# 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**

Prediction performance often does not always reflect the estimation behaviour of a method. High error in estimation not necessarily results in high prediction error but can lead to an unreliable prediction when test data are in a different direction than the training data. In addition, poor estimation error leads to unstable estimates, and consequently the effect of predictors on the response can not have a valid interpretation. Many research fields are more interested in the effect of predictor variables than performing prediction. This study compares some newly-developed (envelope) and well-established (PCR, PLS) estimation methods using simulated data with specifically designed properties such as: multicollinearity in the predictor variables, the correlation between multiple responses and the position of principal components corresponding to predictors that are relevant for the response. This study aims to give some insights into these methods and help the researchers to understand and use them for further study. Here we have found that no single method is superior to others, but each has their strength on some specific nature of data. In addition, newly developed envelope method has shown the impressive results in finding the relevant information out of the data using significantly few numbers of components.

*Keywords:* model-comparison,multi-response,simrel,estimation,estimation error,meta modeling

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

## 1. Introduction

Estimation of parameters in linear regression models is an integral part of many research studies. Research fields such as social science, econometrics, chemometrics, psychology and medical study are more interested in measuring the impact of certain indicators or variable rather than performing prediction. Such studies have a large influence on people's perception and also help in policy making and decisions. A transparent, valid and robust research is critical in order to improve the trust in the findings of modern data science research (High-Level Expert Group on Artificial Intelligence, 2019). It makes the assessment of the error of the measurement, inference and prediction even more essential. Technology has facilitated researchers to collect large amounts of data, however often times, such data either contains irrelevant information or are highly collinear. Researchers are devising new estimators to extract information and identify their inter-relationship. Some estimators are robust towards fixing the multicollinearity problem, while some are targeted to model only the relevant information contained in the response variable.

This study extends the (Rimal et al., 2019) with similar multi-response linear regression model setting and compares some well-established estimators such as Principal Components Analysis (PCR), Partial Least Squares (PLS) together with two new methods based on envelope estimation: Envelope estimation in predictor space (Xenv) (Cook et al., 2010) and simultaneous estimation of the envelope (Senv) (Cook and Zhang, 2015). The estimation process of these methods is discussed in Estimation Methods section. The comparison tests the estimation performance of these methods using multi-response simulated data from a linear model with controlled properties. The properties include the number of predictors, level of multicollinearity, the correlation between different response variables and the position of relevant predictor components. These properties are explained in Experimental Design section together with the strategy behind the simulation and data model.

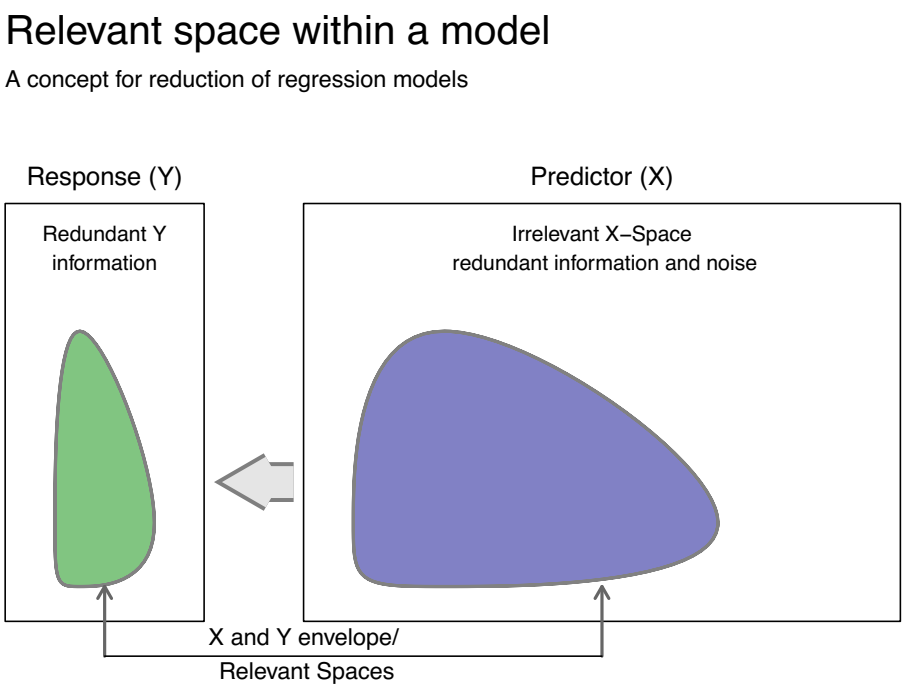


Figure 1: Relevant space in a regression model

## 2. Simulation Model

As a follow-up, this study will continue using the same simulation model used by Rimal et al. (2019). The data are simulated from a multivariate normal distribution where we assume that the variation in response variable $\mathbf{y}$ is partly explained by the predictor variable $\mathbf{x}$. 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 a similar way, for a certain model, we can assume that a subspace in the response space exists and contains the information that the relevant space in predictor can explain (Figure 1).

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, , p.101). See Rimal et al. (2018), Sæbø et al. (2015) and Rimal et al. (2019) for in-depth background on the model.

## 3. Estimation Methods

Consider a joint distribution of $\mathbf{y}$ and $\mathbf{x}$ with corresponding mean vectors $\boldsymbol{\mu}_y$ and $\boldsymbol{\mu}_x$ as,

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

Here, $\boldsymbol{\Sigma}_{xx}$ and $\boldsymbol{\Sigma}_{yy}$ are variance-covariance of $\mathbf{x}$ and $\mathbf{y}$ respectively and $\boldsymbol{\Sigma}_{xy} = \boldsymbol{\Sigma}_{yx}^t$ is the covariance matrix of $\mathbf{x}$ and $\mathbf{y}$. Let $\mathbf{S}_{xx}$, $\mathbf{S}_{yy}$ and $\mathbf{S}_{xy} = \mathbf{S}_{yx}^t$ be the respective estimates of these matrices. A linear regression model based on (1) is

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

where $\boldsymbol{\beta} = \boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}$ is the regression coefficients that defines the relationship between $\mathbf{x}$ and $\mathbf{y}$. With $n$ samples, the least squares estimate of $\boldsymbol{\beta}$ can be written as $\hat{\boldsymbol{\beta}} = \mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}$. Here, in many situations $\mathbf{S}_{xx}$ can either non-invertible or has small eigenvalues. In addition, $\mathbf{S}_{xy}$ is influenced by high level of noise in the data. In order to solve these problems, various methods have used the concept of relevant space to identify the relevant components through the reduction of the dimension in either $\mathbf{x}$ or $\mathbf{y}$ or both. Some of the methods we have used for comparison are discussed below.

*Principal Components Regression* (PCR) uses $k$ eigenvectors of $\mathbf{S}_{xx}$ as the number of components or the number of reduced dimension. Since PCR is based on capturing the maximum variation in predictors on every component it has used to model, this method does not consider the response structure (Jolliffe, 2002). In addition, if the relevant components are not in the initial position, the method requires more number of components to make precise prediction (Almøy, 1996).

*Partial Least Squares* (PLS) regression ~~tries~~ to maximize the covariance between the predictors and the response scores (de Jong, 1993). Broadly, PLS can be divided into PLS1 and PLS2 where the former tries to model the response variables individually and the ~~later~~ ~~using~~ all the response variables together while modelling. Among the three widely used algorithms NIPALS (Wold, 1975), SIMPLS (de Jong, 1993) and KernelPLS (Lindgren et al., 1993), for this study we will be using KernelPLS which gives equivalent results to the other two and is default in R-package `pls` (Mevik and Wehrens, 2007).

*Envelopes* is first introduced by (Cook et al., 2007) as the smallest subspace that includes the span of true regression coefficients. *Predictor Envelopes* (Xenv) identifies the envelope as a smallest subspace in predictor space, by separating the predictor covariance $\mathbf{\Sigma}_{xx}$ into relevant (material) and irrelevant (immaterial) parts, such that response $\mathbf{y}$ is uncorrelated with the irrelevant part given the relevant one. In addition, relevant and irrelevant parts are also uncorrelated. Such separation of the covariance matrix is made using the data through optimization of ~~the~~ objective function. Further, the regression coefficients are estimated only using the relevant part. Cook et al. (2010), Cook et al. (2013) and Cook (2018) have extensively discussed the foundation and various mathematical constructs together with properties related to Predictor Envelope.

*Simultaneous Predictor-Response Envelope* (Senv) implements the envelope in both response and predictor spaces. It separates the material and immaterial part in response space and predictor space such that the material part of response does not correlate with the immaterial part of predictor and the immaterial part of response does not correlate with the material part of the predictor. The regression coefficients are computed using only the material part of the response and predictor spaces. The number of components specified in both of these methods during the fit influences the separation of these spaces. If the number of response components equals to the number of responses, simultaneous envelope reduces to the predictor envelope, and if the number of predictor components equals to the number of predictors, the result will be equivalent to the ordinary least squares. Cook and Zhang (2015) and Cook (2018) have discussed the method in detail. Further, Helland et al. (2018) have discussed how the population models of PCR, PLS and Xenv are equivalent.

## 4. Experimental Design

An R (R Core Team, 2018) package `simrel` (Rimal et al., 2018; Sæbø et al., 2015) is used to simulate the data for comparison. In the simulation, the number of observation is fixed at $n = 100$ and following four simulation parameters are used to obtain the data with a wide range of properties.

**Number of predictors: (p)** In order to cover both tall $(n > p)$ and wide $(p > n)$ cases, $p = 20$ and $p = 250$ number of predictors are simulated.

**Multicollinearity in predictor variables: (gamma)** A parameter `gamma` $(\gamma)$ controls the exponential decline of eigenvalues $(\lambda_i, i = 1, \ldots p)$ corresponding to predictor variables as,

$$\lambda_i = e^{-\gamma(i-1)}, \gamma > 0 \text{ and } i = 1, 2, \ldots p \tag{3}$$

Two levels 0.2 and 0.9 of `gamma` are used for simulation so that level 0.2 simulates the data with low multicollinearity and 0.9 simulates the data with high multicollinearity.

**Position of relevant components: (relpos)** Initial principal components of a non-singular covariance matrix are larger than the later ones. If the principal components corresponding to predictors with larger variation is not relevant for a response, this will just increase the noise in the data. Here we will use two different levels of position index of predictor components (`relpos`): a) 1, 2, 3, 4 and b) 5, 6, 7, 8. Predictor components irrelevant for a response makes prediction difficult (Helland and Almøy, 1994). When combined with multicollinearity, this factor can create both an easy and difficult model for both estimation and prediction.

**Correlation in response variables: (eta)** Many estimators also uses the structure of response for their estimation. Here the correlation between the responses are varied through a simulation parameter `eta` $(\eta)$. The parameter controls the exponential decline of eigenvalues $\kappa_j, j = 1, \ldots m($ number of responses$)$ corresponding to response variables as,

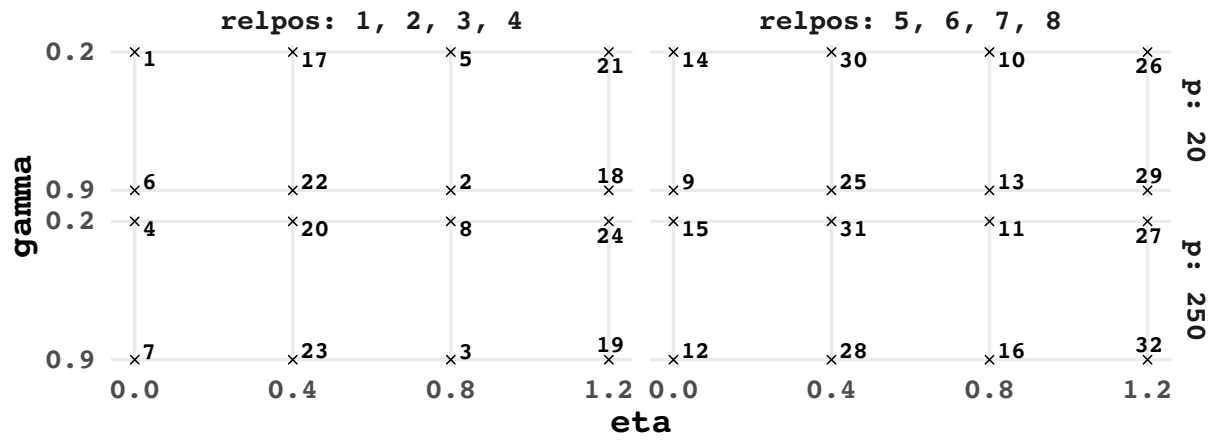


Figure 2: Experimental Design of simulation parameters. Each point represents an unique data property.

$$\eta_j = e^{-\kappa(j-1)}, \kappa > 0 \text{ and } j = 1, 2, \ldots m \tag{4}$$

Four levels 0, 0.4, 0.8 and 1.2 of `eta` are used in the so that level 0 simulates the data with uncorrelated response variables, while 1.2 simulates the highly correlated response variables.

Here we have assumed that there is only one informative response component. In the final dataset, all predictors together span the same space as the relevant predictor components and all responses together span the same space as the one informative response component. In addition, the coefficient of determination is fixed at 0.8 for all datasets.

A complete factorial design is adopted using different levels of factors discussed above to create 32 designs (Figure 2), each of which gives datasets with unique properties. From each of these design and each estimation method, 50 different datasets are simulated so that each of them has the same true population structure. In total, $5 \times 32 \times 50$ i.e., 8000 datasets are simulated.

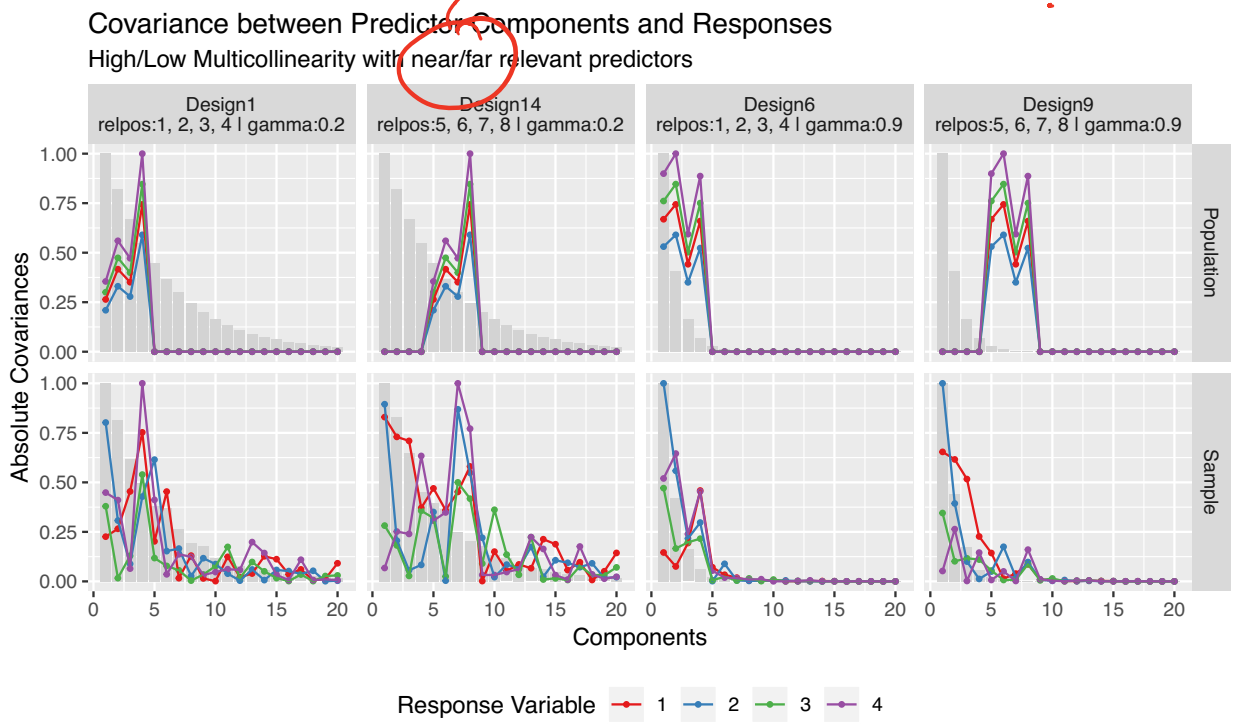


Figure 3: Covariance between predictor components and response variables in the population (top), and in the simulated data (bottom) for four different designs. The bar in the background represents the variance of the corresponding components.

The simulation properties are directly reflected in the simulated data. For example, in Figure 3, design pairs 1 and 14 as well as 6 and 9 differ in their properties only in terms of relevant predictor components, while the design pairs 1 and 6 as well as 9 and 14 differ only in terms of the level of multicollinearity. The properties in population are also reflected in the simulated samples. The combination of these factor levels creates datasets that are easy or difficult for them to model.

## 5. Basis of Comparison

The focus of this study is to extend the exploration of Rimal et al. (2019) to compare the estimation performance of PCR, PLS1, PLS2, Xenv and Senv methods. The performance is measured on the basis of,

a) average estimation error computed as in (6)

b) the average number of components used by the methods to give minimum estimation error

Let us define the expected estimation error as

$$\text{MSE}\left(\hat{\boldsymbol{\beta}}_{ijkl}\right) = \text{E}\left[\frac{1}{\sigma^2_{y_j}}\left(\boldsymbol{\beta}_{ij} - \hat{\boldsymbol{\beta}}_{ijkl}\right)^t\left(\boldsymbol{\beta}_{ij} - \hat{\boldsymbol{\beta}}_{ijkl}\right)\right] \tag{5}$$

for response $j = 1,\ldots 4$ in 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 $\sigma^2_{y_j}$ is the variance of response $j$. Since both the expectation and the variance of $\hat{\boldsymbol{\beta}}$ are unknown, the estimation error are estimated using data from 50 replications as follows,

$$\widehat{\text{MSE}\left(\hat{\beta}\right)}_{ijkl} = \frac{1}{50}\sum_{r=1}^{50}\left[\widehat{\text{MSE}_\circ\left(\hat{\beta}\right)}_{ijkl}\right] \tag{6}$$

where, $\widehat{\text{MSE}\left(\hat{\beta}\right)}_{ijkl}$ is the estimated prediction error averaged over $r = 50$ replicates and,

$$\text{MSE}_\circ\left(\hat{\boldsymbol{\beta}}\right)_{ijklr} = \frac{1}{\sigma^2_{y_j}}\left[\left(\boldsymbol{\beta}_{ij} - \hat{\boldsymbol{\beta}}_{ijklr}\right)^t\left(\boldsymbol{\beta}_{ij} - \hat{\boldsymbol{\beta}}_{ijklr}\right)\right]$$

Our further discussion revolves around *Error Dataset* and *Component Dataset*, as in the prediction comparison paper Rimal et al. (2019). For a given estimation method, design, and response, the component that gives the minimum of estimation error averaged over all replicates is selected as,

$$l_\circ = \underset{l}{\text{argmin}}\left[\frac{1}{50}\sum_{r=1}^{50}\widehat{\text{MSE}_\circ\left(\hat{\beta}\right)}_r\right] \tag{7}$$

The estimation error $\widehat{\text{MSE}_\circ\left(\hat{\beta}\right)}$ for every methods, design and response corresponding to $l_\circ$ component, computed as (7), is then regarded as *error dataset* in the subsequent analysis. Let $\mathbf{u}_{8000\times 4} = (u_j)$, where $u_j$ is the $j^{\text{th}}$ column of $\mathbf{u}$ denoting the estimation error corresponding to response $j = 1,\ldots 4$ in the context of this dataset. Further, let the number of components that result in minimum estimation error in each replication be computed

as (8) will be considered as *component dataset*. Let $\mathbf{v}_{8000\times 4} = (v_j)$ where $v_j$ is the $j^{\text{th}}$ column of $\mathbf{v}$ denoting the outcome variable measuring the number of components used to obtain minimum estimation error corresponding to response $j = 1, \ldots 4$.

$$l_\circ = \underset{l}{\operatorname{argmin}} \left[ \widehat{\mathrm{MSE}}_\circ \left( \hat{\beta} \right) \right] \tag{8}$$

## 6. Exploration

This section explores the variation in the *error dataset* and the *component dataset* for which we have used Principal Component Analysis (PCA). Let $t_u$ and $t_v$ be column vectors denoting the principal component scores corresponding to $\mathbf{u}$ and $\mathbf{v}$ matrices, respectively. The scores density in Figure 4 and Figure 5 correspond to the first principal component of $\mathbf{u}$ and $\mathbf{v}$, i.e. the first column of $t_u$ and $t_v$, respectively. Here higher scores correspond to the larger estimation error and vice versa.

The plot in Figure 4 shows a clear difference between the effect of low and high multicollinearity on estimation error. In the case of low multicollinearity (`gamma: 0.2`), the estimation errors are smaller and have lesser variation compared to high multicollinearity (`gamma: 0.9`). In addition, the envelope methods have increased their estimation error in the case of highly correlated responses (`eta: 1.2`), whereas there is no effect of this correlation in other methods.

Furthermore, position of the relevant predictor components has a noticeable effect on estimation error. When relevant predictors are at position 5, 6, 7, 8, the components at 1, 2, 3, 4, which carry most of the variation, become irrelevant. These irrelevant components with large variation add noise to the model and consequently increases the estimation error. The effect intensifies on highly collinear predictors. Designs with high multicollinearity and relevant predictors at position 5, 6, 7, 8 are relatively difficult to model for all the methods. Although these difficult designs have a large effect on estimation error, their effect on prediction error is less influential (Rimal et al., 2019).

In the case of the *component dataset* (Figure 5), PCR, PLS1 and PLS2 methods have used more components in the case of high multicollinearity compared to low. Surprisingly,


Figure 4: Scores density corresponding to first principal component of *error dataset* (**u**) subdivided by `methods`, `gamma` and `eta` and grouped by `relpos`.

the envelope methods (Senv and Xenv) mostly have used a distinctly ~~lesser~~ number of components in both ~~the~~ cases of multicollinearity compared to other methods.

The plot also shows that there is no clear effect ~~due to~~ the correlation ~~of~~ response variables on the number of components used to obtain minimum estimation error.

A clear interaction between the position of relevant predictors and the multicollinearity is visible in the plot, suggests that the methods use a larger number of components when the relevant components are at position 5, 6, 7, 8. Additionally, the use of components escalate and the difference between the two levels of `relpos` becomes wider in the case of high multicollinearity in the ~~model.~~ Such performance is also seen the case of prediction

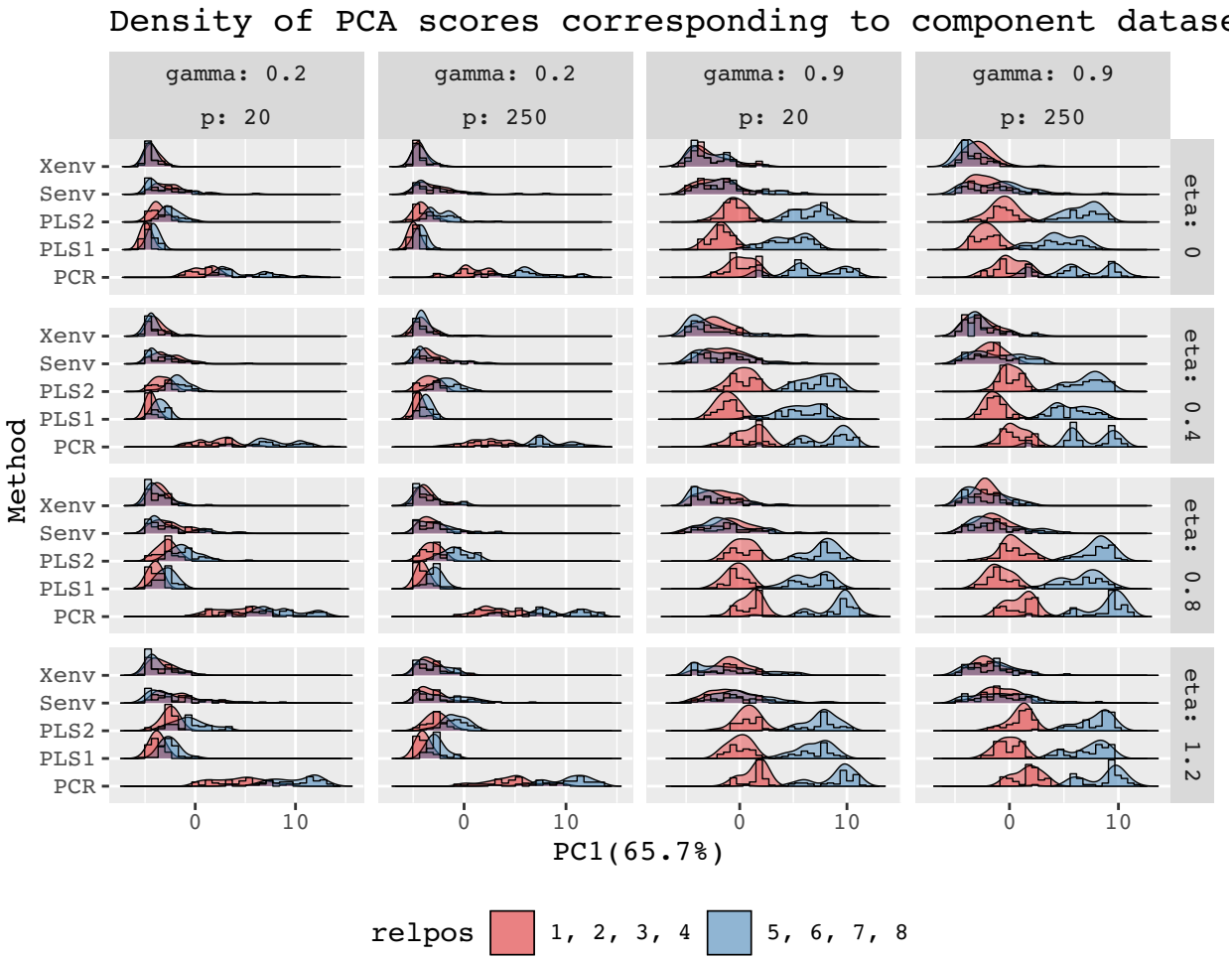


Figure 5: Score density corresponding to the first principal component of *component dataset* (**v**) subdivided by `methods`, `gamma` and `eta` and grouped by `relpos`.

error (See Rimal et al. (2019)), however the number of components used ~~in that case~~ is ~~lesser~~ than in ~~this~~ case. Even when the relevant ~~predictors~~ are at position 5, 6, 7, 8, the envelope methods, in contrast to other methods, have used an almost similar number of components as in the case of relevant ~~predictor~~ at position 1, 2, 3, 4. This shows that the envelope methods identify the predictor space relevant to the response differently and with very few numbers of latent components. Following sub-section explores the prediction and estimation error ~~together with~~ the regression coefficient of Simultaneous Envelope and Partial Least Squares for a design having high multicollinearity with predictor components at position 5, 6, 7, 8. Here we

will use design with $n > p$ and two levels of correlation between the responses. These correspond to Design-9 and Design-29.

Figure 7 shows a clear distinction between the modelling approach of PLS2 and Senv methods for the same model based on Design 9 (top) and Design 29 (bottom). In both of the designs, PLS2 has both minimum prediction error and minimum estimation error obtained using seven to eight components and the estimated regression coefficients approximate the true coefficients. In contrast, the Senv method has approached the minimum prediction and minimum estimation error using one to two components and the corresponding estimated regression coefficients approximate the true coefficients (Figure 6). Despite having contrasted modelling results for a dataset with similar properties, the minimum errors produced by them are comparable in the case of Design 9 (See Table 1). However, in the case of Design 29, estimation error corresponding to PLS1 and envelope methods are much higher than PCR and PLS2. It is interesting to see that despite having large estimation error, the prediction error corresponding to the envelope methods are much smaller in this design. Since the increased correlation in the response is responsible for the elevated estimation error in envelopes, it is certain that the performance of envelope methods, which tries to find relevant space with few number of components, is affected more than the other methods. Here the response dimension for the simultaneous envelope has been fixed at two components, which might have affected its performance, however, both envelope methods had performed much better with the same restriction in the case of prediction.

Figure 7 also shows that, in both designs, Senv has resulted in huge estimation error when the number of components is not optimal. This is also true for the PLS2 model, however, the extent of this variation is noticeably larger for the Senv method. A similar observation as Senv is also found in Xenv method while PCR and PLS1 are closer to the PLS2 in terms of their use of components in order to produce the minimum error (See Table 1).

In addition to the prediction and estimation error, Figure 6 gives a closer view of how the average coefficients corresponding to these methods approximate to true values. Here PLS2 has used seven to eight components to reach the closest approximation to the true

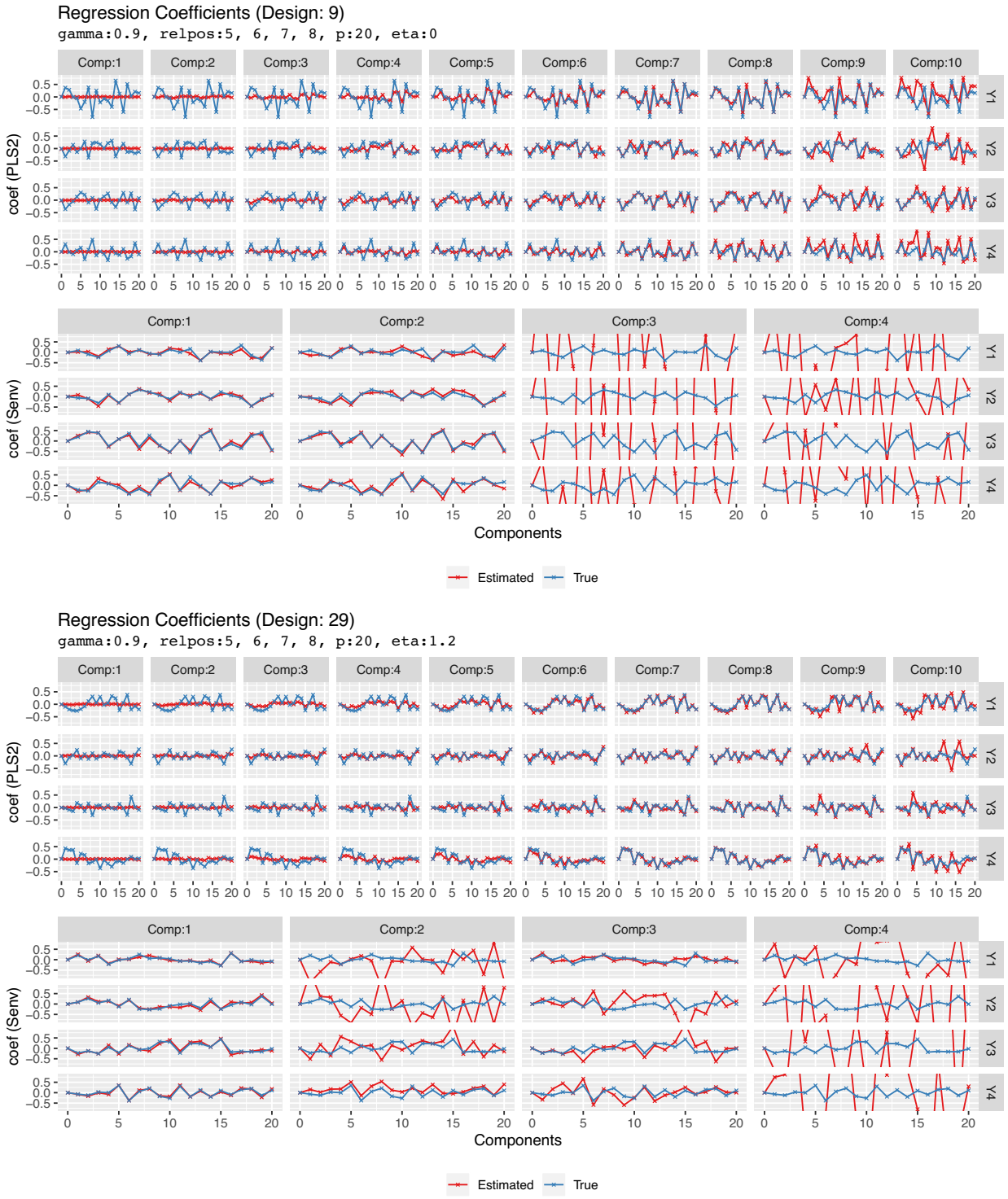


Figure 6: Regression Coefficients (coef) estimated by PLS2 and Simultaneous Envelope methods on the data based on Design 9 and 29.

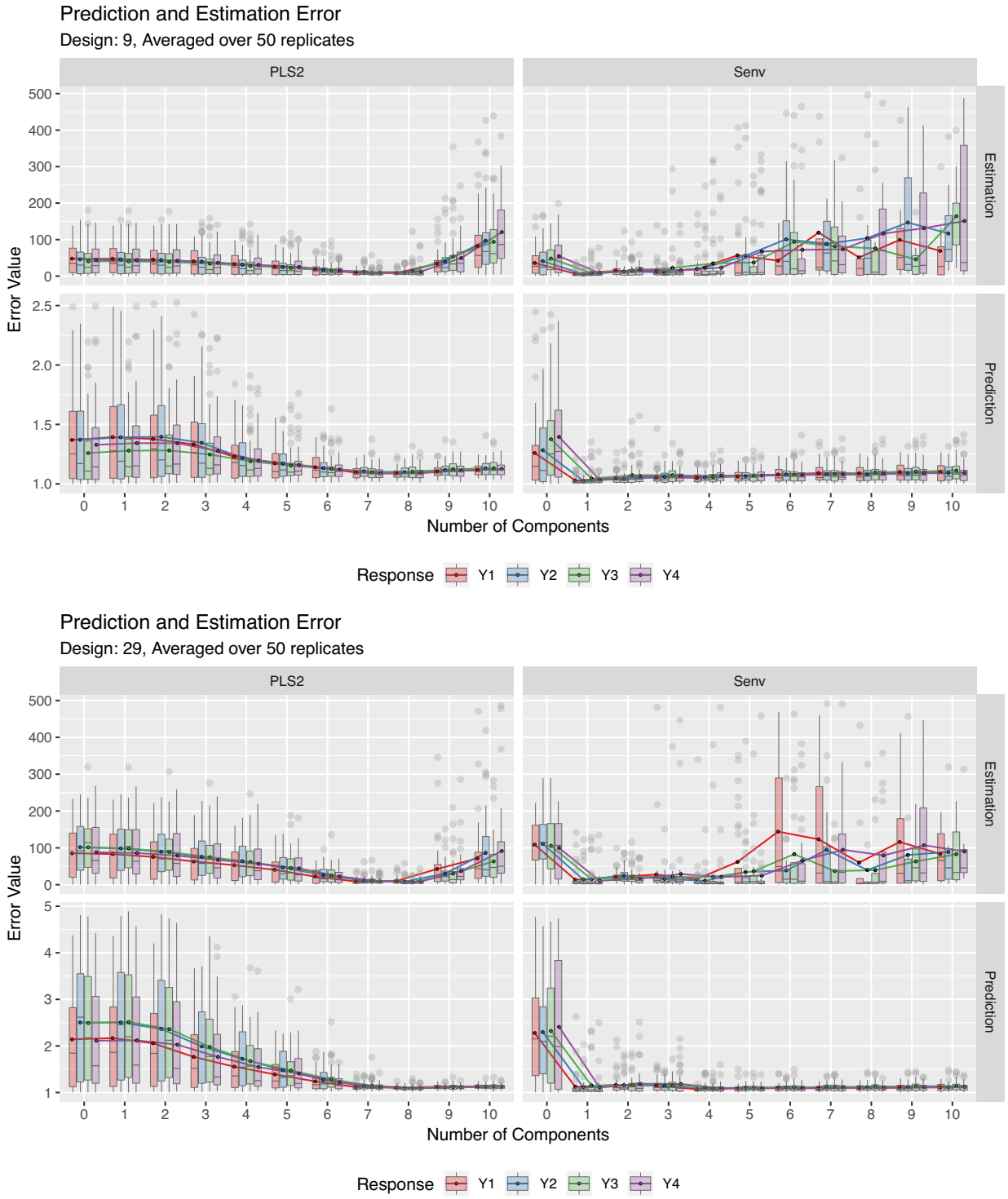


Figure 7: Minimum prediction and estimation error for PLS2 and Simultaneous Envelope methods. The point and lines are averaged over 50 replications.

Table 1: Minimum Prediction and Estimation Error for Design 9 (component numbers in parantheses)

| Design | Response | PCR | PLS1 | *PLS2* | *Senv* | Xenv |
|---|---|---|---|---|---|---|
| **Design 9** | | | | | | |
| **Estimation Error** | | | | | | |
| 9 | 1 | 8.56 (8) | 13.23 (6) | ***8.17 (8)*** | ***6.65 (1)*** | 5.73 (1) |
| 9 | 2 | 7.94 (8) | 14.42 (6) | ***10.65 (8)*** | ***5.06 (1)*** | 5.35 (1) |
| 9 | 3 | 7.02 (8) | 15.9 (6) | ***8.22 (7)*** | ***8.55 (1)*** | 5 (1) |
| 9 | 4 | 9.26 (8) | 13.14 (7) | ***8.29 (7)*** | ***8.19 (1)*** | 4.78 (1) |
| **Prediction Error** | | | | | | |
| 9 | 1 | 1.08 (8) | 1.1 (7) | ***1.09 (8)*** | ***1.03 (1)*** | 1.03 (1) |
| 9 | 2 | 1.09 (8) | 1.11 (7) | ***1.1 (8)*** | ***1.03 (1)*** | 1.03 (1) |
| 9 | 3 | 1.08 (8) | 1.1 (7) | ***1.1 (7)*** | ***1.04 (1)*** | 1.03 (1) |
| 9 | 4 | 1.09 (8) | 1.1 (7) | ***1.09 (7)*** | ***1.04 (1)*** | 1.03 (1) |
| **Design 29** | | | | | | |
| **Estimation Error** | | | | | | |
| 29 | 1 | 6.16 (8) | 13.64 (7) | ***8.67 (7)*** | ***13.45 (1)*** | 13.05 (1) |
| 29 | 2 | 6.29 (8) | 12.3 (7) | ***8.49 (8)*** | ***13.62 (1)*** | 10.98 (1) |
| 29 | 3 | 6.73 (8) | 13.03 (7) | ***6.54 (8)*** | ***14.72 (1)*** | 16.24 (1) |
| 29 | 4 | 6.28 (8) | 12.51 (7) | ***8.66 (8)*** | ***10.76 (1)*** | 10.27 (1) |
| **Prediction Error** | | | | | | |
| 29 | 1 | 1.09 (8) | 1.1 (8) | ***1.1 (8)*** | ***1.07 (4)*** | 1.1 (5) |
| 29 | 2 | 1.1 (8) | 1.11 (8) | ***1.09 (8)*** | ***1.1 (5)*** | 1.11 (1) |
| 29 | 3 | 1.1 (8) | 1.1 (8) | ***1.1 (8)*** | ***1.09 (4)*** | 1.13 (5) |
| 29 | 4 | 1.09 (8) | 1.11 (8) | ***1.09 (8)*** | ***1.09 (5)*** | 1.11 (1) |

coefficients which started to increase after including components larger than eight. This departure from true coefficients is usual for PLS when the relevant components are at 1, 2, 3, 4 whereas PCR has shown the more stable result in such situations. Further, the envelope methods have presented their ability to converge their estimates to the true value in just one or two components. However, one should be cautious about determining the optimal components in these methods due to a highly unstable and large error in non-optimal cases.

Despite having a large variation in prediction and estimation error, the envelope based methods have produced a better result even in the difficult model as obtained from Design 9.

## 7. Analysis

A statistical analysis using Multivariate Analysis of variance (MANOVA) model is performed on *error dataset* and *component dataset* in order to understand the association between data properties and the estimation methods. Let the corresponding MANOVA models be *error model* (9) and *component model* (10). In the MANOVA models we will consider the interaction of simulation parameters (`p`, `gamma`, `eta`, and `relpos`) and `Method` The models is fitted using corresponding *error dataset* (**u**) and *component dataset* (**v**).

**Error Model:**

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

**Component Model:**

$$\mathbf{v} = \boldsymbol{\mu} + (\texttt{p} + \texttt{gamma} + \texttt{eta} + \texttt{relpos} + \texttt{Methods})^3 + \boldsymbol{\varepsilon} \tag{10}$$

where, **u** corresponds to the estimation errors in *error dataset* and **v** corresponds to the number of components used by a method to obtain minimum estimation error in the *component dataset*. To make the analysis equivalent to Rimal et al. (2019), we have also used Pillai's trace statistic for accessing the result of MANOVA. Figure 8 plots the Pillai's trace statistics as bars with corresponding F-values as text labels. The left plot corresponds to the *error model* and the right plot corresponds to the *component model*.

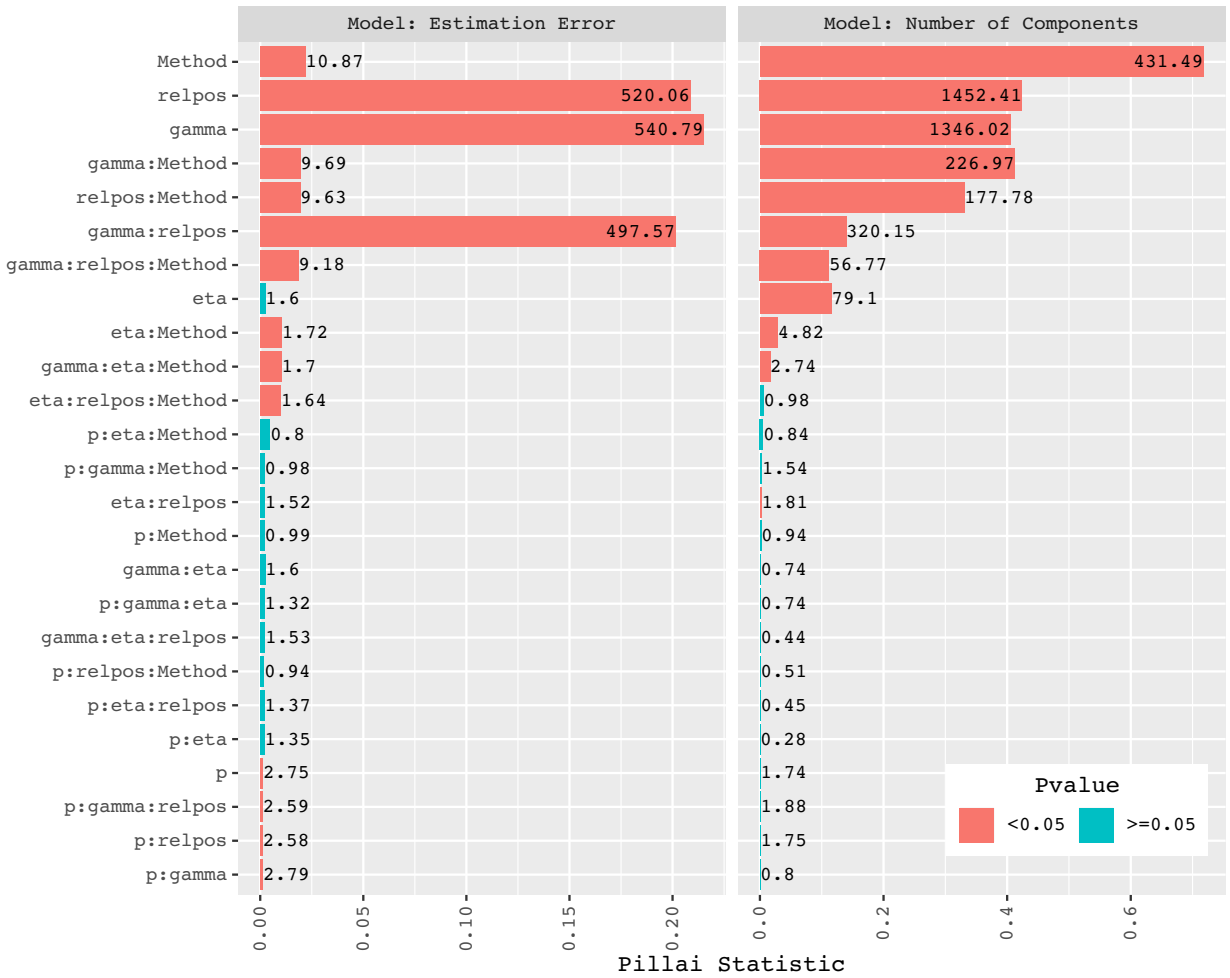


Figure 8: Pillai Statistic and F-value for the MANOVA model. The bar represents the Pillai Statistic and the text labels are F-value for the corresponding factor.

**Error Model:** Unlike prediction error in Rimal et al. (2019), `Method` has a lesser effect while the amount of multicollinearity controlled by `gamma` parameter has a huge effect in the case of estimation error (Figure 8). In addition, the position of relevant predictors and its interaction with the `gamma` parameters also have a substantial effect on the estimation error. This also supports the results seen in the Exploration section where relevant predictors at position 5, 6, 7, 8 with high multicollinearity design creates a large uninformative variance in the components 1, 2, 3, 4 making the design difficult. The effect of this on the estimation error is much larger than on the prediction error. Furthermore, the `eta` factor controlling the correlation between the responses, and

its second-order interaction with other factors except for the number of predictors is significant. The effect is also comparable with the main effect of `Method` and `eta`.

**Component Model:** Although the `Method` does not have a large impact on the estimation error, the *component model* in Figure 8 (right) shows that the methods are significantly different and has a huge effect on the number of components they use to obtain the minimum estimation error. The result also corresponds to the case of prediction error in Rimal et al. (2019). However, the F-value corresponding the `relpos` and `gamma` shows that the importance of these factors is much stronger compared to the case of prediction error.

The following section will further explore the effects of individual levels of different factors.

### *7.1. Effect Analysis of the Error Model*

In figure 9 (left), the effect of correlation between the responses controlled by the `eta` parameter has a clear influence on the estimation error. For the envelope methods, high values of `eta`, i.e. highly correlated responses, have a negative effect on estimation error whereas the effect is negligible for other methods. In all methods, the error in the case of relevant predictors at position 5, 6, 7, 8 is huge as compared to the case where relevant predictors are at position 1, 2, 3, 4. Among these methods, PLS1 has the highest estimation error in both levels of relevant predictors since the method models the response variables independently and does not consider the correlation structure in them.

Figure 9 (right) shows the large difference in the effect of two levels of the position of relevant predictors especially in the designs with high multicollinearity. In the case of high multicollinearity, all methods have noticeable poor performance compared to the case of low multicollinearity.

Figure 9 also shows that in the case of designs with uncorrelated responses, envelope methods have least average estimation error, While PCR and PLS2, being somewhat invariant to the effect of this correlation structure, they have performed better than envelope in the designs with correlated responses. In addition, the average estimation

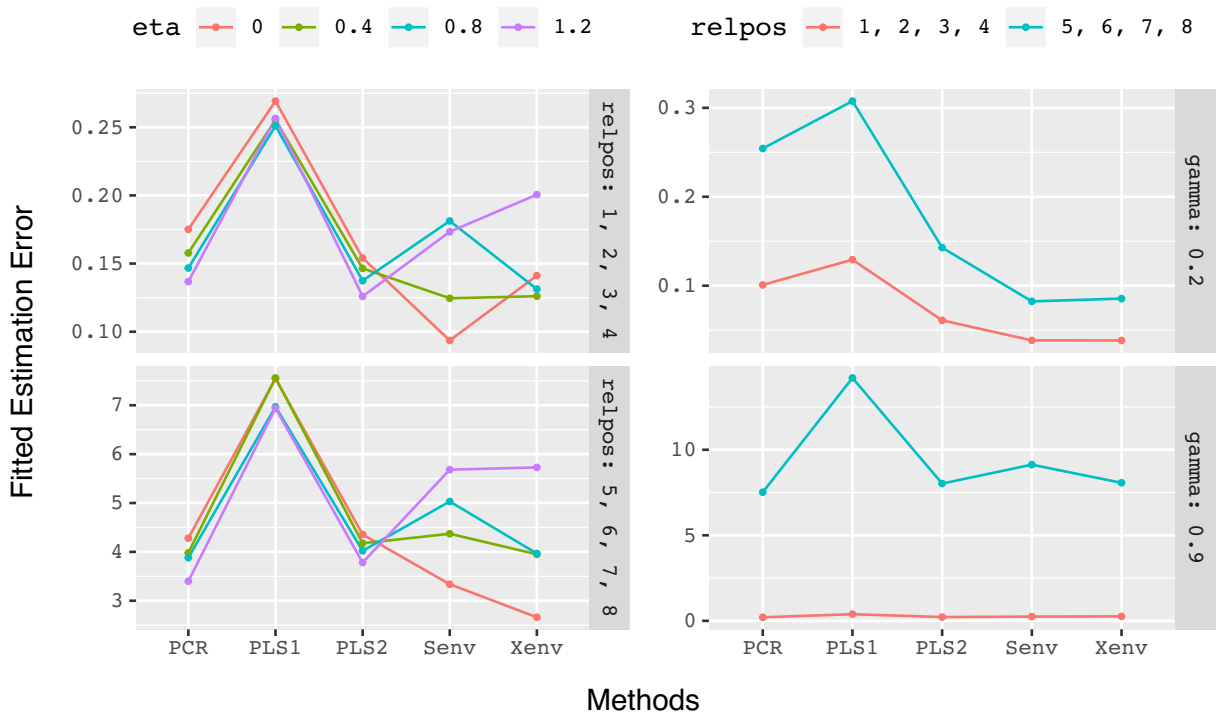


Figure 9: Effect plot of some interactions of the MANOVA corresponding to fitted *error model*

error corresponding to envelop methods in the designs with low multicollinearity is smaller than other methods.

### 7.2. *Effect Analysis of the Component Model*

In the case of fitted *component model*, envelope methods are the clear winner in almost all designs. In the case of low multicollinearity and position of relevant predictors at 1, 2, 3, 4, PLS1 has obtained the minimum estimation error similar to the envelope methods, however in the case of high multicollinearity PLS1 has also used a fairly large number of components to obtain the minimum prediction error. Although the envelope methods have comparable and minimum estimation error in some of the designs, in almost all the designs these methods have used 1-2 components on average. The effect of the correlation in the response has minimal effect on the number of components used by the methods. The design nine, which we have considered in the previous section, has minimum estimation error from envelope methods using only one predictor components by both Xenv and

Senv. In design 29 where the envelope methods have poorer performance than the other methods due to highly correlated responses, the number of components used by them is still one. This corresponds to the results seen in Figure 10.

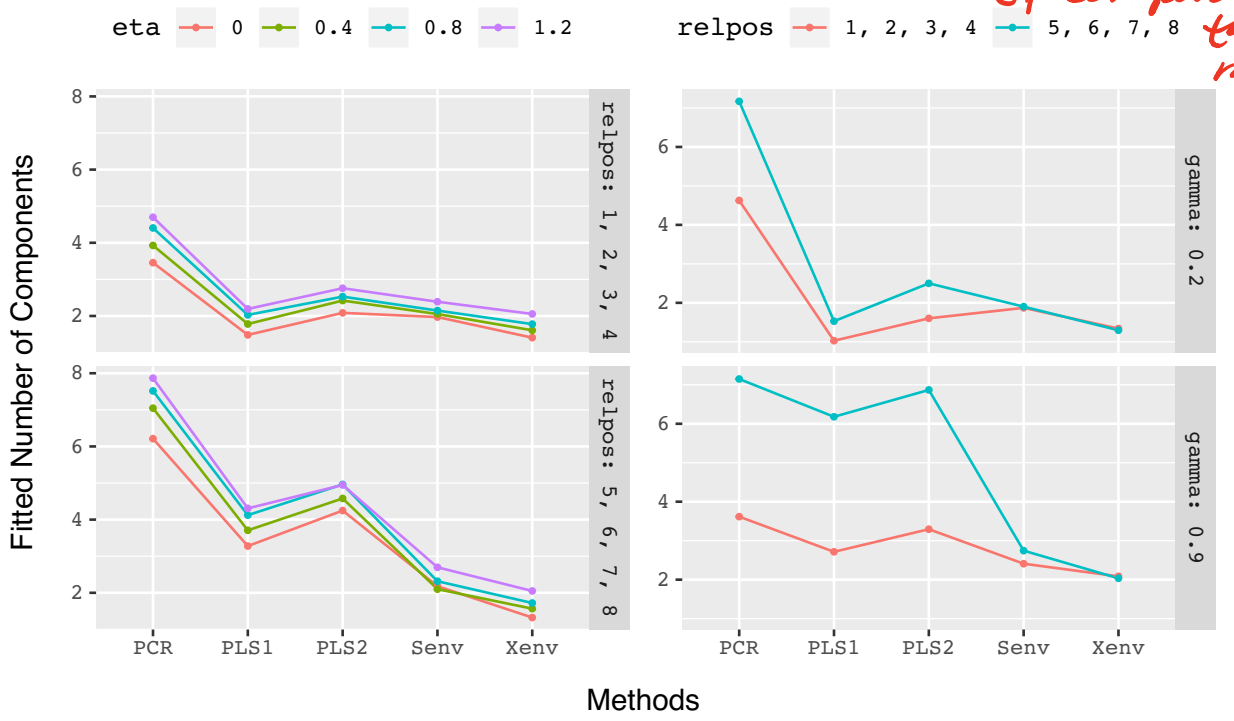


Figure 10: Effect plots of some interactions of the multivariate linear model corresponding to *component model*.

## 8. Discussion and Conclusion

The overall performance of methods highly depends on the nature of the data. The MANOVA plots show that most of the simulation parameters, except p, has significant interaction with methods. In addition, the high interaction of gamma and the relpos suggest to estimate or identify the relevant predictor components in the case of highly multicollinearity data before any analysis. Although the interaction does not have this extent of influence in prediction, one should be careful about interpreting the estimates. Designs with low multicollinearity and independent responses are in favour of envelope methods. The methods have produced the smallest prediction and estimation error with significantly few numbers of components in these designs. However, as the correlation

in the responses increases, the estimation error in envelope methods in most cases also increases noticeably. Despite the interaction of the `eta` parameter with the method is significant, the extent of its effect is rather small compared to both main and interaction effect of `gamma` and `relpos`.

The effect of the number of variables is negligible in all cases for all designs. Here the use of principal components for reducing the dimension of $n < p$ designs, as in prediction paper, has been useful so that we were able to model using envelope methods without losing much variation in the data.

Both prediction and estimation corresponding to PCR methods are found to be stable even when the non-optimal number of components are used. The PLS1 method is in general performed poorer than other methods. Unlike in prediction comparison, the performance of envelope methods in this comparison is not impressive except the use of the number of components to get the minimum estimation error. The envelope methods have used 1-2 number of components in almost all designs which are quite impressive. However, non-optimal number of components can lead to large estimation error so one should be careful in this respect while using the envelope methods. Both PLS1 and PLS2 uses less number of components when the relevant components are at position 1, 2, 3, 4. However, both methods used 7-8 components for the designs with relevant components at position 5, 6, 7, 8.

We expect the results from this study help researchers, working on theory, application and modelling, to understand these methods and their performance on various nature of the data. Based on the performance of envelope methods in these two studies, we would like to encourage readers to consider using new methods such as envelope for their research and analysis.

The first part of this study (Rimal et al., 2019) on prediction comparison should be considered to obtain a comprehensive view of this comparison. A shiny (Chang et al., 2018) web application at http://therimalaya.shinyapps.io/Comparison allows readers to explore all the visualizations for both prediction and estimation comparisons. In addition, a GitHub repository at https://github.com/therimalaya/04-estimation-comparison can be used to

reproduce this study.

## 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.

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: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.

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.

High-Level Expert Group on Artificial Intelligence, 2019. Ethics Guidelines for Trustworthy AI. Technical Report. The European Commission.

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.

Lindgren, F., Geladi, P., Wold, S., 1993. The kernel algorithm for pls. Journal of Chemometrics 7, 45–59. URL: http://dx.doi.org/10.1002/cem.1180070104, doi:doi:10.1002/cem.1180070104.

Mevik, B.H., Wehrens, R., 2007. Theplspackage: Principal component and partial least squares regression

inr. Journal of Statistical Software 18, nil. URL: https://doi.org/10.18637/jss.v018.i02, doi:doi:10.18637/jss.v018.i02.

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.

R Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

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.

Rimal, R., Almøy, T., Sæbø, S., 2019. Comparison of Multi-response Prediction Methods. arXiv e-prints , arXiv:1903.08426arXiv:1903.08426.

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.

Wold, H., 1975. Soft modelling by latent variables: the non-linear iterative partial least squares (nipals) approach. Journal of Applied Probability 12, 117–142. URL: https://doi.org/10.1017/s0021900200047604, doi:doi:10.1017/s0021900200047604.