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

[*]Corresponding Author

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

more interested in measuring the impact of certain indicator or variable rather than performing prediction. Such studies have a large influence on people's perception and also help in policy making and decisions.

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

- Reduction of the regression model
- Include the figure from previous paper
- How the covariance and coefficients are related
- From the construction of the covariance matrix of latent variables to the simulated data
- How and what simulation parameters are related to properties of data

3. Estimation Methods

A regression model is written as,

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

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,

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

3.1. Principal Components Regression

Principal Components are new set of variables from the transformation of original dataset such that they are uncorrelated with each other and the variation in the original data are ordered from first to last of these new variables. Let us define a transformation of $\mathbf{x}$ as,

$$\underset{(1\times p)}{\mathbf{x}} = \underset{(1\times k)(k\times p)}{\mathbf{z} \quad \mathbf{R}^t} \tag{3}$$

where $\mathbf{R}$ is the eigenvectors corresponding to the covariance of $\mathbf{x}$ and $\mathbf{z}$ are the principal components. A regression model can be defined in terms of $\mathbf{z}$ as,

$$\underset{(1\times m)}{\mathbf{y}} = \underset{(1\times k)(k\times m)}{\mathbf{z} \quad \boldsymbol{\alpha}} + \underset{(1\times m)}{\boldsymbol{\varepsilon}} \tag{4}$$

Since the variation is ordered in z, only $k \leq p$ columns of z are used so that $p-k$ uninformative components are not used for modeling. The regression coefficient of (4) can be estimated as,

$$\underset{(k\times m)}{\hat{\boldsymbol{\alpha}}} = \left(\underset{(k\times n)(n\times k)}{\mathbf{z}^t \quad \mathbf{z}}\right)^{-1} \underset{(k\times n)(n\times m)}{\mathbf{z}^t\mathbf{y}} \tag{5}$$

Using (3) in (2), we get,

$$\underset{(p\times m)}{\hat{\beta}} = \left[\underset{(p\times k)(k\times n)}{\mathbf{Rz}^t}\underset{(n\times k)(k\times p)}{\mathbf{zR}^t}\right]^{-1}\underset{(p\times k)(k\times n)}{\mathbf{Rz}^t}\underset{(n\times m)}{y}$$

3.2. *Partial Least Squares Regression*

- How beta coefficients are constructed
- How it is dependent on the variance-covariance matrices
- In what way PLS1 and PLS2 differ

3.3. *Envelope Estimations*

- How beta coefficients are constructed
- How it is dependent on the variance-covariance matrices
- In what way Xenv and Senv differ

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 $(\lambda_i, i = 1, \ldots p)$ corresponding to predictor variables as,

$$\lambda_i = e^{-\gamma(i-1)}, \gamma > 0 \text{ and } i = 1, 2, \ldots p \tag{6}$$

Two levels 0.2 and 0.8 of `gamma` are used for simulation so that level 0.2 simulates the data with low multicollinearity and 0.8 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 model. Here we will use two different levels of position index of predictor components: 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 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 $\kappa_j, j = 1, \ldots m($ number of responses$)$ corresponding to response variables as,

$$\eta_i = e^{-\kappa(j-1)}, \kappa > 0 \text{ and } j = 1, 2, \ldots m \tag{7}$$

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 1) 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 × 32 × 50 i.e., 8000 datasets are simulated.

The simulation properties are directly reflected in the simulated data. For example, in

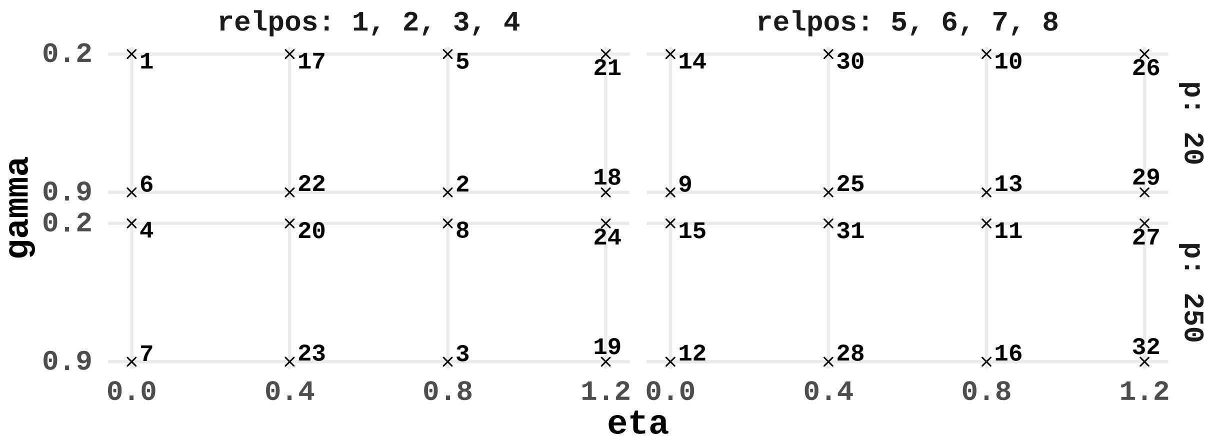

Figure 1: Experimental Design of simulation parameters. Each point represents an unique data property.

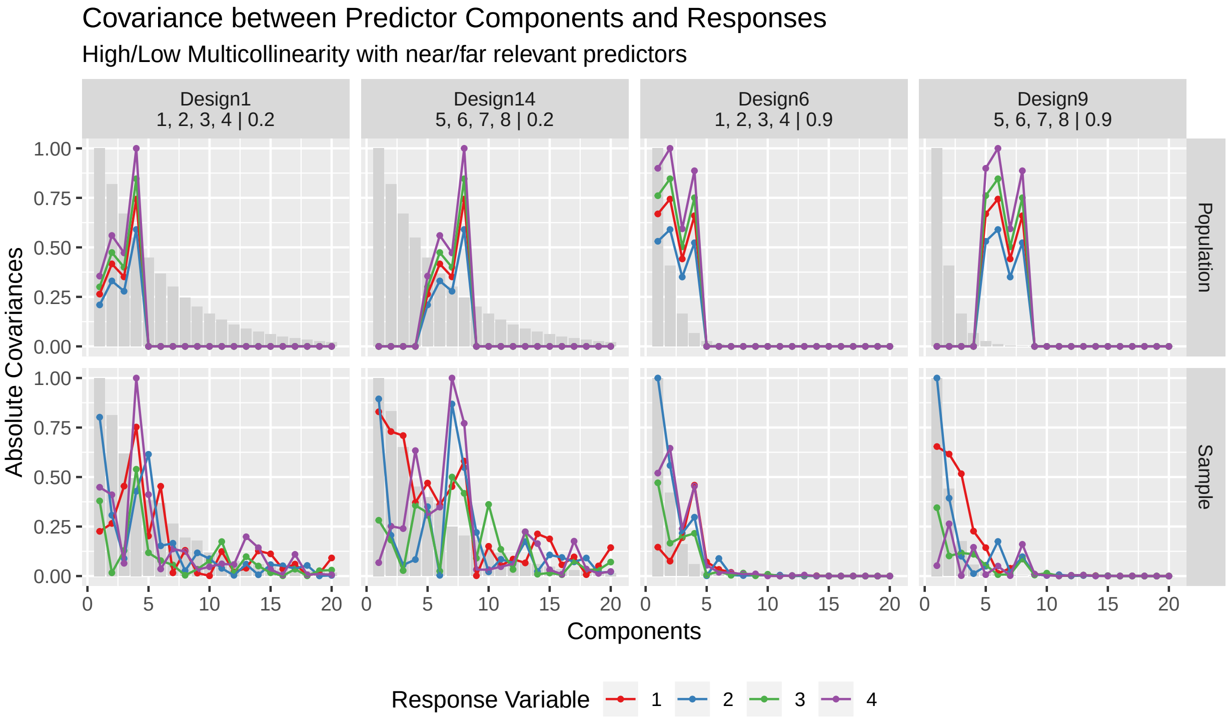

Figure 2: Covariance between predictor components and response variables in population (top) and in the simulated data (bottom) for four different designs. The Bar in the background represents the variance of corresponding components.

Figure 2, 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 only in-terms of level of multicollinearity. The properties in population are also reflected in the simulated samples.

May be we need to write few thing on how easy or difficult data are simulated with the interaction of these properties

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,

$$\mathcal{E}\mathcal{E}_{ijkl} = \mathsf{E}\left[\left(\boldsymbol{\beta}_{ij} - \hat{\boldsymbol{\beta}}_{\boldsymbol{ijkl}}\right)^t \left(\boldsymbol{\beta}_{ij} - \hat{\boldsymbol{\beta}}_{\boldsymbol{ijkl}}\right)\right] \tag{8}$$

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,

$$\widehat{\mathcal{E}\mathcal{E}_{ijkl}} = \frac{1}{50}\sum_{r=0}^{50}\left[\left(\boldsymbol{\beta}_{ij} - \hat{\boldsymbol{\beta}}_{\boldsymbol{ijklr}}\right)^t \left(\boldsymbol{\beta}_{ij} - \hat{\boldsymbol{\beta}}_{\boldsymbol{ijklr}}\right)\right] \tag{9}$$

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

5.1. Data Preparation

A simulation strategy as in Rimal et al. (2019) is adopted for preparation of data where the estimation error $(\beta_j - \hat{\beta}_j)^t(\beta_j - \hat{\beta}_j))$ is measured for each method, design and replication using number of components ranging from 0 to 10.

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 u and v matrices respectively. The scores density in Figure 3 and Figure 4 correspond to the first principal component of u and v, i.e. the first column of t_u and t_v respectively.

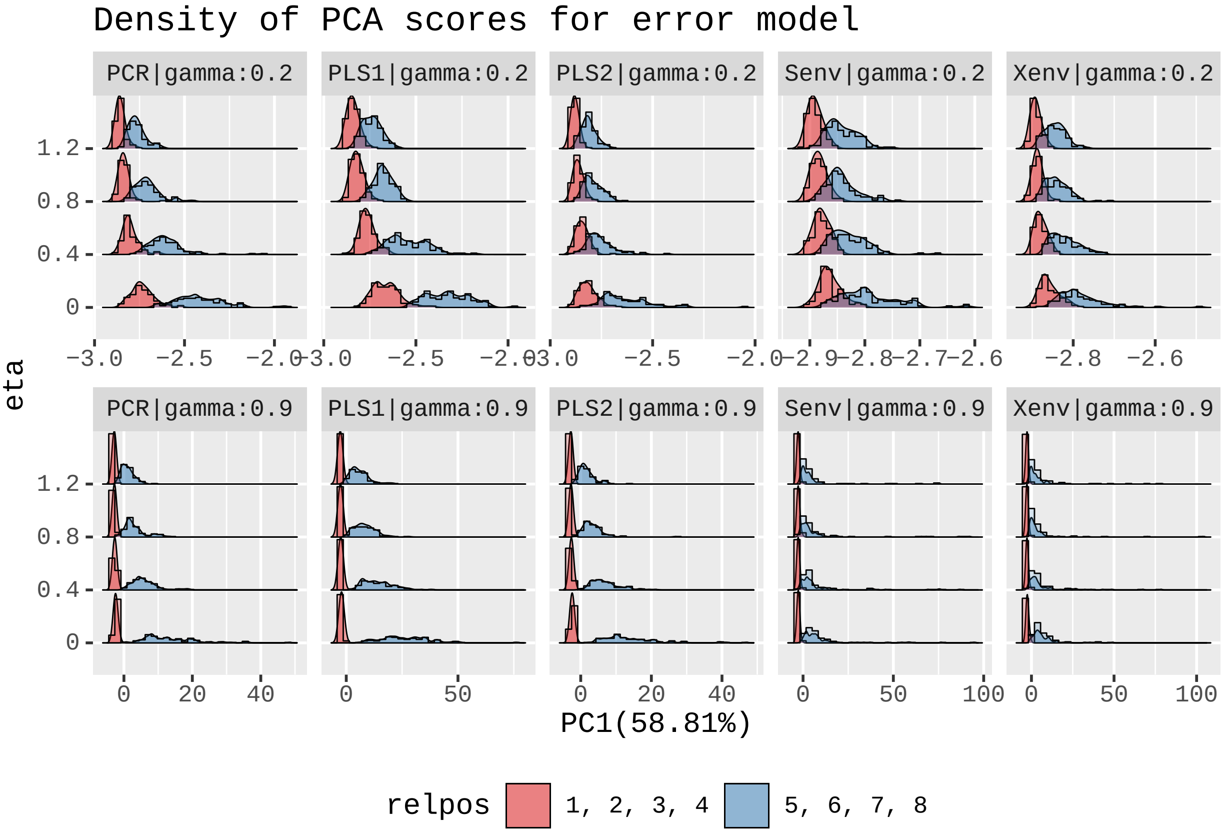

Figure 3: Scores density corresponding to first principal component of *error dataset* (**u**) subdivided by `methods`, `gamma` and `eta` and grouped by `relpos`.

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 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 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

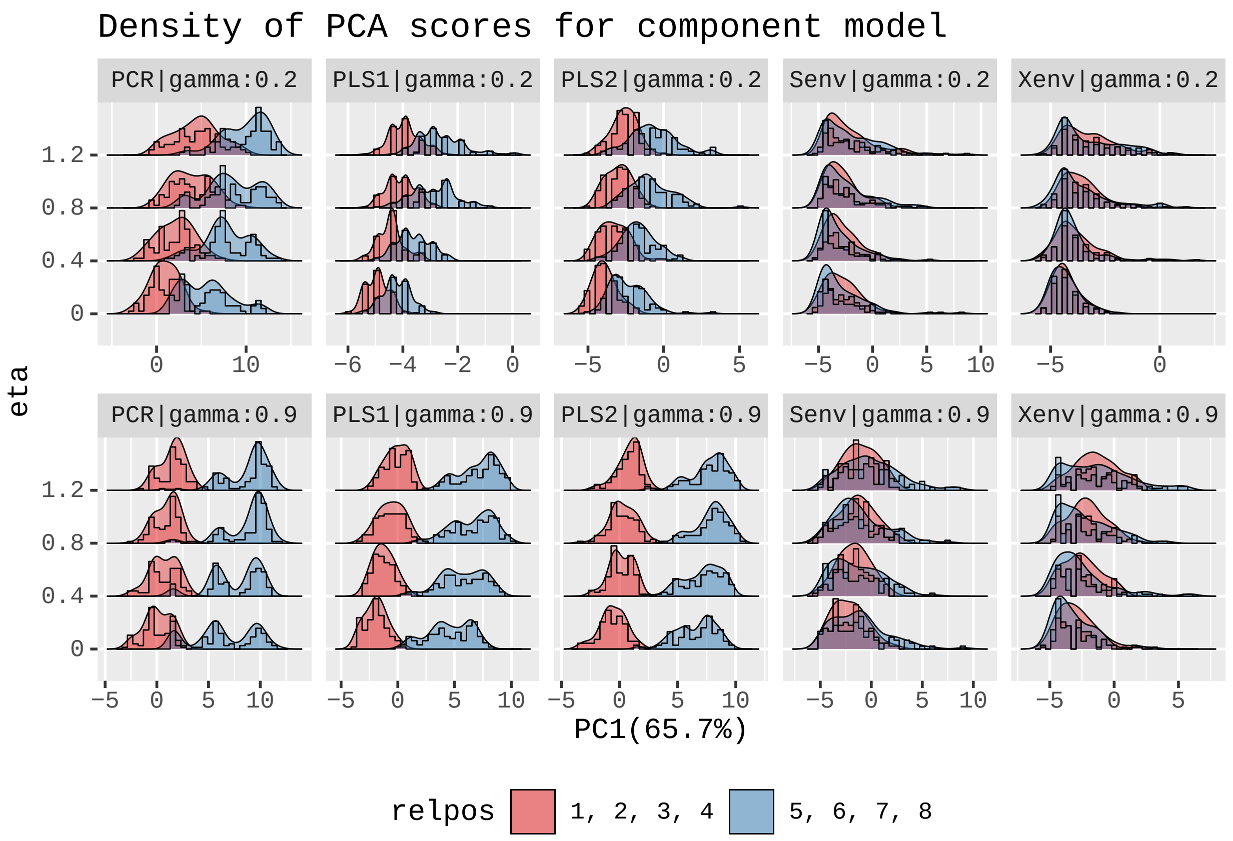

Figure 4: Score density corresponding to first principal component of *component dataset* (**v**) subdivided by `methods`, `gamma` and `eta` and grouped by `relpos`.

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 6 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 com-

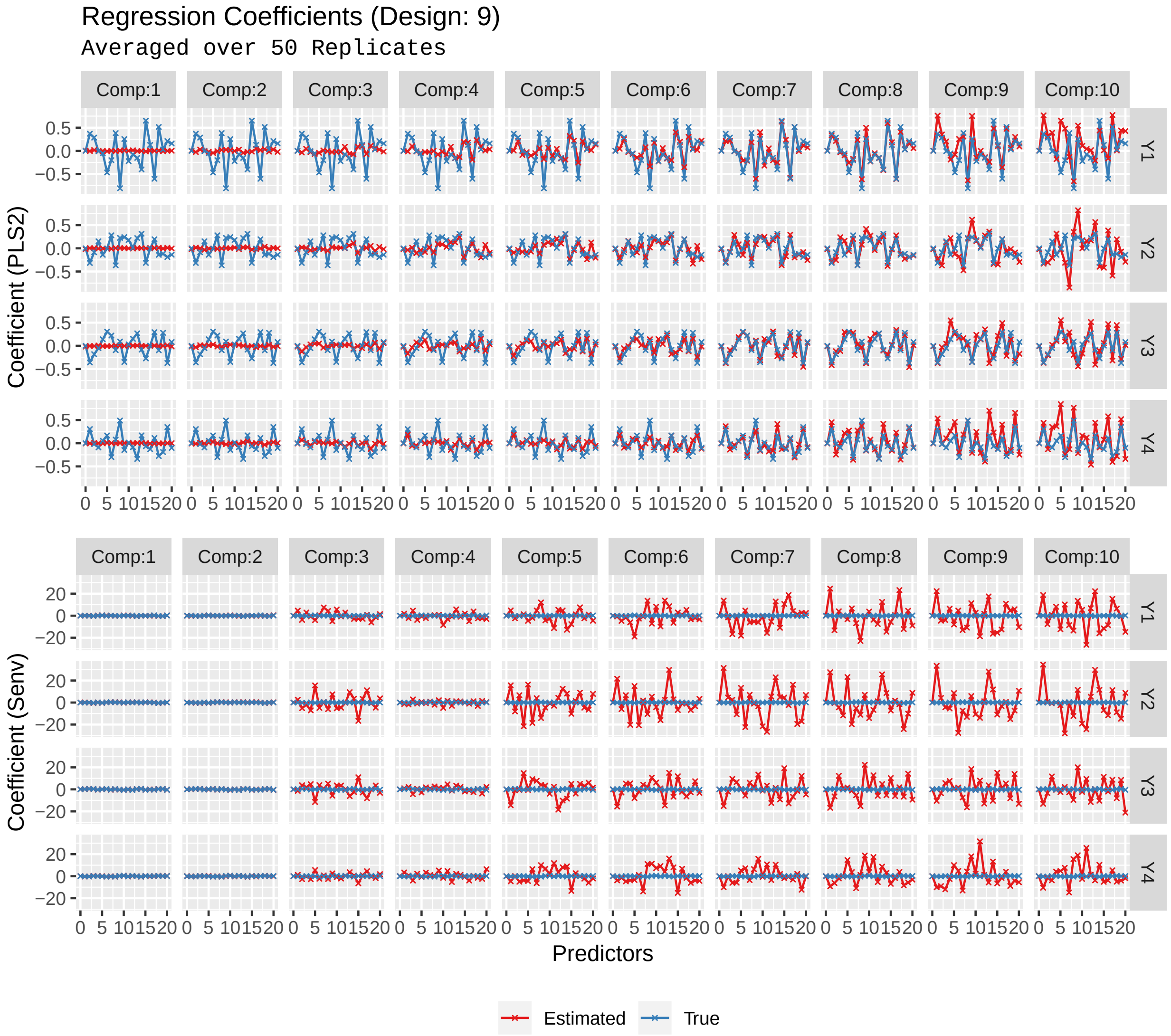

Figure 5: Regression Coefficients estimated by PLS2 and Simultaneous methods on the data based on Design 9.

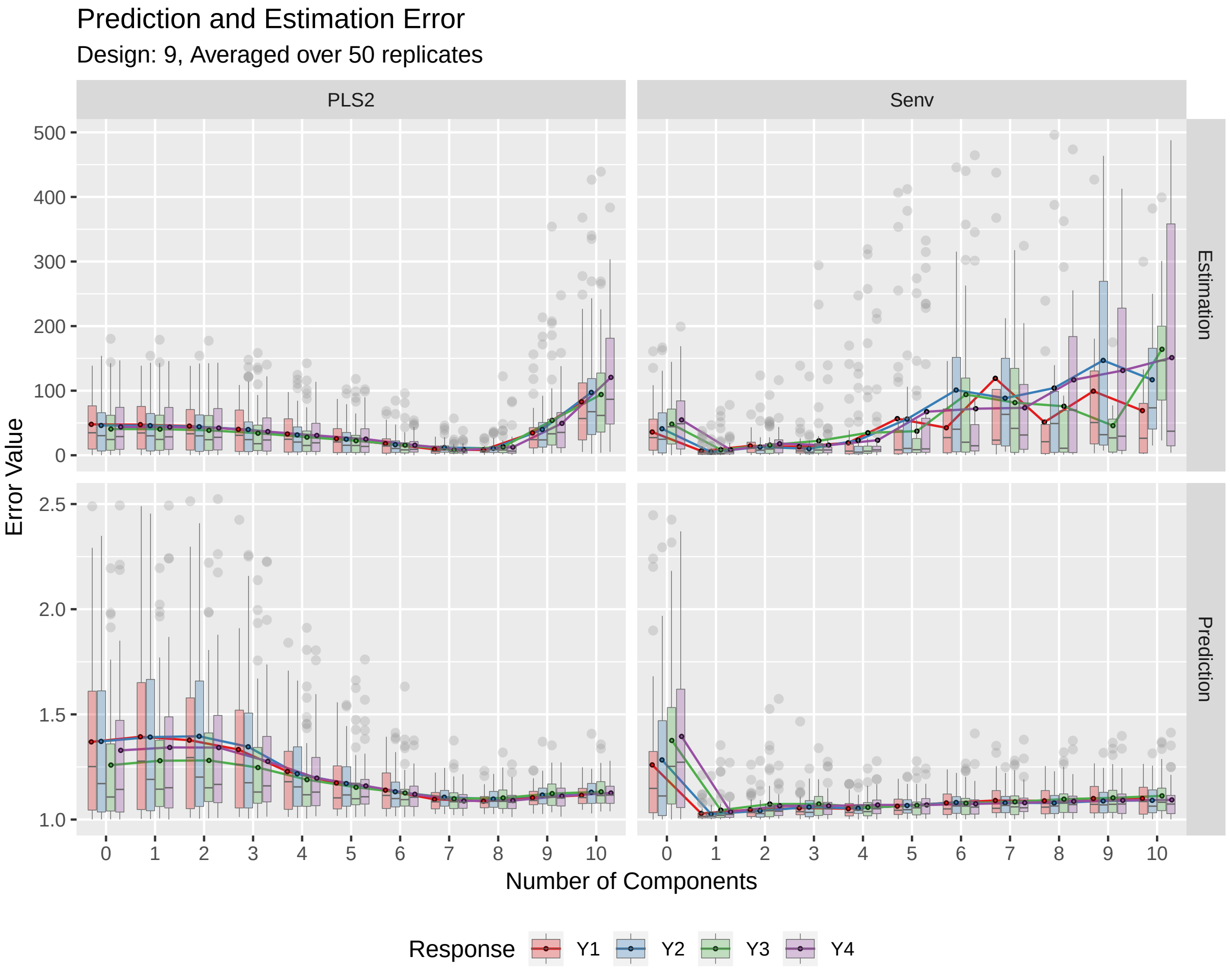

Figure 6: Minimum prediction and estimation error for PLS2 and Simultaneous methods. The point and lines are averaged over 50 replications.

ponents 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 5). Despite having contrast modelling result for a dataset with similar properties, the minimum errors produced by them are comparable (See Table 1).

The Figure 6 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

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)

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).

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

7.1. Dataset for Analysis

7.2. Discuss Model

7.3. Interpretation of the fitted MANOVA Model

7.4. Effect of MANOVA

7.4.1. Error Model

- Effect analysis of estimation error model
- Tie up these results with prediction error in previous paper

7.4.2. Component Model

- Effect analysis of number of component model
- Tie up these results in previous paper

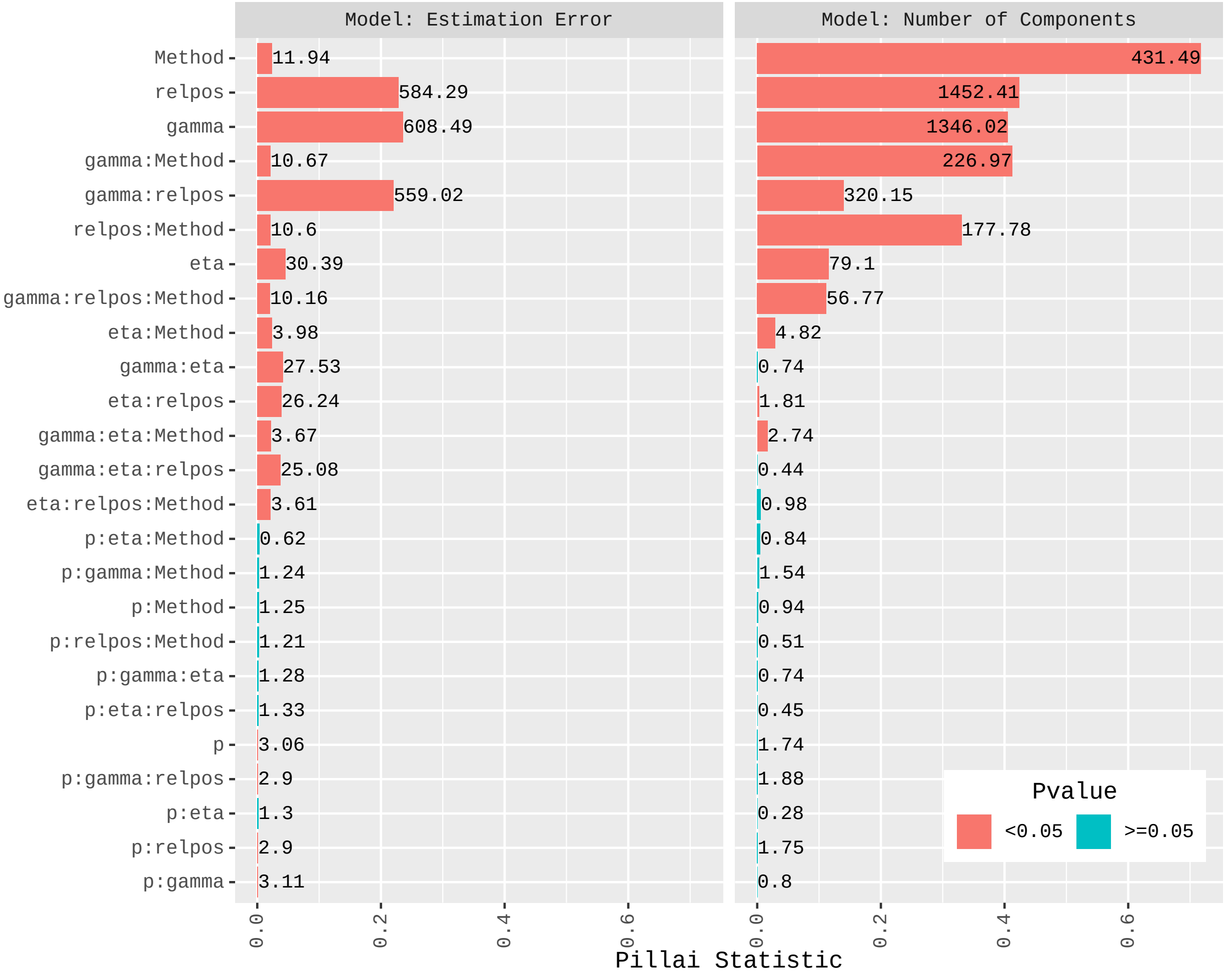

Figure 7: Pillai Statistic and F-value for the MANOVA model. The bar represents the Pillai Statistic and the text labels are F-value for corresponding factor.

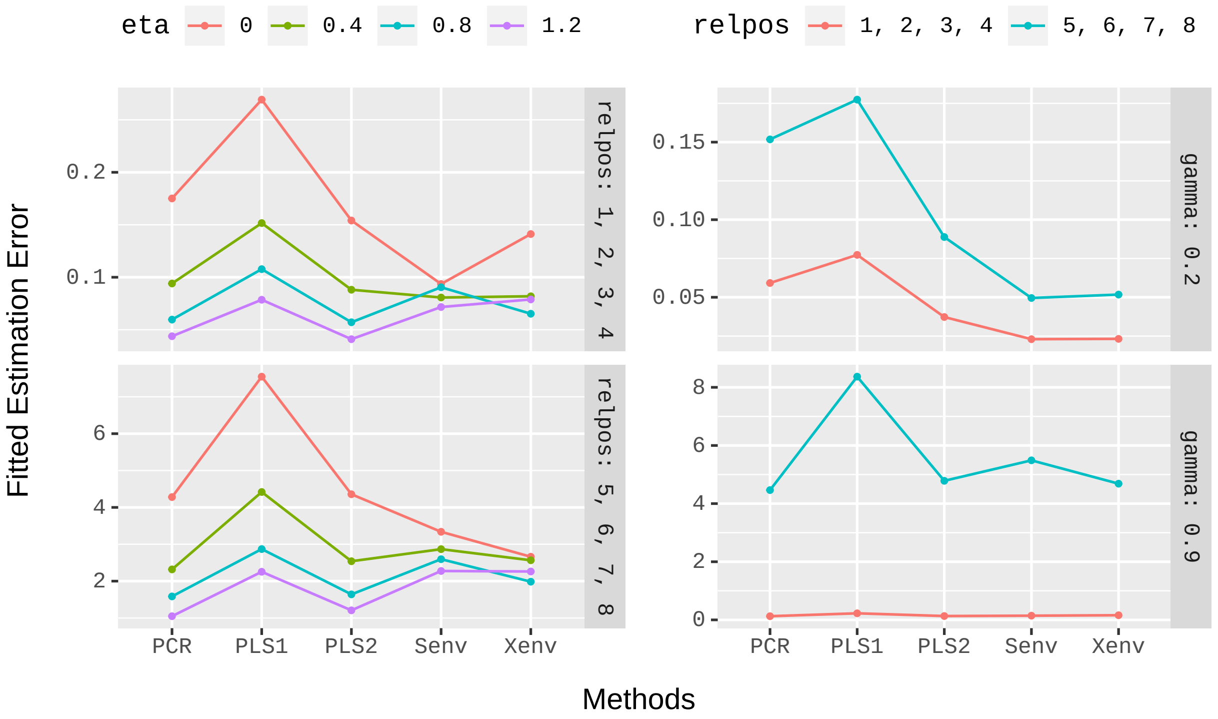

Figure 8: Effect plot of some interactions of the multivariate linear model of estimation error

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., 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., 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.

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/.

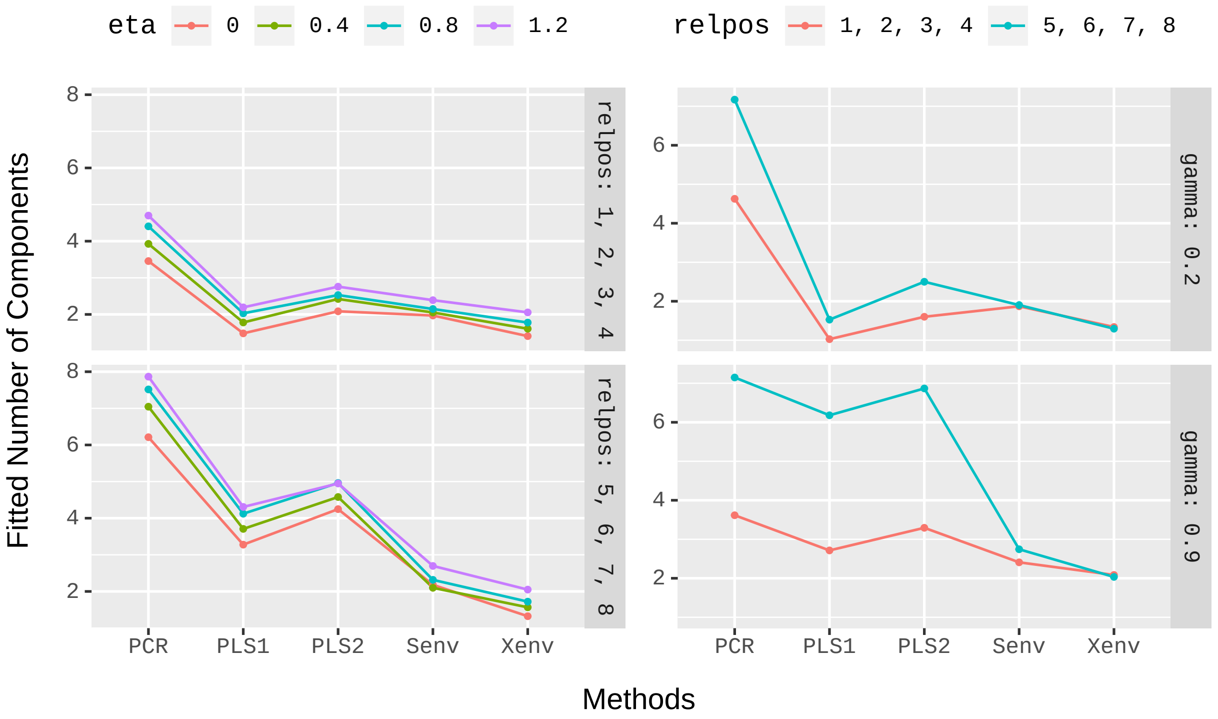

Figure 9: Effect plot of some interactions of the multivariate linear model of number of components to get minimum prediction error

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.