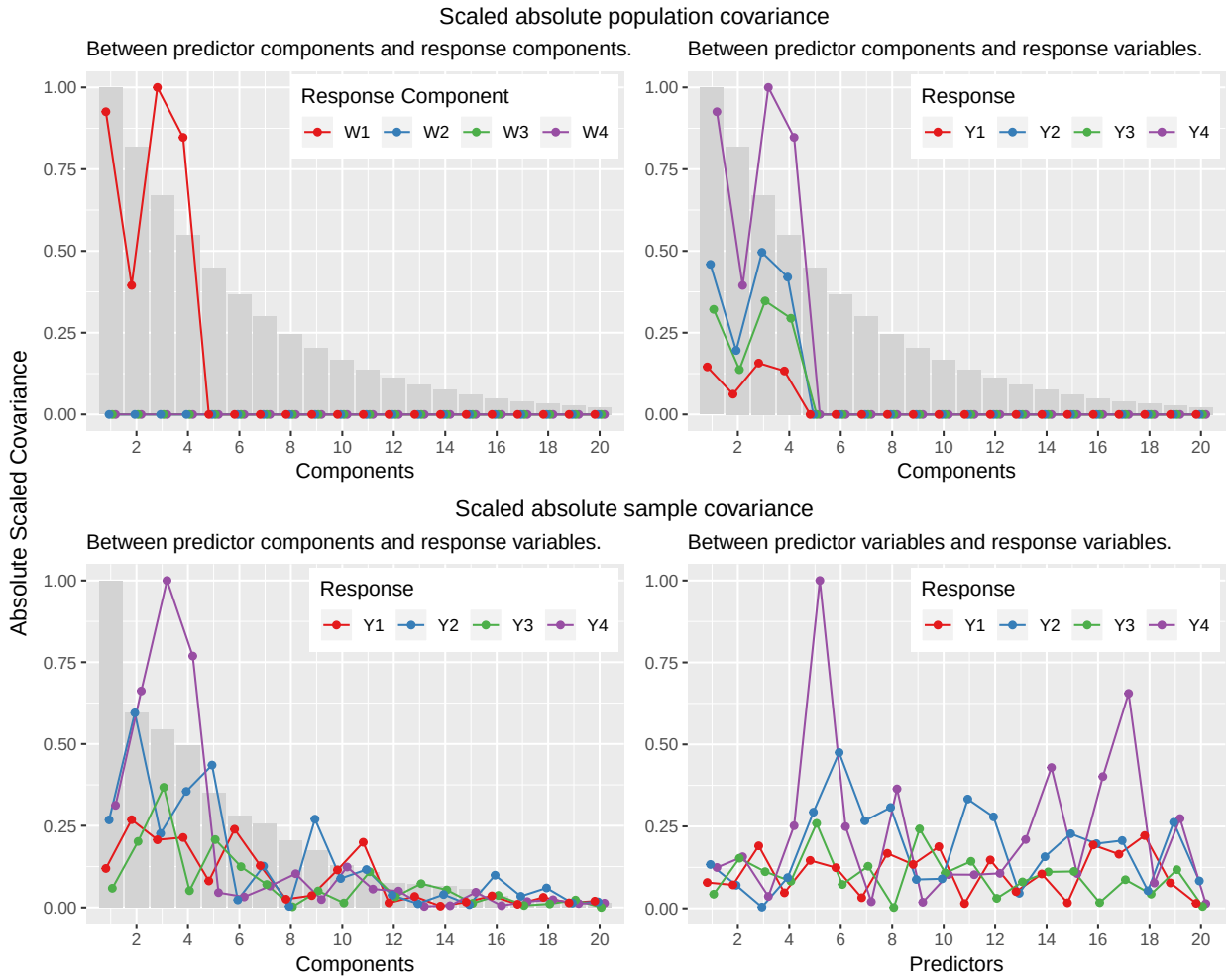

Figure 4: Expected Scaled absolute covariance between predictor components and response components (top left). Expected Scaled absolute covariance between predictor components and response variables (top right). Sample scaled absolute covariance between predictor components and response variables (bottom 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). One can compare the top-right plot (true covariance of the population) with bottom-left (covariance in the simulated data) which shows a similar pattern for different components.

simulated in order to give them equal opportunity to capture the dynamics in the data.

5. 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 on the following basis:

a) The average prediction error that a method can give using arbitrary number of components and
b) The average number of components used by the method to give the minimum prediction error

Let us define,

$$\mathcal{PE}_{ijkl} = \frac{1}{\sigma^2_{y_{ij}|x}} \mathsf{E}\left[\left(\boldsymbol{\beta}_{ij} - \hat{\boldsymbol{\beta}}_{\boldsymbol{ijkl}}\right)^t (\boldsymbol{\Sigma}_{xx})_i \left(\boldsymbol{\beta}_{ij} - \hat{\boldsymbol{\beta}}_{\boldsymbol{ijkl}}\right)\right] + 1 \tag{7}$$

as a prediction error of response $j = 1, \ldots 4$ for a given design $i = 1, 2, \ldots 32$ and method $k = 1(PCR), \ldots 5(Senv)$ using $l = 0, \ldots 10$ number of components. Here, $(\boldsymbol{\Sigma}_{xx})_i$ is the true covariance matrix of the predictors, unique for a particular design i and $\sigma^2_{y_j|x}$ for response $j = 1, \ldots m$ is the true model error. Here prediction error is scaled by the true model error to remove the effects of influencing residual variances. Since both the expectation and the variance of $\hat{\boldsymbol{\beta}}$ are unknown, the prediction error are estimated using data from 50 replications as follows,

$$\widehat{\mathcal{PE}_{ijkl}} = \frac{1}{\sigma^2_{y_{ij}|x}} \sum_{r=0}^{50} \left[\left(\boldsymbol{\beta}_{ij} - \hat{\boldsymbol{\beta}}_{\boldsymbol{ijklr}}\right)^t (\boldsymbol{\Sigma}_{xx})_i \left(\boldsymbol{\beta}_{ij} - \hat{\boldsymbol{\beta}}_{\boldsymbol{ijklr}}\right)\right] + 1 \tag{8}$$

where, $\widehat{\mathcal{PE}_{ijkl}}$ is the estimated prediction error averaged over $r = 50$ replicates.
Following section focuses on the data for the estimation of these prediction error that are used for the two models discussed above in a) and b) of this section.

6. Data Preparation

A dataset for estimating (7) is obtained from simulation which contains five factors corresponding to simulation parameters, prediction methods, number of components, replications and prediction error for four responses. The prediction error is computed using 0 to 10 predictor components for each 50 replicates as,

$$\left(\widehat{\mathcal{PE}_\circ}\right)_{ijklr} = \frac{1}{\sigma^2_{y_{ij}|x}}\left[\left(\boldsymbol{\beta}_{ij} - \hat{\boldsymbol{\beta}}_{ijklr}\right)^t \left(\boldsymbol{\Sigma}_{xx}\right)_i \left(\boldsymbol{\beta}_{ij} - \hat{\boldsymbol{\beta}}_{ijklr}\right)\right] + 1$$

.

Thus there are 32 (design) $\times$ 5 (methods) $\times$ 11 (number of components) $\times$ 50 (replications), i.e. 88000 observations corresponding to the response variables from `Y1` to `Y4`.
Since we will focus our discussion on the average minimum prediction error that a method can obtain and the average number of components they use to get the minimum prediciton error in each replicates, the dataset discussed above is summarized to construct following two smaller datasets. Let us call them *Error Dataset* and *Component Dataset*.

***Error Dataset*:** For each prediction method, design and response, an average prediction error is computed over all replicates for each components. Next, a component that gives the minimum of this average prediction error is selected, i.e.,

$$l_\circ = \underset{l}{\text{argmin}}\left[\frac{1}{50}\sum_{i=1}^{50}\left(\mathcal{PE}_\circ\right)_{ijklr}\right] \tag{9}$$

Using the component $l_\circ$, a dataset of $\left(\mathcal{PE}_\circ\right)_{ijkl_\circ r}$ is used as the *Error Dataset*. Let $\mathbf{u}_{(8000\times 4)} = (u_j)$ for $j = 1, \ldots 4$ be the outcome variables measuring the prediction error corresponding to the response number j in the context of this dataset.

***Component Dataset*:** The component number that gives the minimum prediction error in each replication is referred to as the *Component Dataset*, i.e.,

$$l_\circ = \underset{l}{\operatorname{argmin}} \left[\mathcal{PE}_{ijklr}\right] \tag{10}$$

Here $l_\circ$ is the number of components that gives minimum prediction error $(\mathcal{PE}_\circ)_{ijklr}$ for design i, response j, method k and replicate r. Let $\mathbf{v}_{(8000\times 4)} = (v_j)$ for $j = 1, \ldots 4$ be the outcome variables measuring the number of components used for minimum prediction error corresponding the response j in the context of the component dataset.

7. Exploration

This section focuses on exploring the variation in the *error dataset* and the *component dataset* for which we will use Principal Component Analysis (PCA). Let $\mathbf{t}_u$ and $\mathbf{t}_v$ be the principal component score sets corresponding to PCA run on the $\mathbf{u}$ and $\mathbf{v}$ matrices respectively. Figure-5 plots the scores density corresponding to the first principal component of $\mathbf{u}$, i.e. the first column of $\mathbf{t}_u$.

Since higher prediction error here corresponds to high scores, the plot shows that the PCR, PLS1 and PLS2 methods are influenced by the two levels of position of relevant predictor components. When the relevant predictors are at positions 5, 6, 7, 8, the eigenvalues corresponding to them are relatively smaller. This also suggest that PCR, PLS1 and PLS2 depends h[illegible]y on the position of the relevant components and the variation of these components affect their prediction performance. However, the envelope methods appe[illegible]o be less influenced by `relpos` in this regard.

In addition, the plot also shows that the effect of `gamma`, i.e., the level of multicollinearity, has smaller effect when the relevant predictors are at positions 1, 2, 3, 4. This indicates that the methods are somewhat robust to handle collinear predictors. Although, when the relevant predictors are at positions 5, 6, 7, 8, high multicollinearity results in small variance of these relevant components and consequently gives poor prediction. This is in accordance with the findings by Helland and Almøy (1994).

Further, the density curves for PCR, PLS1 and PLS2 are similar for different levels of `eta`,

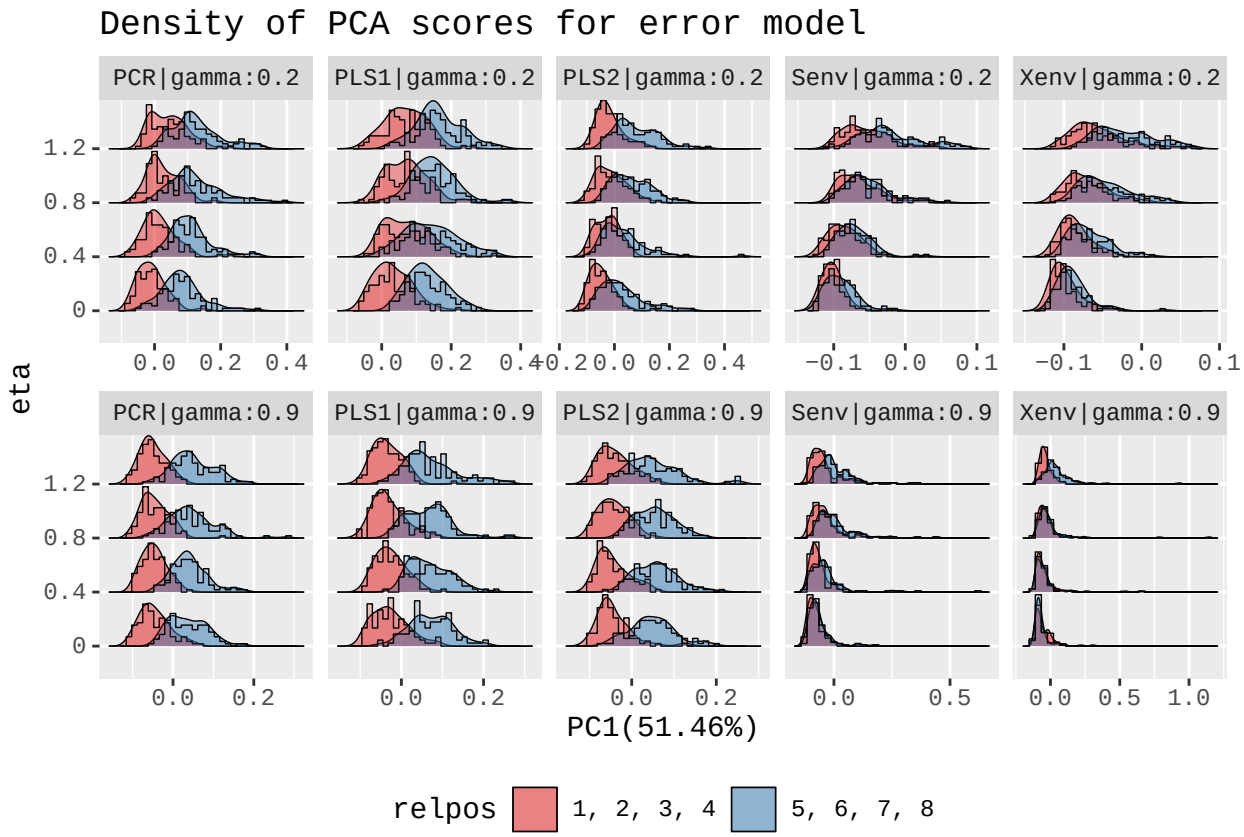

Figure 5: Scores density corresponding to first principal component of *error dataset* (**u**) subdivided by `methods`, `gamma` and `eta` and grouped by `relpos`.

i.e., the factor controlling the correlation between responses. However, this is not true for the envelope models. The envelope methods have shown to have distinct interaction between position of relevant components and `eta`. Here higher levels of `eta` has given larger scores and clear separation between two level of `relpos`.

However in the case of high multicollinearity, envelope methods have resulted in some large outliers. This suggests that in the case of multicollinearity, the methods can give unexpected prediction.

In the Figure 6, the higher scores suggest the different methods have used a large number of components to give minimum prediction error. The plot also shows that the relevant predictor components at 5, 6, 7, 8 gives larger prediction error than those which are at the position 1, 2, 3, 4. The pattern is more distinct in large multicollinearity case and PCR

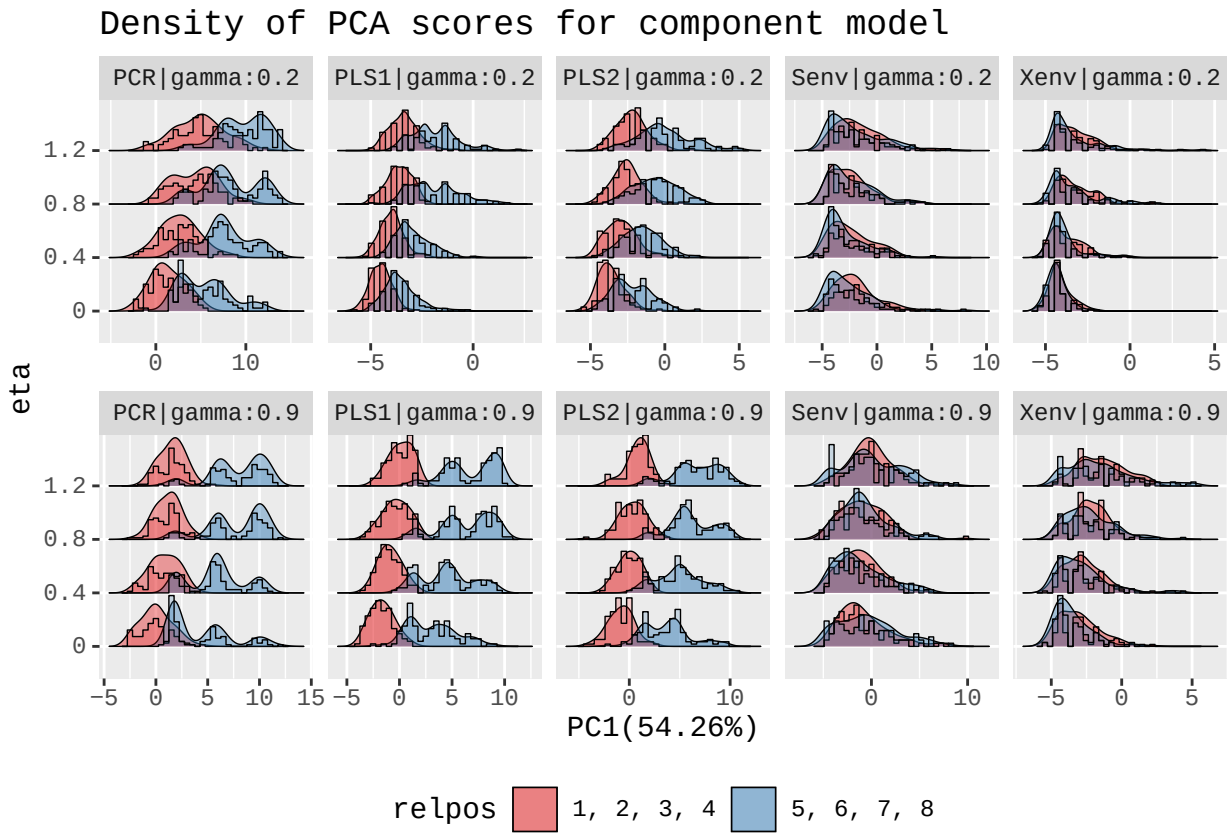

Figure 6: Score density corresponding to first principal component of *component dataset* (**v**) subdivided by `methods`, `gamma` and `eta` and grouped by `relpos`.

and PLS methods. Both the envelope methods have shown equally better performance at both levels of `relpos` and `gamma`. However, for data with low multicollinearity ($\gamma = 0.2$), the envelope methods have used fewer number of components on average than in the high multicollinear cases to achieve minimum prediction error.

8. Statistical Analysis

In this section we will model the *error data* and the *component data* as function of the simulation parameters in order to better understand the connection between data properties and prediction methods.

Let us consider a model with third order interaction of the simulation parameters (`p`, `gamma`, `eta` and `relpos`) and `Methods` as in (11) and (12) using datasets **u** and **v**, respec-

tively. Let us refer them as the *error model* and the *component model*.

Error Model:

$$\mathbf{u}_{abcdef} = \boldsymbol{\mu}_u + (\mathtt{p}_a + \mathtt{gamma}_b + \mathtt{eta}_c + \mathtt{relpos}_d + \mathtt{Methods}_e)^3 + (\boldsymbol{\varepsilon}_u)_{abcdef} \tag{11}$$

Component Model:

$$\mathbf{v}_{abcdef} = \boldsymbol{\mu}_v + (\mathtt{p}_a + \mathtt{gamma}_b + \mathtt{eta}_c + \mathtt{relpos}_d + \mathtt{Methods}_e)^3 + (\boldsymbol{\varepsilon}_v)_{abcdef} \tag{12}$$

where, $\mathbf{u}_{abcdef}$ is a vector of prediction errors in the *error model* and $\mathbf{v}_{abcdef}$ is a vector of number of components used by a method to obtain minimum prediction error in the *component model*.

Although there are several test-statistic for MANOVA, for large samples all are essentially equivalent (Johnson and Wichern, 2018). Here we will use Pillai's trace statistic which is defined as,

$$\text{Pillai statistic} = \text{tr}\left[(\mathbf{E}+\mathbf{H})^{-1}\mathbf{H}\right] = \sum_{i=1}^{m} \frac{\nu_i}{1+\nu_i} \tag{13}$$

Here the matrix $\mathbf{H}$ holds between-sum-of-squares and sum-of-products for each of the predictors. The matrix $\mathbf{E}$ has a within sum of squares and sum of products for each of the predictors. ν_i represents the eigenvalues corresponding to $\mathbf{E}^{-1}\mathbf{H}$ (Rencher, 2003).

For both the models (11) and (12), Pillai's trace statistic is used for accessing the effect of each factor and returns an F-value for the strength of their significance. Figure 7 plots the Pillai's trace statistics as bars with corresponding F-values as text labels for both models.

***Error Model*:** Figure 7 (left) shows the Pillai's trace statistic for factors of the *error model*. The main effects of `Method` has largest influence on the model followed by `relpos`, `eta` and `gamma`. A highly significant two interaction of `Method` with `eta` followed by `relpos` and `gamma` clearly shows that methods perform differently for different levels of these data properties. Further, the significant third order interaction between

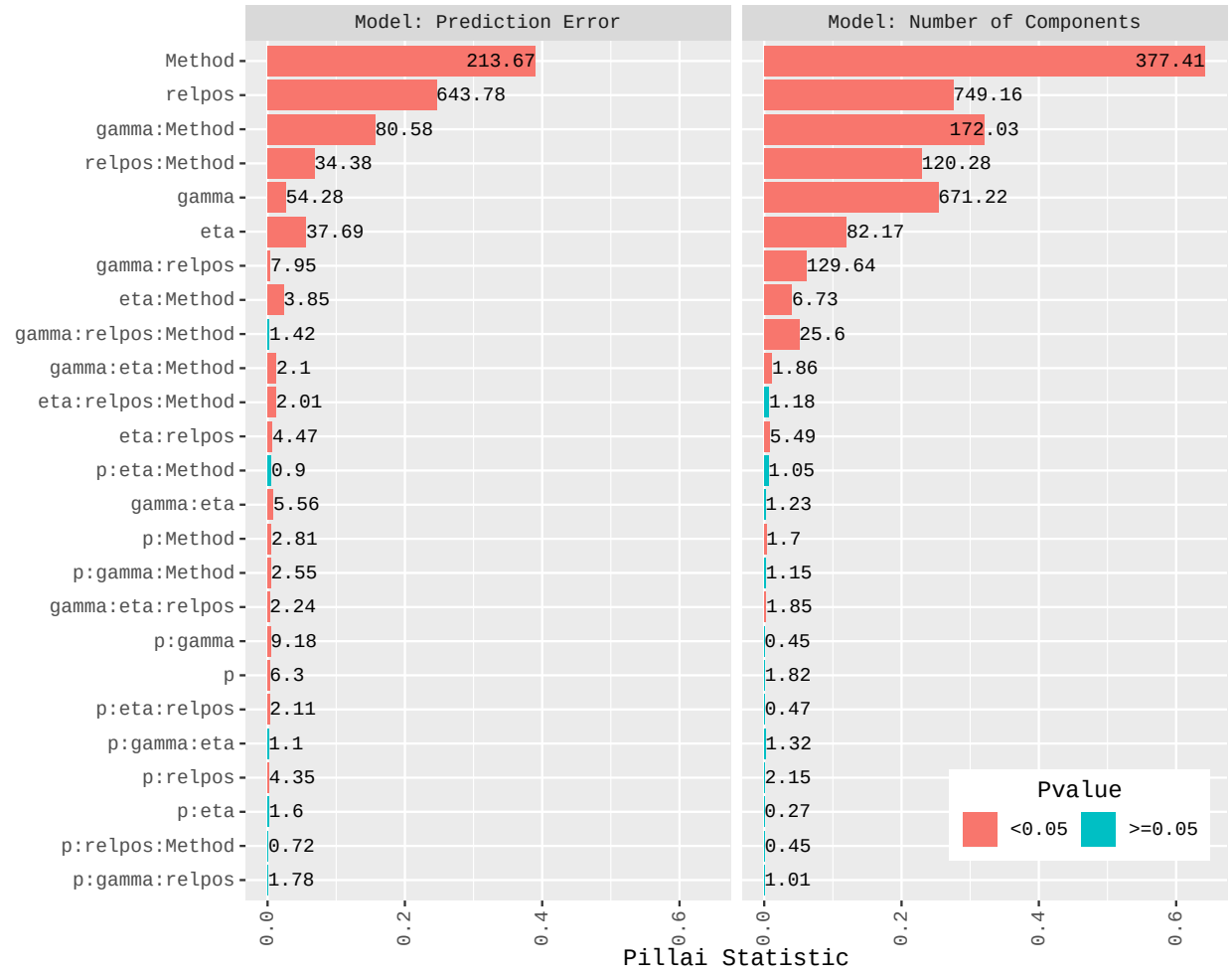

Figure 7: Pillai Statistic and F-value for the MANOVA model. The bar represents the Pillai Statistic and the text labels are F-value for corresponding factor.

`Method`, `eta` and `gamma` suggests that a method performs differently for a given level of multicollinearity and the correlation between the responses. Since, only some methods consider modelling predictor and response together, the prediction is affected by the level of correlation between the response (`eta`) for a given method.

***Component Model*:** Figure 7 (right) shows the Pillai's trace statistic for factors of the *component model*. As in the *error model*, the main effects of the Method, `relpos`, `gamma` and `eta` have significantly large effect on the number of components that a method has used to obtain minimum prediction error. The two factor interactions of `Method` with simulation parameters are larger in this case. This shows that the Methods and these interactions have larger effect on the use of number of component than the prediction error itself. In addition, a similar significant high third order interaction as in *error model* is also observed in this model.

The following section will continue exploring the effects of different levels of the factors in the case of these interactions.

8.1. Effect Analysis of Error Model

Figure 8 (left) shows that, for the envelope models, the differences in the prediction error is large. These differences are intensified when position of relevant predictor at 6, 7, 8. The results also sh that the envelope methods i re sensible to the levels of `eta` m han the rest of the methods. In the case of PCR and PLS, the difference in the effect of levels of `eta` is small.

In Figure 8 (right), we can see that the multicollinearity affected all the methods, however, PCR, PLS1 and PLS2 are more robust for the condition than the envelope methods. Rather these methods have shown better performance when high multicollinearity is present in the data. Envelope methods on the other hand are better at handling the model when relevant positions are at 5, 6, 7, 8 in both high and low multicollinearity cases.

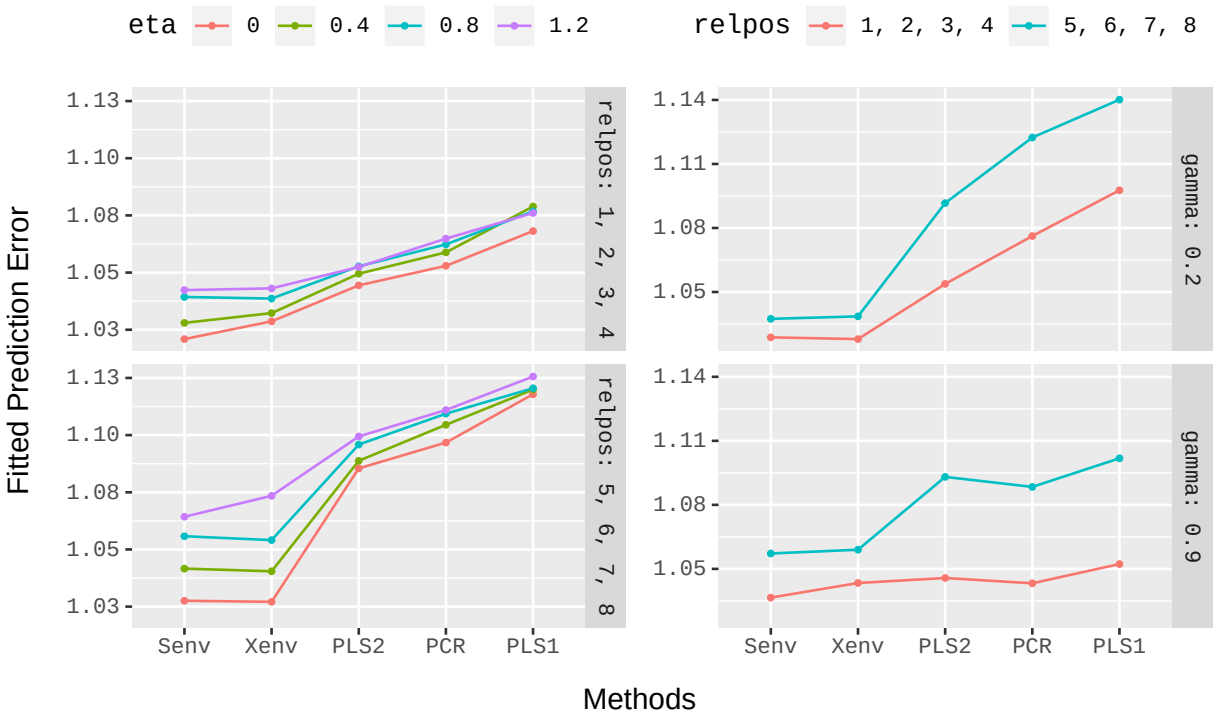

Figure 8: Effect plot of some interactions of the multivariate linear model of prediction error

8.2. *Effect Analysis of* Component Model

Unlike for prediction errors, Figure 9 (left) shows that the number of components used by the methods to obtain minium prediction error is less affected by the levels of `eta`. However, simulteneous envelope has used slightly larger number of components when there is no correlation between the response variables than the cases with moderate correlation. This pattern is distinct within the simultaneous envelope method. Envelope methods are able to obtain minimum prediction error by using components ranging from 1 to 3 in both the cases of `relpos`. This value is much higher in the case of PCR as its prediction is based only on the principal components of p ctors. The number of components used by this method ranges from 3 to 5 when relevant components are at positions 1, 2, 3, 4 and 5 to 8 when relevant components are at positions 5, 6, 7, 8.

We can also see that at `relpos` 1, 2, 3, 4 for PLS1 and PLS2 have used fewer components than simulteneous envelope. However, in the case when relevant components are at position 5, 6, 7, 8, simulteneous envelope manage to obtain smaller prediction error using

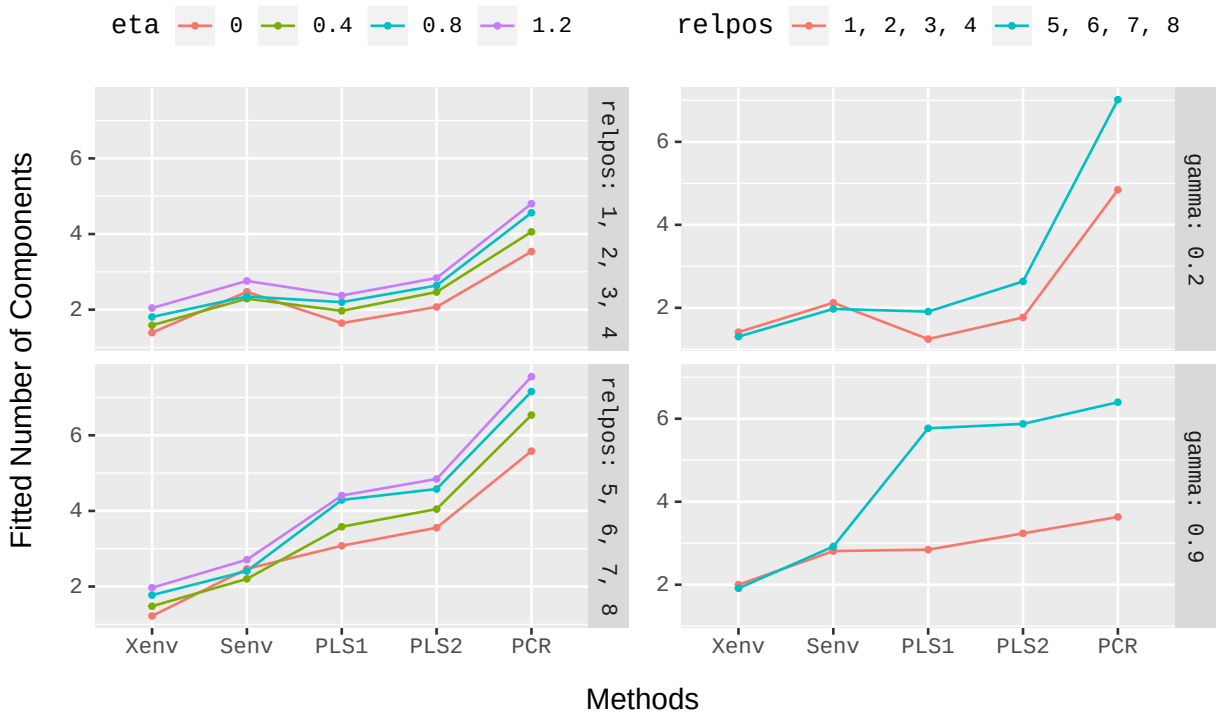

Figure 9: Effect plot of some interactions of the multivariate linear model of number of components to get minimum prediction error

fewer components than that of the PLS models. This is the case when eigenvalues of relevant predictors are small and responses are relatively difficult to predict.

With regard to the interaction effect between `gamma`, `relpos` and `Method` (Figure 9 (right)), PCR, PLS1 and PLS2 methods have used a fairly large number of components cases of high multicollinearity and relevant predictors at positions 5, 6, 7, 8. The number of components used by the envelope methods in both cases of `relpos` is similar, although slightly higher for the models with high multicollinearity than models with low multicollinearity.

In the case of `PCR`, prediction relies heavily on the position of relevant components (`relpos`). Here the method has used 3 to 5 number of components when the relevant components are at positions 1, 2, 3, 4; and 7 to 8 nber of components when the relevant components are at positions 5, 6, 7, 8. The number of components used by `PCR` is expected as in Figure 9. s reinforces the conclusion from the density plot (Figure 6)

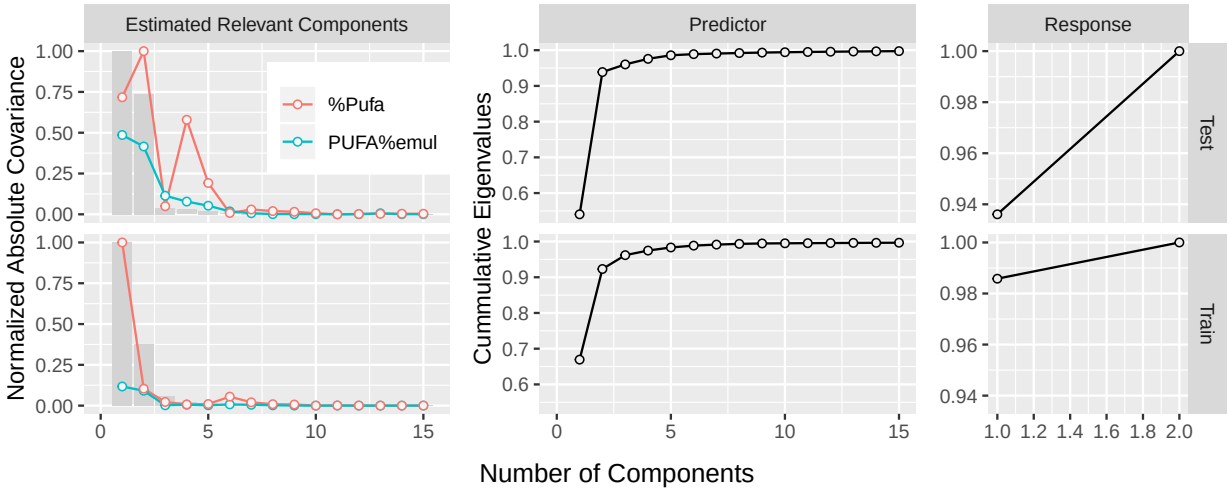

Figure 10: (Left) Bar represents the eigenvalues corresponding to Raman Spectra. The points and line are the covariance between response and the principal components of Raman Spectra. All the values are normalized to scale from 0 to 1. (Middle) Cumulative sum of eigenvalues corresponding to predictors. (Right) Cumulative sum of eigenvalues corresponding to responses.

in the previous section.

9. Examples

In addition to the analysis with the simulated data, wing two examples explores prediction performance of the methods using real datasets.

9.1. Raman spectra analysis of contents of polyunsaturated fatty acids (PUFA)

This dataset contains 44 training samples and 25 test samples of fatty acid information expressed as: a) percen total sample weight and b) percer total fat content. The dataset is us rom Næs et al. (2013) where more information can be found. The samples were analysed using Raman spectroscopy from which 1096 bles are obtained as predictors. Raman spectroscopy provi etailed chemical information from minor components in food. The aim of this example is to compare how well the prediction methods that we have considered are able to predict the contents of PUFA using these Raman spectra.

Figure 10 (left) shows that the first few predictor components are somewhat correlated with response variables. In addition the most variation in predictors are explained by

less than five components (middle). Further, the response variables are highly correlated suggesting that a single latent dimension explain most of the variation (right) is resembles wi e design 19 (Figure 2) from our simulation.

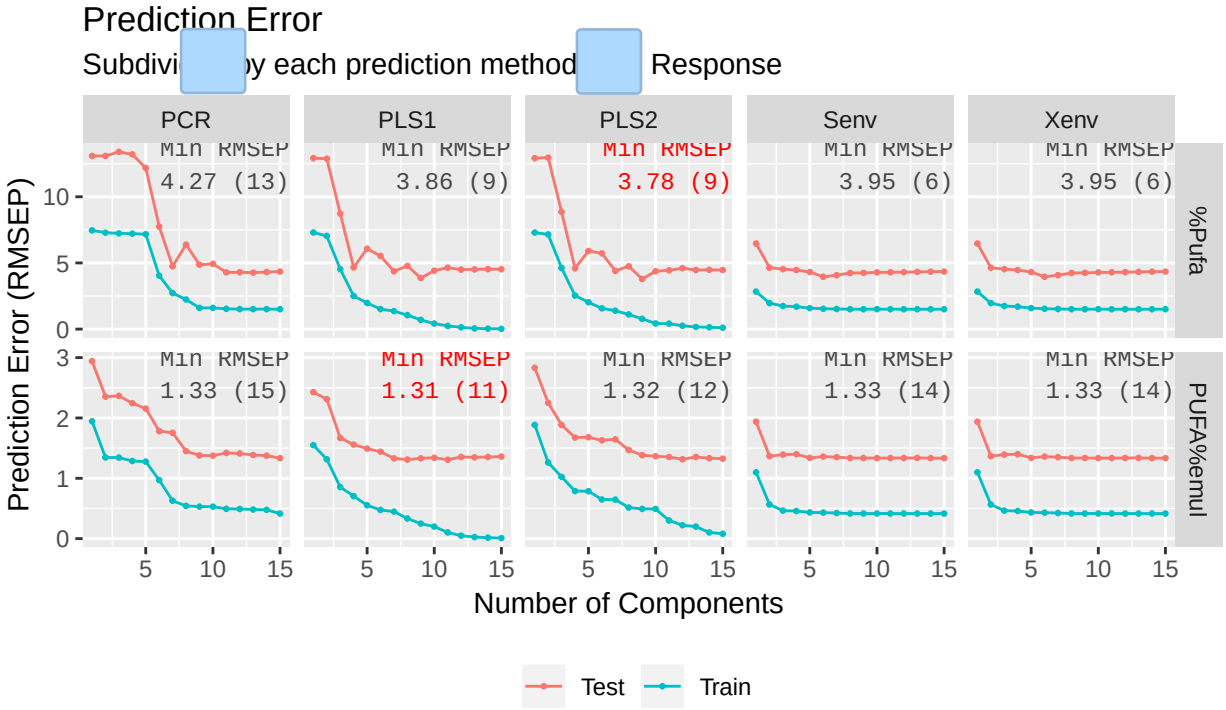

Figure 11: Prediction Error of different prediction methods using different number of components.

Using components from 1 to 15, a regression mode itted using each of the methods. The fitted models ar ed to predict the test observation and the root mean squared error of prediction (RMSEP) is culated. Figure 11 shows that PLS2 has n minimum prediction error of 3.783 using 9 components in the case of response %Pufa while PLS1 has given mum prediction error of 1.308 using 11 components in the case of response PUFA%emul. However the figure also shows that both envelope methods have reached to almost minimum predicton error in fe mber of components. This pattern is also visible in the simulation results (re 8).

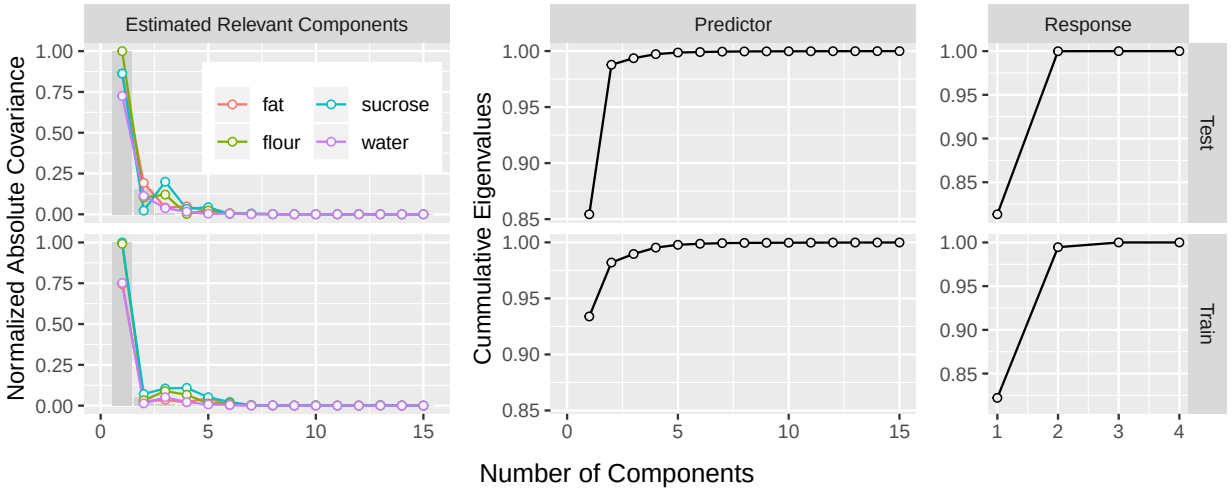

Figure 12: (Left) Bar represents the eigenvalues corresponding to NIR Spectra. The points and line are the covariance between response and the principal components of NIR Spectra. All the values are normalized to scale from 0 to 1. (Middle) Cumulative sum of eigenvalues corresponding to predictors. (Right) Cumulative sum of eigenvalues corresponding to responses.

9.2. *Example-2: NIR spectra of biscuit dough*

The dataset consists of 700 wavelengths of NIR spectra (1100–2498 nm in steps of 2 nm) which we will use as predictor variables. There are four response variables as the yield percentages of (a) fat, (b) sucrose, (c) flour and (d) water. The measurements are taken from 40 training observation of biscuit dough. A separate set with 32 samples which were created and measured on different occasions are used as test observations. The dataset is used from Indahl (2005) where further information can be obtained.

Figure 12 (left) shows that the first predictor component has largest variance and also has large covariance with all response variables. The second component, however, has larger variance (middle) than the succeeding components but has small covaration with all the response which indicates that the component is less relevant for any of the responses. In addition, two response components have explained most of the variation in response variables (right). This structure is also similar to Design 19.

Similar model is optimized as in the first example and the fitted model with each of 15 components are used for prediction. Figure 13 shows the root mean squared error for both test and train prediction. Here four different methods have minimum test prediction error

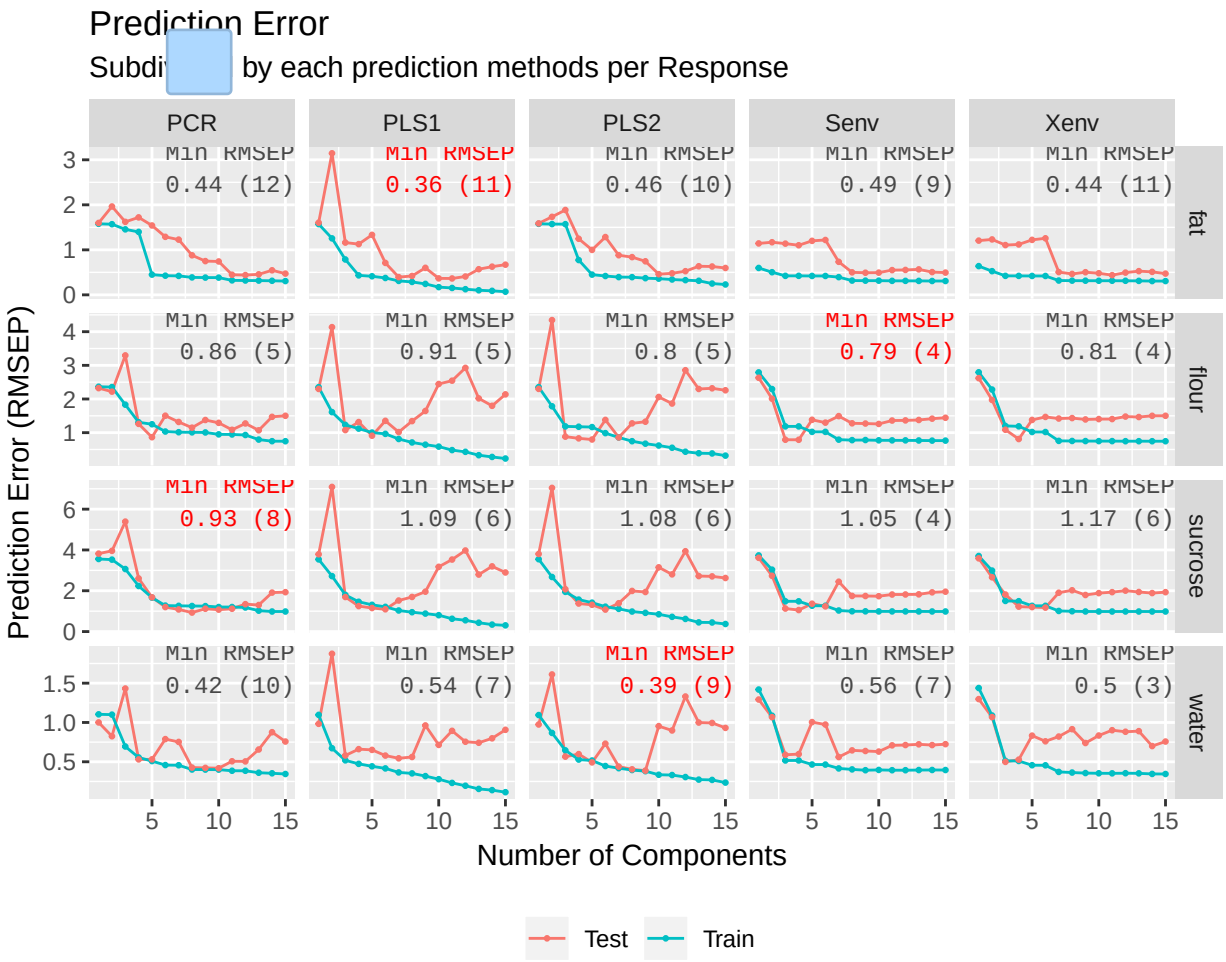

Figure 13: Prediction Error of different prediction methods using different number of components.

for responses. As the structure of the data is similar to the first example, the pattern in the prediction is also similar for all methods.

The results from both of these examples also tes the simulation results in Figure 5 which has shown large variation in prediction error in certain cases of envelope method. However this also collaborate with Figure 6 which suggests that in most of the designs, envelope methods have used less number of components to achieve the minimum prediction error.

10. Discussions and Conclusion

On one hand, the envelope methods have shown better prediction performance with fewer number of components. However, for data with high multicollinearity, the methods have shown some cases of unstable prediction. Here for the prediction error, we have used the components that give the minimum average prediction error for all replicates. This hints that for some replicates, the methods has used non-optimal number of components and consequently resulting in high prediction error. The two examples in the previous section collaborate with this point.

On the other hand, PLS1, PLS2 and PCR methods have smaller prediction error in the model with high multicollinearity which suggest their robustness in the particular case. However, they have shown poor performance in modelling information at relevant components with small variance. These methods have used larger number of components, in general, than envelopes. Although resulting higher prediction error than envelope methods in most situations, these methods are more stable especially in the cases of high multicollinearity.

Further, we have fixed the coefficient of determination (R^2) constant throughtout all the designs. Initial simulations (not shown) indicated that low R^2 affect all methods in similar manner and the MANOVA is hightly dominated by R^2. Keeping the value of R^2 fixed has allowed us to analyze other factors properly.

Two clear comment can be made about the effect of correlation of response on the prediction methods. Highly correlated response has shown highest prediction error in general and the effect is [redacted]nct in envelope methods. Since the envelope methods identifies the relevant space as the span of relevant eigenvectors, the methods are able to obtain the minimum average prediction error by using fewer number of components for all levels of `eta`.

As of our knowledge, the effect of correlation in the response on PCR and PLS methods are less explored. In this regards, it is interesting to see that these methods have used large number of components and returned large prediction error than envelope methods

in the case of highly correlated responses. In order to fully understand the effect of `eta`, it is necessary to study the estimation performance of these methods at different number of components.

In addition, since using principal components or actual variables as predictors in envelope methods have shown similar results, we have used principal components that have explained 97.5% of the variation as mentioned previously in the cases of envelope methods. As the envelope methods are based on MLE and this can be an alternative way of using the methods in data with wide predictors. The results from this study will help researcher to understand these methods for different nature of data. We encourage researcher to use newly developed methods such as envlope based on the nature of data they are working on.

Since, this study have focused entirely on prediction performance, further analysis of their estimative properties of these methods is required. A study of estimation error and the behaviour of methods on non-optimal number of components can give deeper understanding of these methods.

A shiny application (Chang et al., 2018) is availiable at `http://therimalaya.shinyapps.io/Comparison` where all the results related to this study can be visualized. In addition, a github repository at `https://github.com/therimalaya/03-prediction-comparison` can be used to reproduce this study.

11. Acknowledgment

We are grateful to Inge Helland on his inputs on this paper throughout the period. His guidance on the envelope models and review on this paper helped us to shape this paper extensively. We would gratefully like to thank Kristian Lillan, Ulf Indahl, Tormod Næs, Ingrid Måge and the team for providing the data for analysis.

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:10.1016/0167-9473(95)00006-2.

Chang, W., Cheng, J., Allaire, J., Xie, Y., McPherson, J., 2018. shiny: Web Application Framework for R. URL: https://CRAN.R-project.org/package=shiny. r package version 1.2.0.

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:10.1111/rssb.12018.

Cook, R.D., Li, B., Chiaromonte, F., 2007. Dimension reduction in regression without matrix inversion. Biometrika 94, 569–584. 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: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: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:10.2307/4616159.

Helland, I.S., 2000. Model Reduction for Prediction in Regression Models. Scandinavian Journal of Statistics 27, 1–20. 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: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: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:10.1111/j.1467-9469.2011.00770.x.

Indahl, U., 2005. A twist to partial least squares regression. Journal of Chemometrics 19, 32–44. doi:10.1002/cem.904.

Johnson, R., Wichern, D., 2018. Applied Multivariate Statistical Analysis (Classic Version). Pearson Modern Classics for Advanced Statistics Series, Pearson Education Canada. URL: https://books.google.no/books?id=QBqlswEACAAJ.

Jolliffe, I.T., 2002. Principal Component Analysis, Second Edition. 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:10.1016/0169-7439(93)85002-X.

Lee, M., Su, Z., 2018. Renvlp: Computing Envelope Estimators. URL: https://CRAN.R-project.org/package=Renvlp. r package version 2.5.

Mevik, B.H., Wehrens, R., Liland, K.H., 2018. pls: Partial Least Squares and Principal Component Regression. URL: https://CRAN.R-project.org/package=pls. r package version 2.7-0.

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:10.1080/03610918508812458.

Næs, T., Tomic, O., Afseth, N.K., Segtnan, V., Måge, I., 2013. Multi-block regression based on combinations of orthogonalisation, pls-regression and canonical correlation analysis. Chemometrics and Intelligent Laboratory Systems 124, 32–42.

Rencher, A.C., 2003. Methods of multivariate analysis. volume 492. John Wiley & Sons.

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: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:10.1016/j.chemolab.2015.05.012.