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 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, the effect of a variable becomes unstable and can not be interpreted in such situations. Many research fields are more interested in these estimates 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, the correlation between multiple responses and position of principal components of 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. *Write some specifics from the results to show what we have found.*

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

1. Introduction

Estimation of parameters in a regression model is an integral part of many research study. Research fields such as social science, econometrics, psychology and medical study are more interested in measuring the impact of certain indicator or variable rather than

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

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 measurment, inference and prediction even more essential.

Technology has facilitated researcher to collect a large amount 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) and compares some well-established estimators such as Principal Components Analysis (PCA), 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 envelope (Senv) (Cook and Zhang, 2015). The estimation process of these methods is discussed in [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.

2. Simulation Model

As a follow-up, this study will continue using the same simulation model used by Rimal et al. (2019). The data is 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

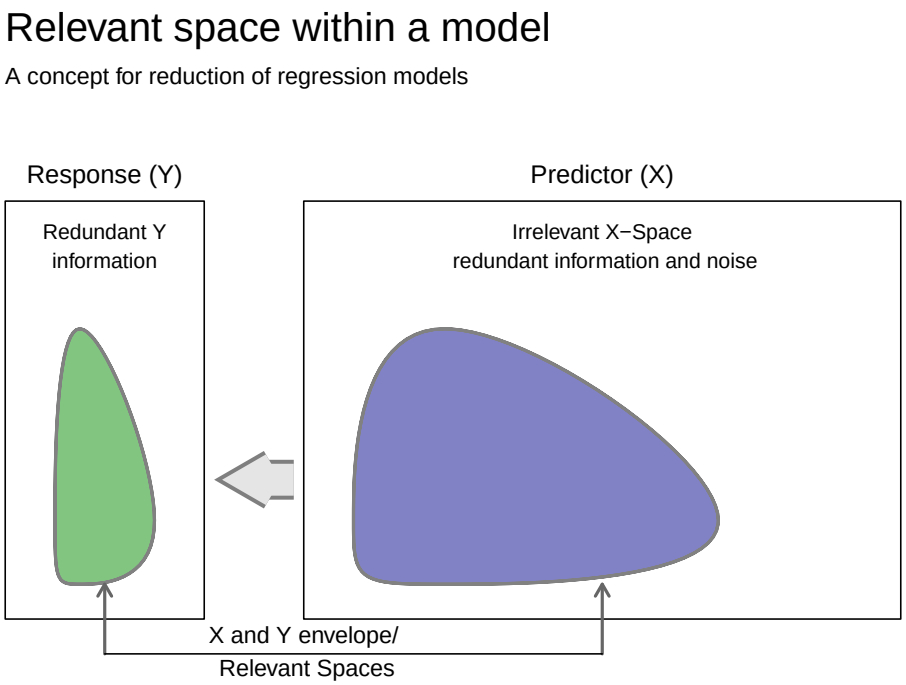

Figure 1: Relevant space in a regression model

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

Let us departure from the linear model (1),

$$\underset{(1\times m)}{\mathbf{y}} = \underset{(1\times p)(p\times m)}{\mathbf{x}\boldsymbol{\beta}} + \underset{(1\times m)}{\boldsymbol{\varepsilon}} \tag{1}$$

where $\mathbf{y}$ is a vector of m responses measured about their means, $\mathbf{x}$ is a vector of p predictors measured about their means, $\boldsymbol{\beta}$ is a matrix of regression coefficients and $\boldsymbol{\varepsilon}$ is a vector of independent error terms with constant variance $\boldsymbol{\Sigma}_{y|x}$. In ordinary least squares, coefficient $\boldsymbol{\beta}$ is estimated as,

$$\underset{(p\times m)}{\hat{\boldsymbol{\beta}}} = \left(\underset{(p\times n)(n\times p)}{\mathbf{x}^t\ \mathbf{x}}\right)^{-1} \underset{(p\times n)(n\times m)}{\mathbf{x}^t\mathbf{y}} = \mathbf{S}_{xx}^{-1}\mathbf{S}_{xy} \tag{2}$$

Let us define a transformation as $\mathbf{z} = \mathbf{x}\mathbf{R}$, where $\mathbf{R}$ is a $p \times k$ matrix with orthogonal columns. Regression model (3) defines a linear relationship of these new variables $\mathbf{z}$ with $\mathbf{y}$ through the coefficients $\boldsymbol{\alpha}$.

$$\underset{(1\times m)}{\mathbf{y}} = \underset{(1\times k)(k\times m)}{\mathbf{z}\ \boldsymbol{\alpha}} + \underset{(1\times m)}{\boldsymbol{\varepsilon}} \tag{3}$$

The coefficients $\boldsymbol{\alpha}$ can be transformed back to oritinal coefficients $\boldsymbol{\beta}$ as $\boldsymbol{\beta} = \mathbf{R}\boldsymbol{\alpha}$. We can also write a general form of regression coefficient (2) as,

$$\underset{(p\times m)}{\hat{\boldsymbol{\beta}}} = \left[\underset{(p\times k)(k\times n)}{\mathbf{R}\mathbf{x}^t}\ \underset{(n\times k)(k\times p)}{\mathbf{x}\mathbf{R}^t}\right]^{-1} \underset{(p\times k)(k\times n)}{\mathbf{R}\mathbf{x}^t}\ \underset{(n\times m)}{\mathbf{y}} = \left[\mathbf{R}\mathbf{S}_{xx}\mathbf{R}^t\right]^{-1}\mathbf{R}\mathbf{S}_{xy}^t$$

Principal Components Regression (PCR) uses k eigenvectors of $\mathbf{S}_{xx}$ as the columns of $\mathbf{R}$. Since PCR is based on capturing the maximum variation in predictors on every components it 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 require more number of components to make precise prediction (Rimal et al., 2019).

Partial Least Squares (PLS) regression on the other hand tries to maximize the covariance between the predictors and the response scores (SIMPLS paper). 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 variable together while modeling. 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 gires results equivalent to the rest and is default in R-package `pls` (Mevik and Wehrens, 2007).
Envelopes is first introduced by (Cook et al., 2007) as a smallest subspace that include the span of regression coefficients. *Predictor Envelopes* (Xenv) identifies the envelope as a smallest subspace in predictor space by separating the predictor covariance $\mathbf{S}_{xx}$ into relevant (material) and irrelevant (immaterial) parts such that response $\mathbf{y}$ is uncorrelated with the irrelevant part given the relevant one. In addition, the relevant and irrelevant parts are also uncorrelated. Such separation of the covariance matrix is made using the data through optimization of 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) implement 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 predictor. The regression coefficients are computed using only the material part of the response and predictor space. The number of components specified in both of these methods during the fit influence the separation of these spaces. If the number of response components is equals to the number of responses, simultaneous envelope reduces to the predictor envelope and if the number of predictor components is 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 model 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 wide range of properties.

Number of predictors: 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: A parameter `gamma` (γ) in simulation controls the exponential decline of eigenvalues $(\lambda_i, i = 1, \dots p)$ corresponding to predictor variables as,

$$\lambda_i = e^{-\gamma(i-1)}, \gamma > 0 \text{ and } i = 1, 2, \dots p \tag{4}$$

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

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

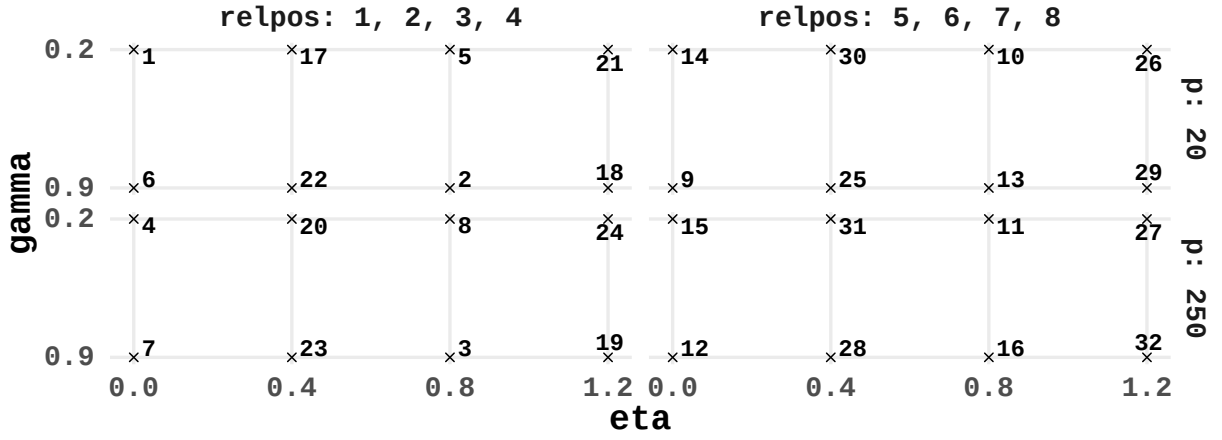

Figure 2: Experimental Design of simulation parameters. Each point represents an unique data property.

$$\eta_i = e^{-\kappa(j-1)}, \kappa > 0 \text{ and } j = 1, 2, \ldots m \tag{5}$$

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 response together span the same space as the one informative response component. In addition, coefficient of determination is fixed at 0.8 for all dataset.

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

The simulation properties are directly reflected in the simulated data. For example, in Figure 3, design pairs 1 and 4 as well as 6 and 9 differs their properties only in terms of relevant predictor components while the design pairs 1 and 6 as well as 14 and 9 differs

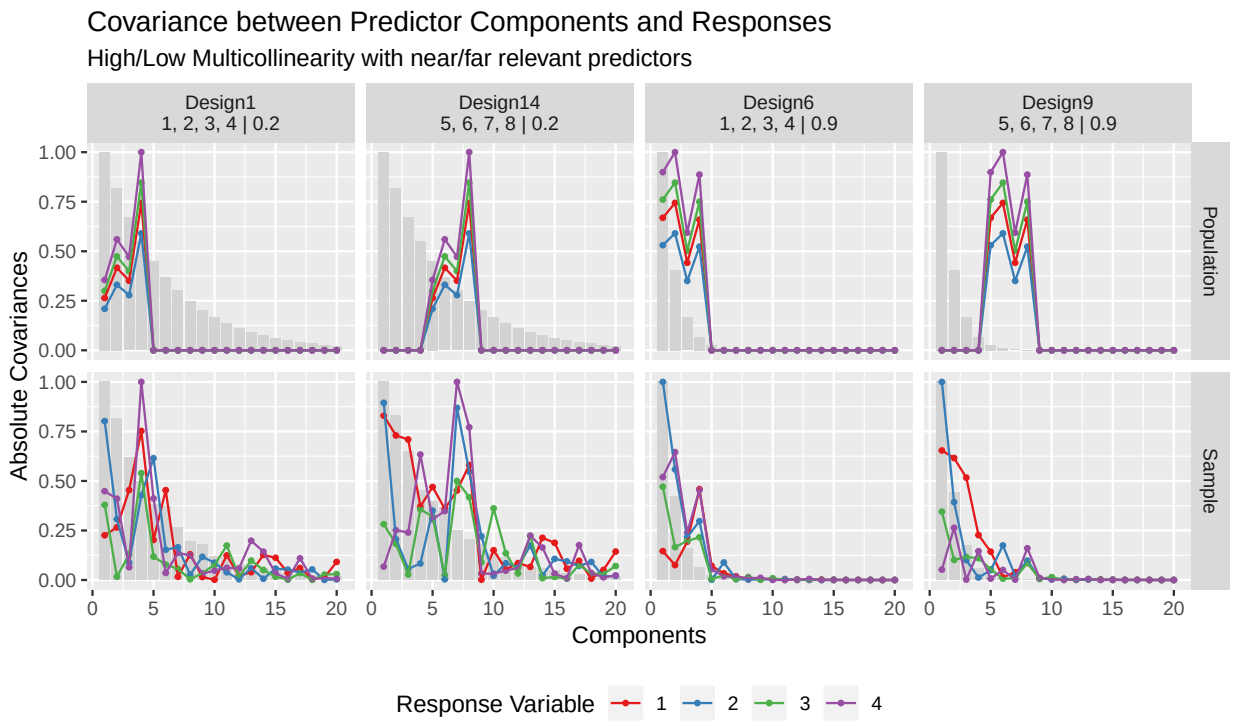

Figure 3: Covariance between predictor components and response variables in population (top) and in the simulated data (bottom) for four different designs. The Bar in the background represents the variance of corresponding components.

only in-terms of level of multicollinearity. The properties in population are also reflected in the simulated samples.

5. Basis of Comparison

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

a) average estimation error of the method using arbitrary number of components
b) average number of components used by the methods to give minimum estimation error

Let us define the expected estimation error as,

$$\mathcal{EE}_{ijkl} = \mathsf{E}\left[\left(\boldsymbol{\beta}_{ij} - \hat{\boldsymbol{\beta}}_{\boldsymbol{ijkl}}\right)^{t}\left(\boldsymbol{\beta}_{ij} - \hat{\boldsymbol{\beta}}_{\boldsymbol{ijkl}}\right)\right] \tag{6}$$

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. 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,

$$\widehat{\mathcal{EE}_{ijkl}} = \frac{1}{50}\sum_{r=1}^{50}\left[(\mathcal{EE}_{\circ})_{ijklr}\right] \tag{7}$$

where, $\widehat{\mathcal{EE}_{ijkl}}$ is the estimated prediction error averaged over $r = 50$ replicates and,

$$(\mathcal{EE}_{\circ})_{ijklr} = \left(\boldsymbol{\beta}_{ij} - \hat{\boldsymbol{\beta}}_{\boldsymbol{ijklr}}\right)^{t}\left(\boldsymbol{\beta}_{ij} - \hat{\boldsymbol{\beta}}_{\boldsymbol{ijklr}}\right)$$

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,

$$l_\circ = \underset{l}{\operatorname{argmin}} \left[\frac{1}{50} \sum_{r=1}^{50} \left(\widehat{\mathcal{E}\mathcal{E}}_\circ \right)_r \right] \tag{8}$$

The estimation error $\widehat{\mathcal{E}\mathcal{E}}_\circ$ for every methods, design and response corresponding to $l_\circ$ component, computed as (8), is then regarded as *error dataset* in the subsequent analysis. Let $\mathbf{u}_{8000\times 4} = (u_j)$ for $j = 1, \ldots 4$ be the outcome variables measuring the estimation error corresponding to the response j in the context of this dataset. Further, let the number of components that result in minimum estimation error in each replication be $l_\circ$ computed as (9) will be considred as *component dataset*.

$$l_\circ = \underset{l}{\operatorname{argmin}} \left[\widehat{\mathcal{E}\mathcal{E}}_\circ \right] \tag{9}$$

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 the principal component score sets corresponding to PCA run on the $\mathbf{u}$ 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.

The plot shows a clear difference between the effect of low and high multicollinearity in 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`). High multicollinearity has a larger influence on all but noticeably in the methods based on envelopes. Some large estimation error in the envelope is more than 100 which in the case of other methods is less than 60.

Furthermore, the relevant predictor components, in general, has a noticeable effect on estimation error. When relevant predictors are at position 5, 6, 7, 8, the predictor components at 1, 2, 3, 4, which carry most of the variation, becomes 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

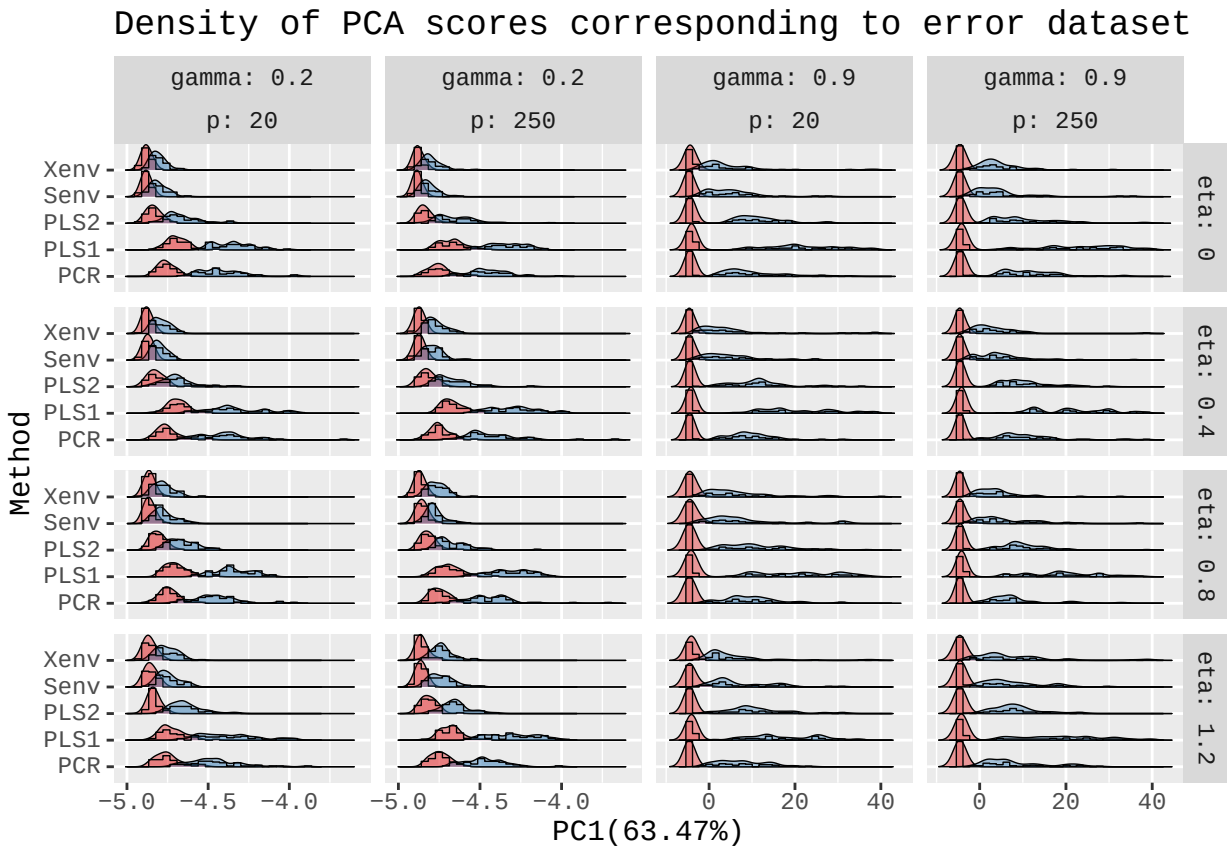

Figure 4: Scores density corresponding to first principal component of *error dataset* (**u**) subdivided by `methods`, `gamma` and `eta` and grouped by `relpos`.

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 Above), PCR, PLS1 and PLS2 methods have used more components in the case of high multicollinearity compared to low. Surprisingly, 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 variable on the number of components used to obtain minimum estimation error.

A clear interaction between the position of relevant predictors and the multicollinearity visible in the plot suggest that the methods use a larger number of components when the

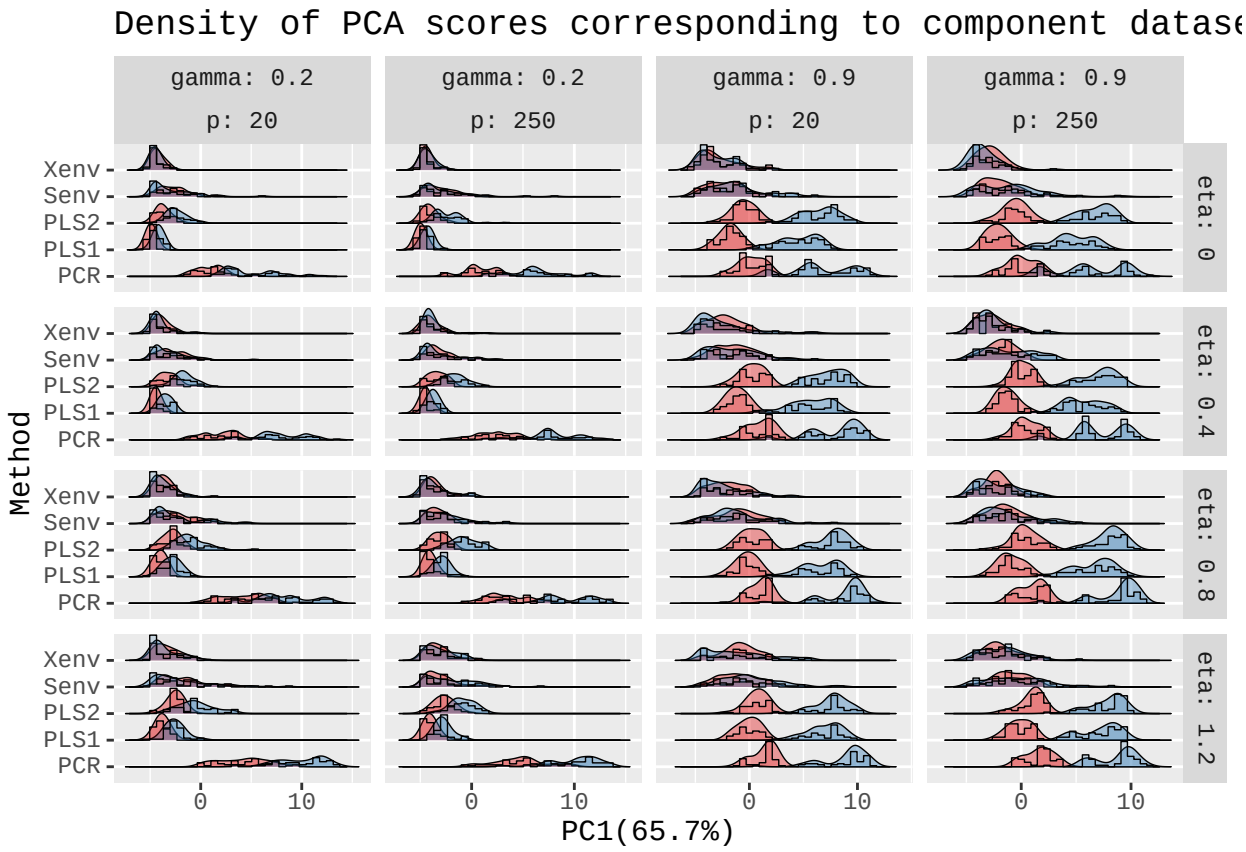

Figure 5: Score density corresponding to first principal component of *component dataset* (**v**) subdivided by `methods`, `gamma` and `eta` and grouped by `relpos`.

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 the case of prediction error (See Rimal et al. (2019)) however the number of components used in that case is lesser than in this case. Envelope methods, however, have shown a distinct result in contrast to the other methods. 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 few numbers of latent components.

Following section explore 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 no correlation between the response which corresponds to Design-9.
Figure 7 shows a clear distinction between the modelling approach of PLS2 and Senv methods for the same model based on Design 9. In the case of PLS2, both minimum prediction error and minimum estimation error are 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 contrast modelling result for a dataset with similar properties, the minimum errors produced by them are comparable (See Table 1).
The Figure 7 also shows that 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 large in 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).

```
List of 10
 $ name      : chr "kePrint"
 $ version   : chr "0.0.1"
 $ src       :List of 1
  ..$ file: chr "/usr/local/lib/R/site-library/kableExtra/kePrint-0.0.1"
 $ meta      : NULL
 $ script    : chr "kePrint.js"
 $ stylesheet: NULL
 $ head      : NULL
 $ attachment: NULL
 $ package   : NULL
```

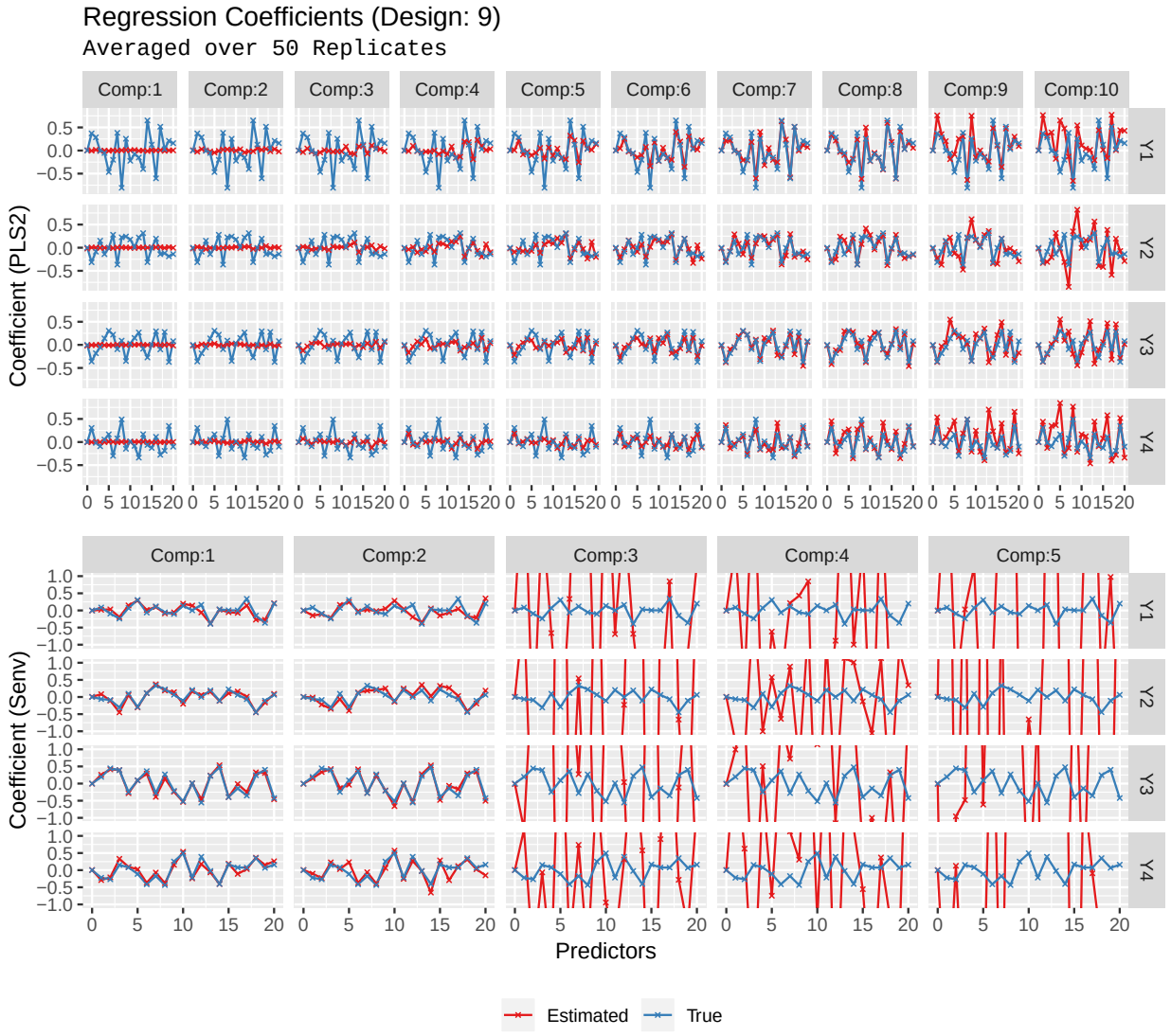

Figure 6: Regression Coefficients estimated by PLS2 and Simultaneous methods on the data based on Design 9.

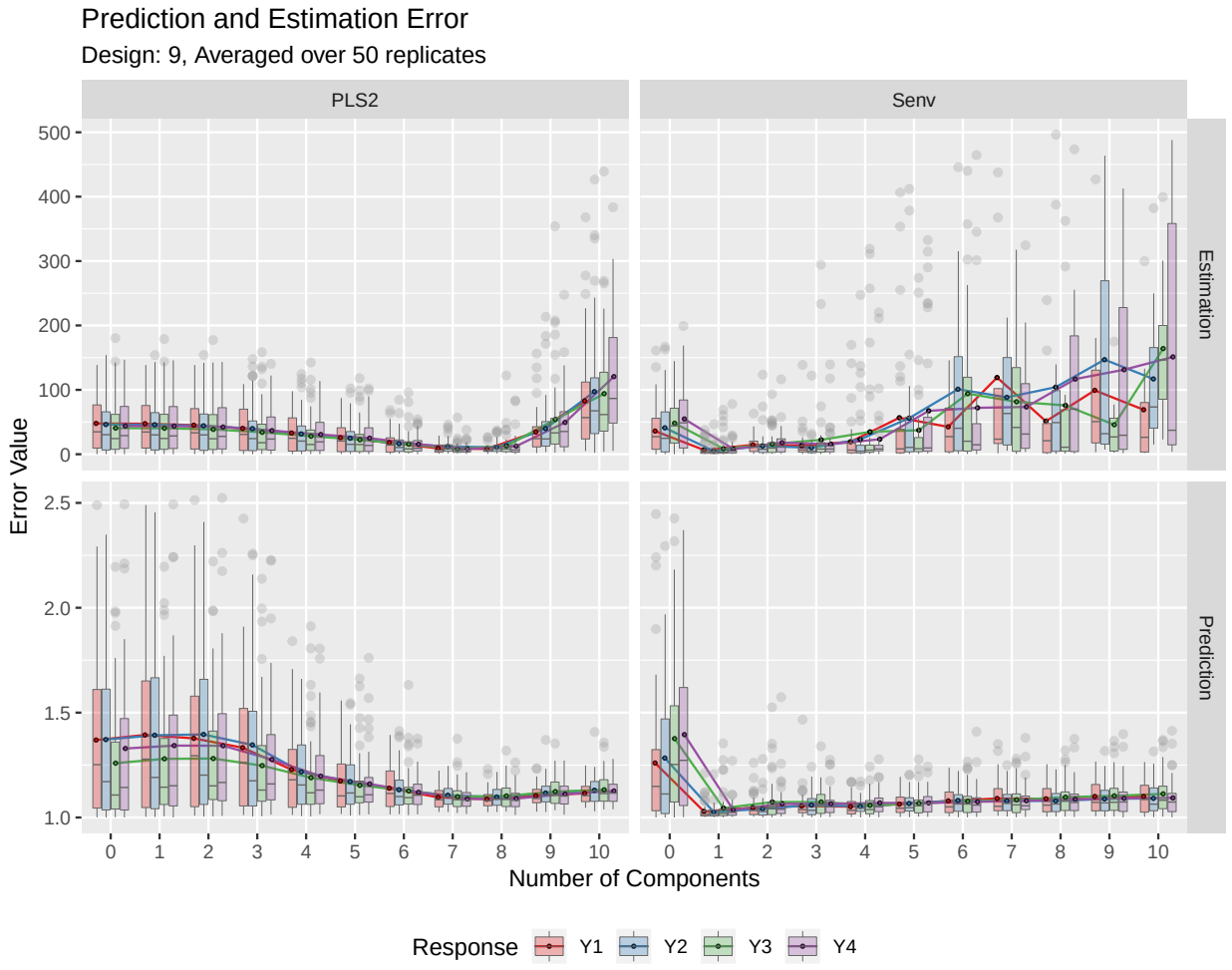

Figure 7: Minimum prediction and estimation error for PLS2 and Simultaneous methods. The point and lines are averaged over 50 replications.

Table 1: Minimum Prediction and Estimation Error for Design 9

Response	PCR	PLS1	*PLS2*	*Senv*	Xenv
Estimation Error					
1	8.56 (8)	13.23 (6)	*8.17 (8)*	*6.65 (1)*	5.73 (1)
2	7.94 (8)	14.42 (6)	*10.65 (8)*	*5.06 (1)*	5.35 (1)
3	7.02 (8)	15.9 (6)	*8.22 (7)*	*8.55 (1)*	5 (1)
4	9.26 (8)	13.14 (7)	*8.29 (7)*	*8.19 (1)*	4.78 (1)
Prediction Error					
1	1.08 (8)	1.1 (7)	*1.09 (8)*	*1.03 (1)*	1.03 (1)
2	1.09 (8)	1.11 (7)	*1.1 (8)*	*1.03 (1)*	1.03 (1)
3	1.08 (8)	1.1 (7)	*1.1 (7)*	*1.04 (1)*	1.03 (1)
4	1.09 (8)	1.1 (7)	*1.09 (7)*	*1.04 (1)*	1.03 (1)

```
$ all_files : logi TRUE
- attr(*, "class")= chr "html_dependency"
```

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 models be *error model* (10) and *component model* (11). In the model, we will consider the interaction of simulation parameters (`p`, `gamma`, `eta`, and `relpos`) and `Method` The model is fitted using corresponding *error dataset* (**u**) and *component dataset* (**v**).

Error Model:

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

Component Model:

$$\mathbf{v} = \boldsymbol{\mu} + (\mathtt{p} + \mathtt{gamma} + \mathtt{eta} + \mathtt{relpos} + \mathtt{Methods})^3 + \boldsymbol{\varepsilon} \tag{11}$$

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*.

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 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 significance of these factors are much stronger compared to the case of prediction error.

We will further explore the effects of individual levels of different factors in the following

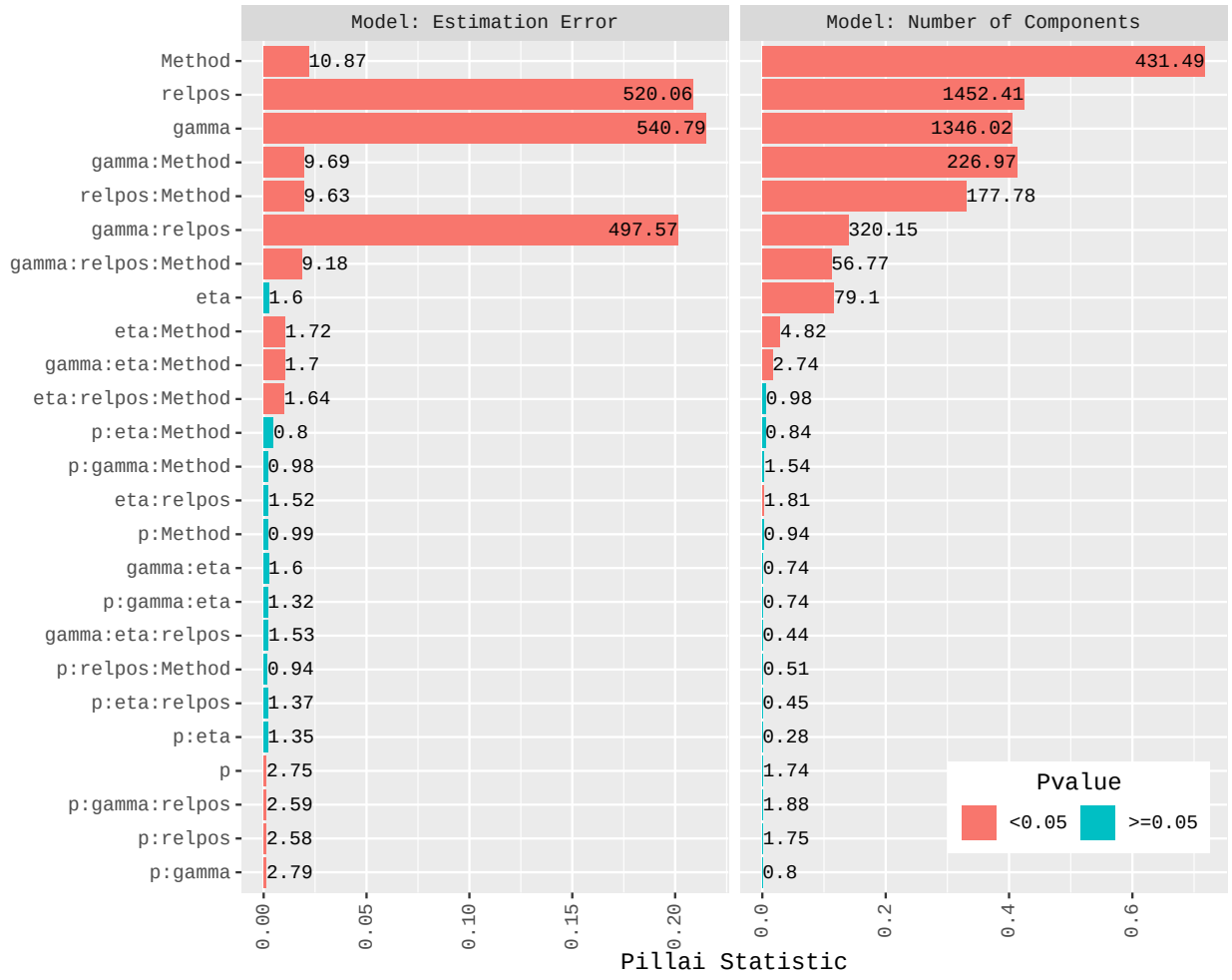

Figure 8: Pillai Statistic and F-value for the MANOVA model. The bar represents the Pillai Statistic and the text labels are F-value for corresponding factor.

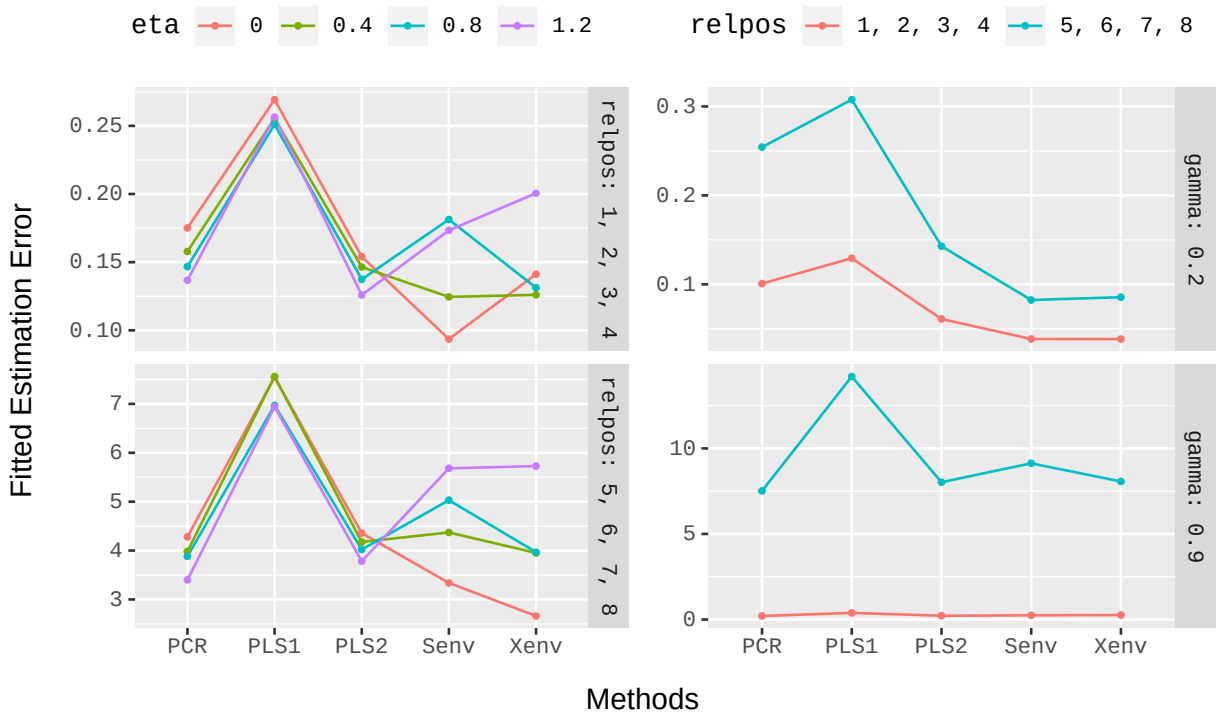

Figure 9: Effect plot of some interactions of the multivariate linear model of estimation error

subsection.

7.1. Effect Analysis of the Error Model

In figure 9 (left), the effect of correlation between the responses controlled by `eta` parameter has a clear influence on the estimation error and the effect is highly dependent on the estimation methods. In the case of envelope methods, the effect of `eta` on estimation error is smaller than other methods and are similar in the case of both levels of the position of relevant predictors (`relpos`). However, PCR, PLS1 and PLS2 have shown a clear difference in the estimation error for four different levels of response correlation. 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 PCR and PLS2 methods have the smallest effect on the estimation error in both levels of relevant predictors in the case of moderate to high level of correlation present in the response. In the case of low correlation, envelope methods have the smallest estimation error. Similarly, the designs with low multicollinearity are favourable to the envelope methods. Their estimation errors in these cases are smaller than other methods.

7.2. Effect Analysis of the Component Model

In the case of *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. It is interesting to observe that the envelope methods have comparable and minimum estimation error in most of the designs 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 at most 2 components by Xenv and 3 components by Senv. This corresponds with the results seen in Figure 10.

8. Discussion and Conclusion

- A similar discussion but based more on why the methods worked in the way we have seen in the results in previous sections
- Some concluding remarks and limitations (or a gate for further exploration)

References

Cook, R.D., 2018. An introduction to envelopes : dimension reduction for efficient estimation in multivariate statistics. 1 ed., Hoboken, NJ : John Wiley & Sons, 2018.

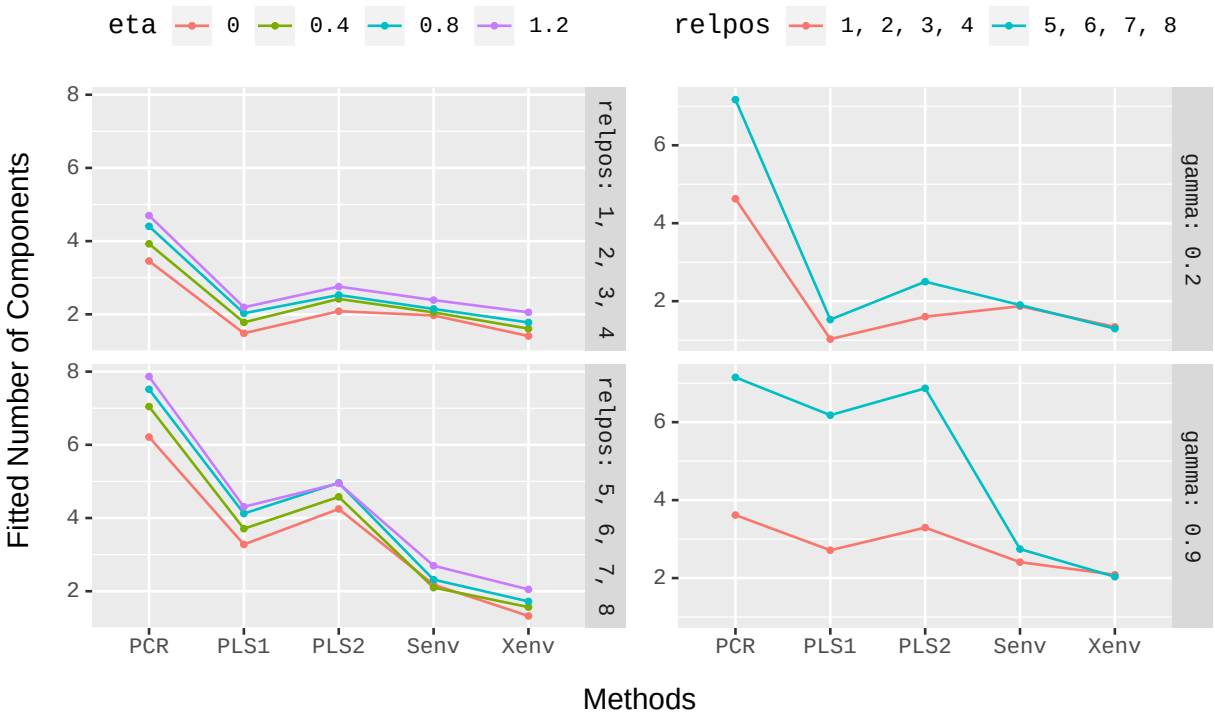

Figure 10: Effect plot of some interactions of the multivariate linear model of number of components to get minimum prediction error

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.