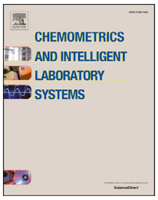
Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics



Comparison of multi-response estimation methods

Raju Rimal [*], Trygve Almøy, Solve Sæbø

Faculty of Chemistry and Bioinformatics, Norwegian University of Life Sciences, Ås, Norway

ARTICLE INFO

Keywords:
Model-comparison
Multi-response
Simrel
Estimation
Estimation error
Meta modelling
Envelope estimation

ABSTRACT

Prediction performance does not always reflect the estimation behaviour of a method. High error in estimation may necessarily not result in high prediction error, but can lead to an unreliable prediction if test data lie in a slightly different subspace than the training data. In addition, high estimation error often leads to unstable estimates, and consequently, the estimated effect of predictors on the response can not have a valid interpretation. Many research fields show more interest in the effect of predictor variables than actual prediction performance. 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, not surprisingly, found that no single method is superior to others, but each has its strength for some specific nature of data. In addition, the newly developed envelope method has shown impressive results in finding relevant information from data using significantly fewer components than the other methods.

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 medicine show more interest in measuring the impact of certain indicators or variable 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 to improving the trust in the findings of modern data science research [13]. This makes the assessment of measurement error, inference and prediction even more essential.

Technology has facilitated researchers to collect large amounts of data, however, often such data either contains irrelevant information or are highly redundant. Researchers are devising new estimators to extract information and identify their inter-relationship. Some estimators are robust towards fixing the multicollinearity (redundancy) problem, while others are targeted to model only the relevant information contained in the response variable.

This study extends [22] with a similar multi-response, linear regression model setting and compares some well-established estimators such as Principal Components Regression (PCR), Partial Least Squares (PLSR) Regression, together with two new methods based on envelope estimation: Envelope estimation in predictor space (Xenv) [6] and simultaneous estimation of the envelope (Senv) [7]. The estimation processes of these methods are discussed in the Estimation Methods section. The comparison is aimed at 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 the 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 as used by Rimal et al. [22]. The data are simulated from a multivariate normal distribution where we assume that the variation in a response vector-variable **y** is partly explained by the predictor vector-variable **x**. However, in many situations, only a subspace of the predictor space is relevant for the variation in the response **y**. This space can be referred to as the relevant space of **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

[*] Corresponding author.
E-mail addresses: raju.rimal@nmbu.no (R. Rimal), trygve.almoy@nmbu.no (T. Almøy), solve.sabo@nmbu.no (S. Sæbø).

https://doi.org/10.1016/j.chemolab.2020.104093
Received 19 July 2019; Received in revised form 22 May 2020; Accepted 22 June 2020
Available online 10 August 2020

relevant space in predictor can explain (Fig. 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. Næs and Martens [19] introduced the concept of relevant components, which was explored further by Helland [9]; Næs and Helland [18]; Helland and Almøy [11] and Helland [10]. The corresponding eigenvectors were referred to as relevant eigenvectors. A similar logic is introduced by Cook et al. [6] and later by Cook et al. [4] as an envelope, as space spanned by the relevant eigenvectors [3]; p.101). See Rimal et al. [21]; Sæbø et al. [23] and Rimal et al. [22] 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,

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

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 define the relationship between $\mathbf{x}$ and $\mathbf{y}$. With n samples, the least-squares estimate of $\boldsymbol{\beta}$ can be written as $\widehat{\boldsymbol{\beta}} = \mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}$. Here, as in many situations, the estimator $\mathbf{S}_{xx}$ for $\boldsymbol{\Sigma}_{xx}$ can either be non-invertible or have small eigenvalues. In addition, $\mathbf{S}_{xy}$, the estimator of $\boldsymbol{\Sigma}_{xy}$, is often influenced by a high level of noise in the data. In order to solve these problems, various methods have adopted the concept of relevant space to identify the relevant components through the reduction of the dimension in either $\mathbf{x}$ or 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 to span the reduced relevant space. Since PCR is based on capturing the maximum variation in predictors for every component it has added to the model, this method does not consider the response structure in the model reduction [14]. In addition, if the relevant components are not corresponding to the largest eigenvalues, the method requires a larger number of components to make precise prediction [1].

Partial Least Squares (PLS) regression aims to maximize the covariance between the predictor and response components (scores) [15]. Broadly speaking, PLS can be divided into PLS1 and PLS2 where the former tries to model the response variables individually, whereas the latter uses all the response variable together while modelling. Among the three widely used algorithms NIPALS [24], SIMPLS [15] and KernelPLS [16], we will be using KernelPLS for this study, which gives equivalent results to the classical NIPALS algorithm and is default in R-package pls [17].

Envelopes was first introduced by Ref. [5] as the smallest subspace that includes the span of true regression coefficients. The *Predictor Envelope* (Xenv) identifies the envelope as a smallest subspace in the predictor space, by separating the predictor covariance $\boldsymbol{\Sigma}_{xx}$ into relevant (material) and irrelevant $\mathbf{y}$ (immaterial) parts, such that the response 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 the optimization of an objective function. Further, the regression coefficients are estimated using only the relevant part. Cook et al. [6]; Cook et al. [4] and Cook [3] have extensively discussed the foundation and various mathematical constructs together with properties related to the Predictor Envelope.

Simultaneous Predictor-Response Envelope (Senv) implements the envelope in both the response and the predictor space. It separates the material and immaterial part in the response space and the predictor space such that the material part of the response does not correlate with the immaterial part of the predictor and the immaterial part of the 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 the number of responses, simultaneous envelope reduces to the predictor envelope, and if the number of predictor components equals the number of predictors, the

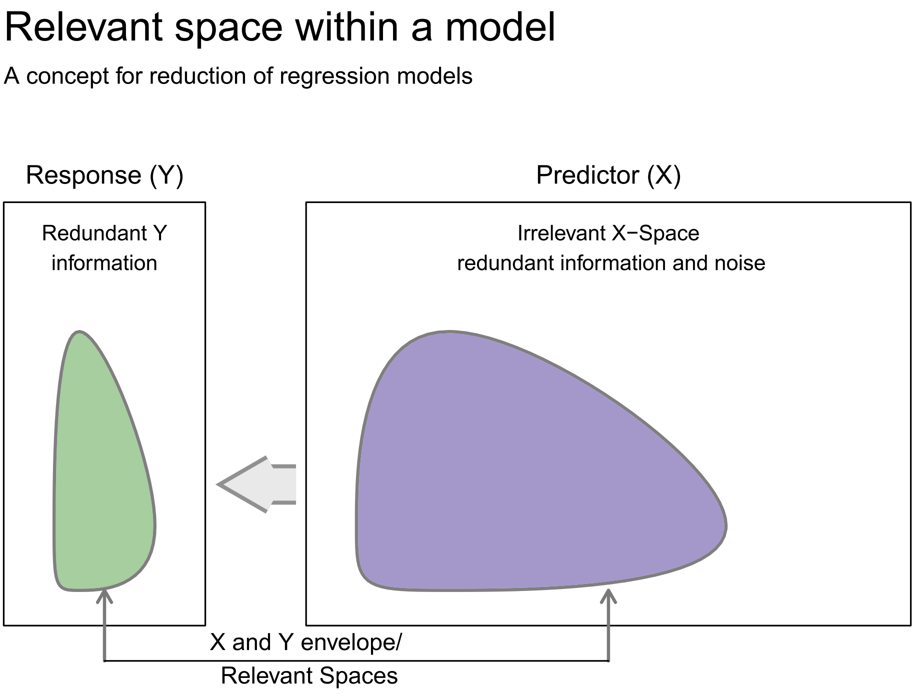

Relevant space within a model
A concept for reduction of regression models
Response (Y)
Predictor (X)
Redundant Y information
Irrelevant X−Space redundant information and noise
X and Y envelope/ Relevant Spaces

Fig. 1. Relevant space in a regression model.

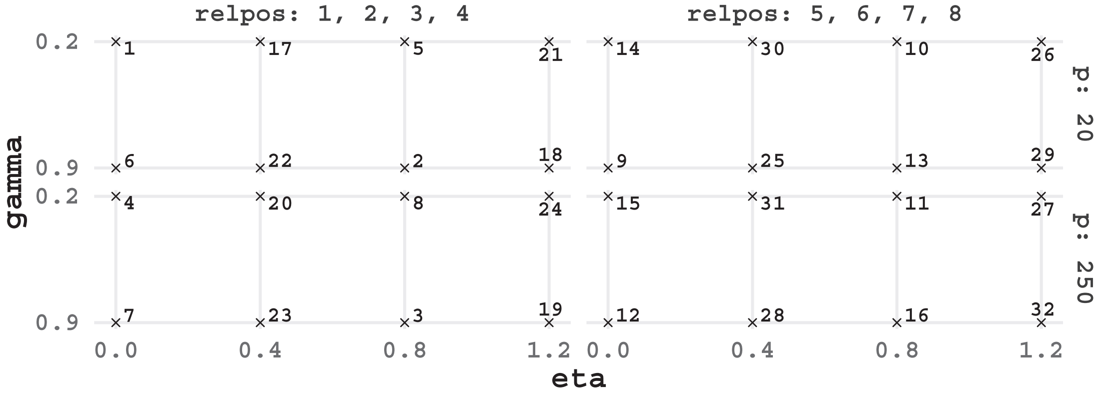


Fig. 2. Experimental Design of simulation parameters. Each point represents a unique data property.

result will be equivalent to ordinary least squares. Cook and Zhang [7] and Cook [3] have discussed the method in detail. Further, Helland et al. [12] have discussed when and under which condition the population models of PCR, PLS and Xenv are equivalent.

Here, each methods uses different strategy for estimating the regression coefficients due to which the optimal number of components they determine will be different. For example, PCR method captures the maximum variation in predictor matrix **y** in every subsequent components while PLS methods focus more on the variation in predictors that are relevant for the responses. The envelope methods construct the envelope as a linear combination of relevant eigenvectors. This allows them to reduce the dimension even further and consequently these methods need fewer components.

4. Experimental Design

An R [20] package simrel [21,23] is used to simulate the data for comparison. In the simulation, number of response variables and number of observations $n = 100$ are fixed, and the following four simulation parameters are varied to obtain 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 (γ) controls the exponential decline of eigenvalues in $\boldsymbol{\Sigma}_{xx} (\lambda_i, i = 1, \dots p)$ as,

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

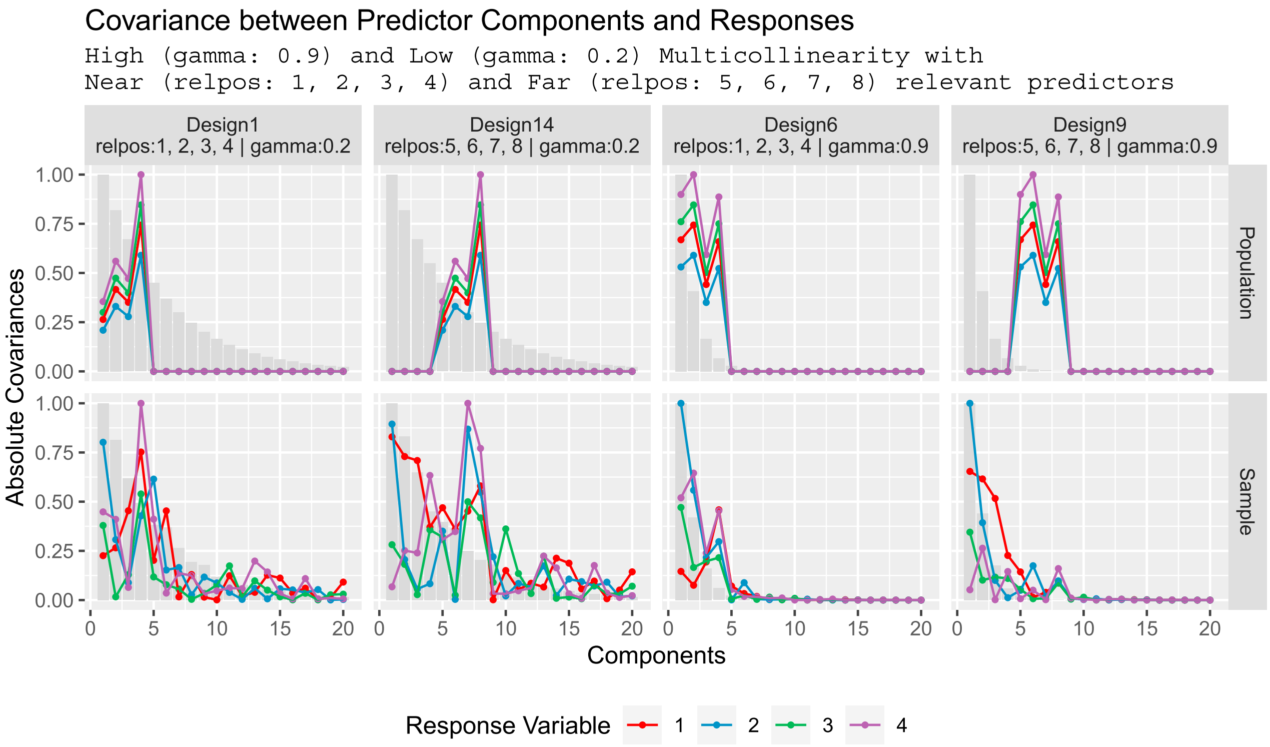


Fig. 3. Covariance between predictor components and each response variable in the population (top), and in the simulated data (bottom) for four different designs. The bars in the background represent the variance of the corresponding components (eigenvalues).

Two levels, 0.2 and 0.9, of gamma are used for simulation so that level 0.2 simulates data with low multicollinearity and 0.9 simulates the data with high multicollinearity in **x** respectively.

Position of relevant components: (relpos) Initial principal components of a non-singular covariance matrix have higher variance than the later ones. If the principal components corresponding to predictors with larger variation are not relevant for a response, this will just increase the noise level in the data. Here $wm = 4e$ will use two different levels of a position index of true predictor components (relpos): a) 1, 2, 3, 4 and b) 5, 6, 7, 8. Predictor components irrelevant for a response make prediction difficult [11]. When combined with multicollinearity, this factor can create both easy and difficult cases for both estimation and prediction.

Correlation in response variables: (eta) Some estimators also use the dependence structure of response for estimation. Here the correlation between the responses is varied through a simulation parameter eta (η). The parameter controls the exponential decline of eigenvalues $\kappa_j, j = 1, \ldots m($ number of responses$)$ of $\mathbf{\Sigma}_{yy}$ as,

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

Four levels 0, 0.4, 0.8 and 1.2 of eta (η) are used in the simulations. Level $\eta = 0$ gives data with uncorrelated response variables, while $\eta = 1.2$ gives highly correlated response variables.

Using these simulation parameters, a latent covariance matrix is constructed as in 5.

$$\begin{bmatrix}\mathbf{w}\\ \mathbf{z}\end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix}\mu_w\\ \mu_z\end{bmatrix}, \quad \begin{bmatrix}\mathbf{\Sigma}_{ww} & \mathbf{\Sigma}_{wz}\\ \mathbf{\Sigma}_{zw} & \mathbf{\Sigma}_{zz}\end{bmatrix}\right) \tag{5}$$

For example, $\eta = 1.2$ gives $\mathbf{\Sigma}_{ww}$ as a diagonal matrix with 1, 0.3, 0.09, 0.03 in its diagonal. However for $\eta = 0$, $\mathbf{\Sigma}_{ww}$ will be an identity matrix. A similar approach is used for covariance matrix $\mathbf{\Sigma}_{zz}$. In addition, when the true relevant components are at position 1, 2, 3, 4, the first row of covariance matrix $\mathbf{\Sigma}_{wz}$ with dimension $m \times p$ will have $\sigma_{11}, \sigma_{12}, \sigma_{13}$ and σ_{14} in its first four columns and the rest are filled with zeros. These σ values are the links that defines the relationship between the latent components of predictors and the first response component. Two random orthogonal rotation matrices **R** and **Q** are used to rotate the latent covariance matrices in order to obtain the covariance matrices in 1. Rimal et al. [21] have discussed the underlying mechanism in details.

Here we have assumed that there is only one informative response component. Hence the relevant space of the response matrix has dimension one. For the predictors, there are 4 true relevant components, so the relevant space for predictor matrix has 4 dimension. In the discussion onwards, *number of components* refers to the number of predictor components unless otherwise stated. In addition, the coefficient of determination is fixed at 0.8 for all datasets.

A complete factorial design is adopted using the different levels of factors discussed above to create 32 designs (Fig. 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 × 32 × 50 i.e., 8000 datasets are simulated.

The simulation properties are directly reflected in the simulated data. For example, in Fig. 3, design pairs 1 and 14 as well as 6 and 9 differ in their properties only in terms of position 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 population properties are also reflected in the simulated samples (bottom row Fig. 3). The

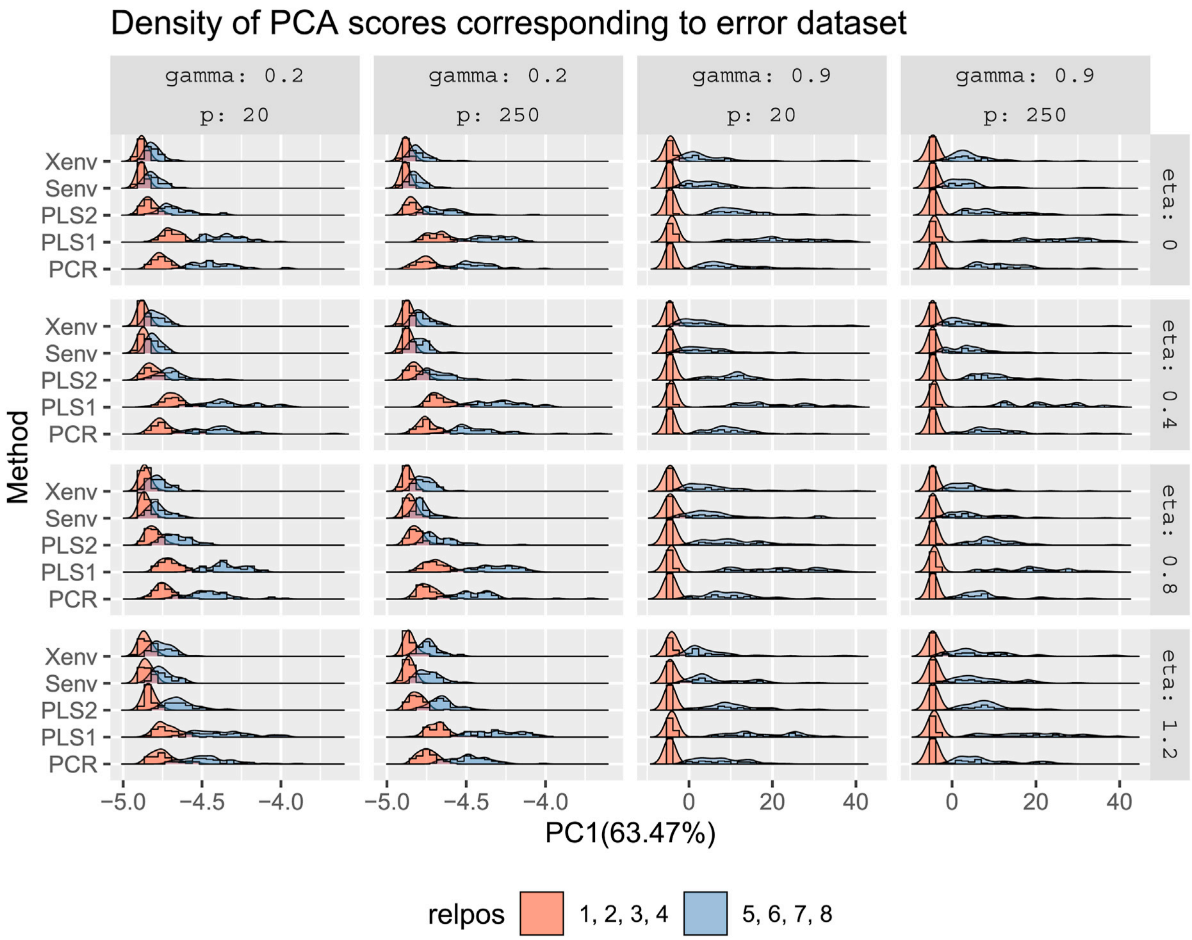


Fig. 4. Scores density corresponding to first principal component of *error dataset* (**u**) subdivided by methods, gamma and eta and grouped by relpos.

combination of these factor levels creates datasets that are easy or difficult with regard to estimation and prediction. We observe from Fig. 3 that it may be difficult to infer the structure of the latent relevant space of **x** from the estimated principal components and their estimated covariances with the observed responses.

5. Basis of comparison

The focus of this study is to extend the exploration of Rimal et al. [22] 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 (7)
(b) the average number of components used by the methods to give minimum estimation error

Let us define the expected estimation error as

$$\text{MSE}(\widehat{\boldsymbol{\beta}})_{ijkl} = \text{E}\left[\frac{1}{\sigma^2_{y_j}}\left(\beta_{ij} - \widehat{\boldsymbol{\beta}}_{\text{ijkl}}\right)^t\left(\beta_{ij} - \widehat{\boldsymbol{\beta}}_{\text{ijkl}}\right)\right] \tag{6}$$

for response $j = 1, \dots 4$ in a given design $i = 1, 2, \dots 32$ and method $k = 1(PCR), \dots 5(Senv)$ using $l = 0, \dots 10$ number of components. Here $\sigma^2_{y_j}$ is the variance of response *j*. Since both the expectation and the variance of $\widehat{\boldsymbol{\beta}}$ are unknown, the estimation error is estimated using data from 50 replications as follows,

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

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

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

Our further discussion revolves around what we will refer to as the *Error Dataset* and the *Component Dataset,* as in the prediction comparison paper Rimal et al. [22]. For a given estimation method, design, and response, the component that gives the minimum estimation error averaged over all replicates is selected as,

$$l_\circ = \underset{l}{\text{argmin}}\left[\frac{1}{50}\sum_{r=1}^{50}\widehat{\text{MSE}_\circ(\widehat{\boldsymbol{\beta}})}_r\right] \tag{8}$$

Here we have skipped further indices on $\widehat{\boldsymbol{\beta}}$ for brevity. The estimation error $\widehat{\text{MSE}_\circ(\widehat{\boldsymbol{\beta}})}$ for every method, design and response corresponding to component $l_\circ$, computed as (8), is then regarded as the *error dataset* in the subsequent analysis. Let $\mathbf{u}_{8000\times 4} = (u_j)$, where u_j is the j^{th} column of $\mathbf{u}$ denoting the estimation error corresponding to response $j = 1, \dots 4$ in the context of this dataset. Further, let the number of components that result in minimum estimation error in each replication and computed as (9), comprise the *component dataset.* Let $\mathbf{v}_{8000\times 4} = (v_j)$ where v_j is the j^{th} column of $\mathbf{v}$ denoting the outcome variable measuring the number of components used to obtain minimum estimation error corresponding to response $j = 1, \dots 4$.

$$l_\circ = \underset{l}{\text{argmin}}\left[\widehat{\text{MSE}_\circ(\widehat{\boldsymbol{\beta}})}\right] \tag{9}$$

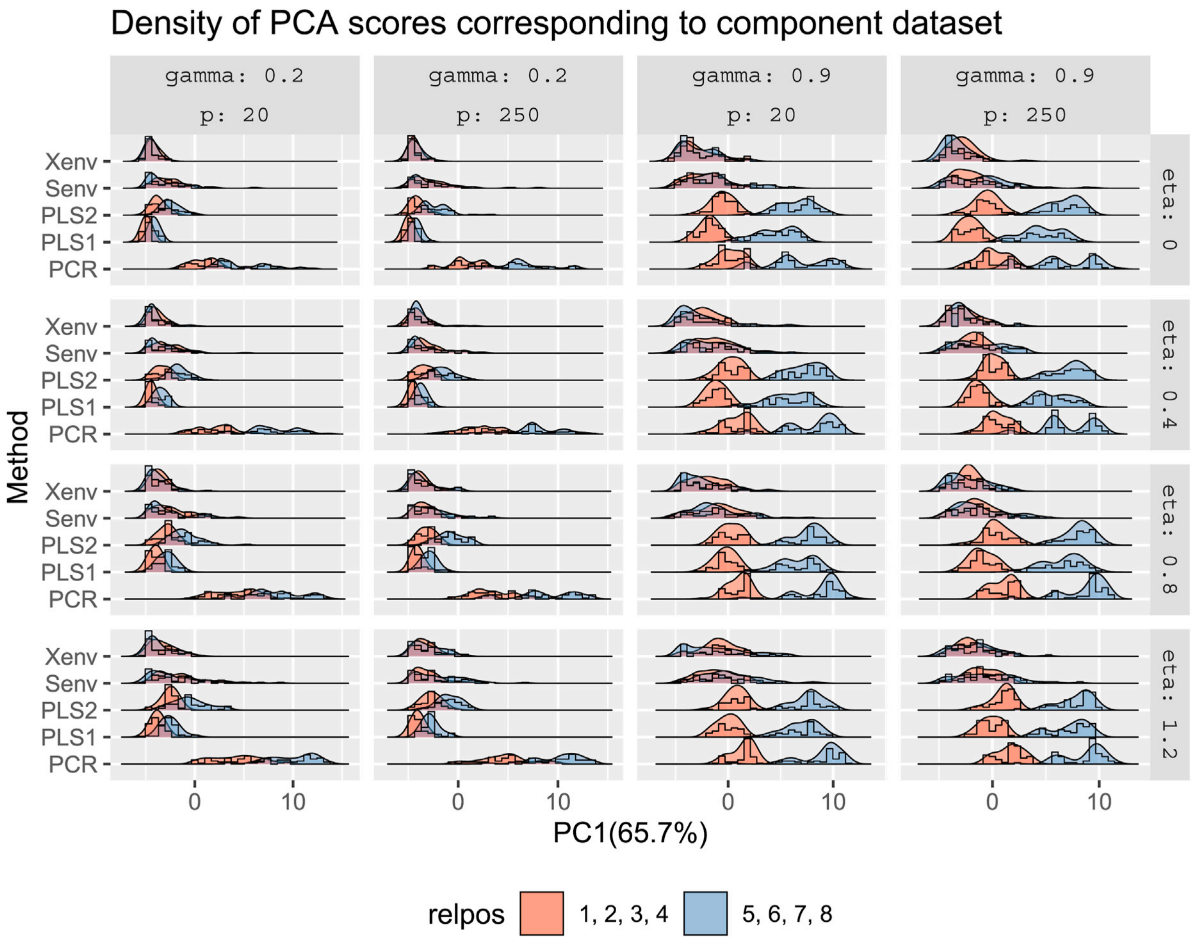


Fig. 5. Score density corresponding to the first principal component of *component dataset* (**v**) subdivided by methods, gamma and eta and grouped by relpos.

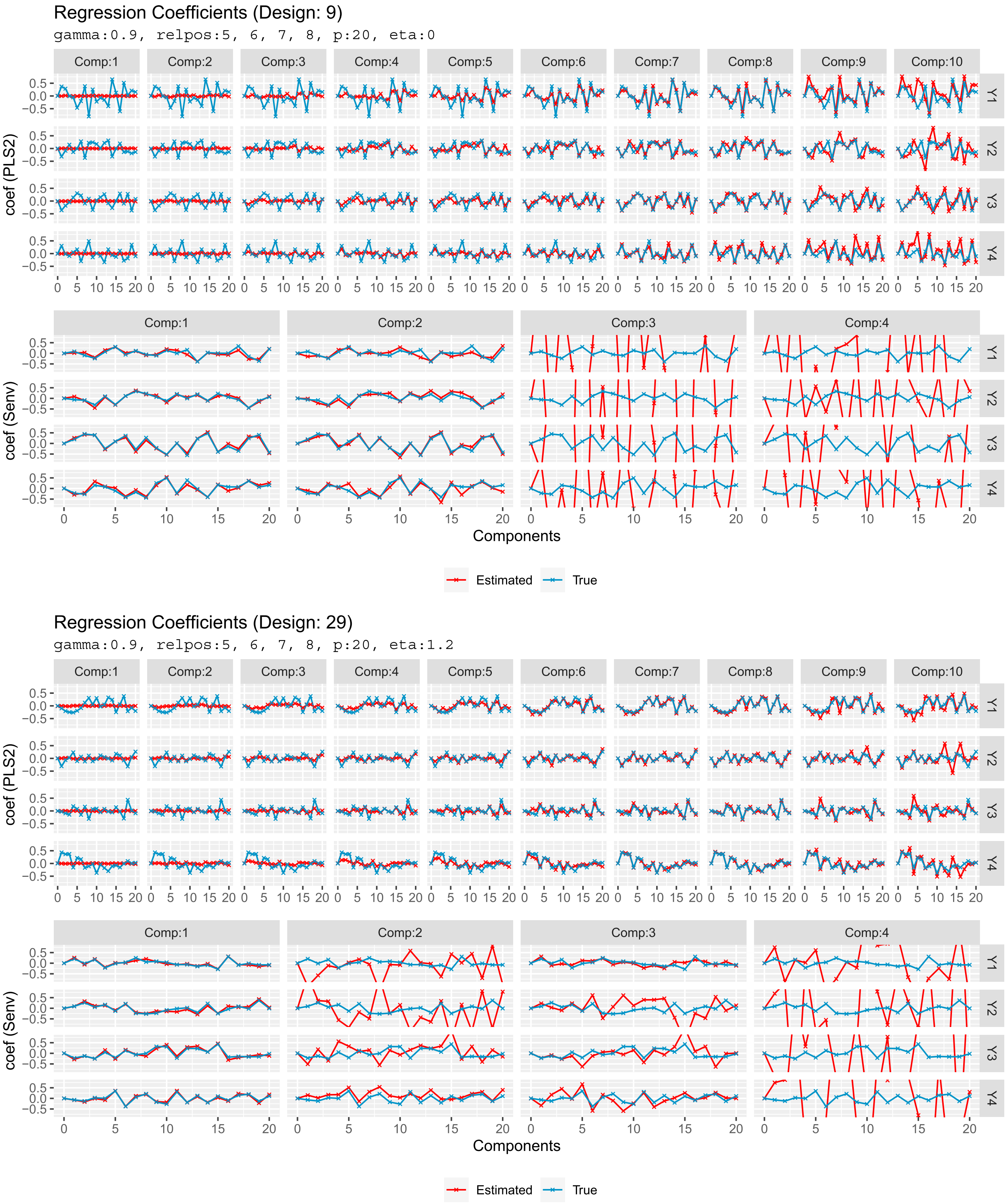


Fig. 6. Regression Coefficients (coef) estimated by PLS2 and Simultaneous Envelope methods on the data based on Design 9 and 29.

6. Exploration

In this section we explore the variation in the *error dataset* and the *component dataset* by means of Principal Component Analysis (PCA). Let $\mathbf{t}_u$ and $\mathbf{t}_v$ be matrices holding the column vectors of the principal component scores corresponding to the **u** and **v** matrices, respectively. The density of the scores in Fig. 4 and Fig. 5 correspond to the first principal component of **u** and **v**, i.e. the first column of $\mathbf{t}_u$ and $\mathbf{t}_v$ respectively. Here higher scores correspond to larger estimation error and vice

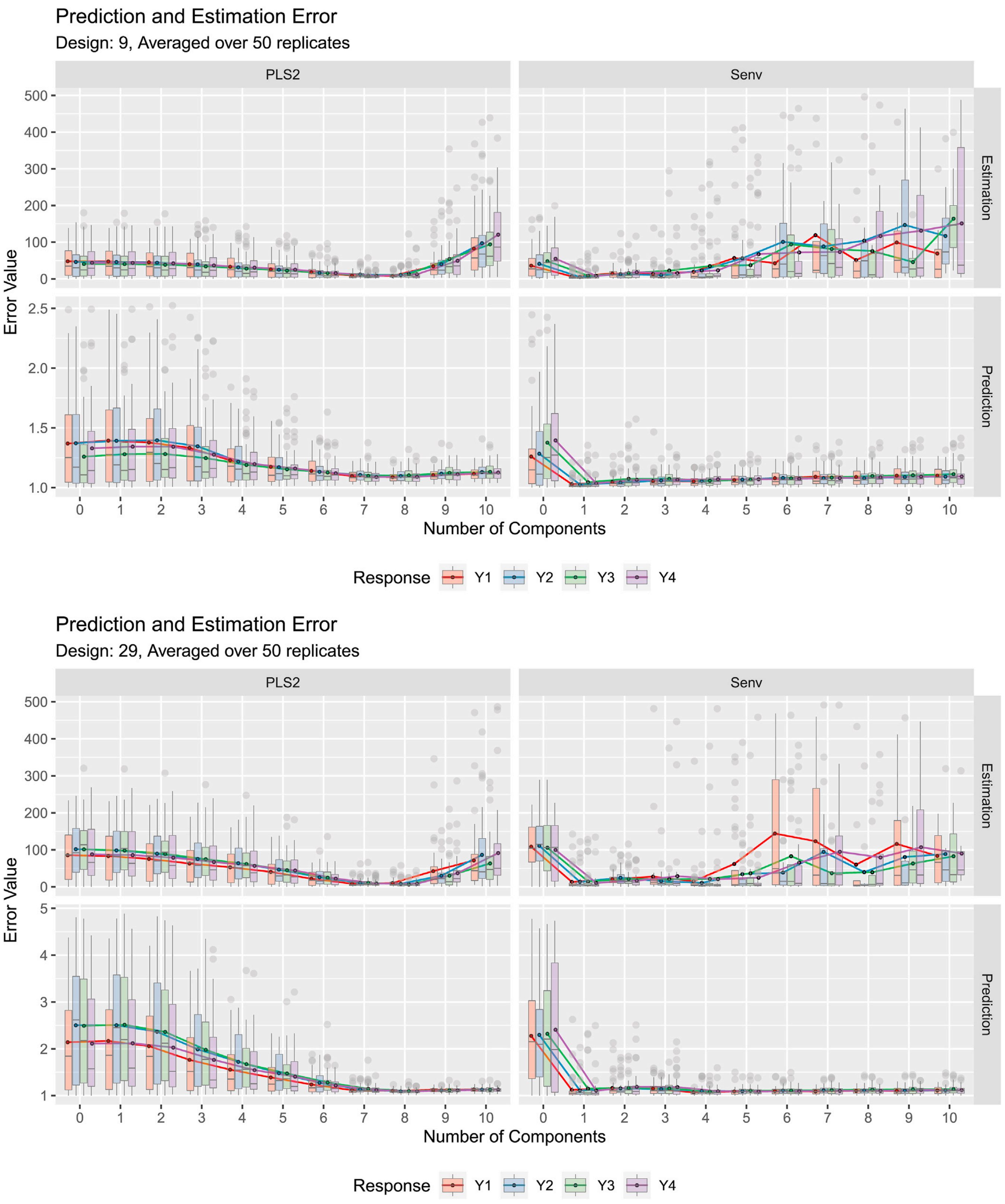


Fig. 7. Minimum prediction and estimation error for PLS2 and Simultaneous Envelope methods. The point and lines are averaged over 50 replications.

versa.

Fig. 4 shows a clear difference in the effect of low and high multicollinearity on estimation error. In the case of low multicollinearity (gamma: 0.2), the estimation errors are in general smaller and have lesser variation compared to high multicollinearity (gamma: 0.9). In particular we observe that the envelope methods have small estimation errors in the low multicollinearity cases compared to the other methods.

Furthermore, position of the relevant predictor components has a noticeable effect on estimation error for all methods. When relevant predictors are at position 5, 6, 7, 8, the components at positions 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 with highly collinear predictors (gamma = 0.9). 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 [22].

In the case of the *component dataset* (Fig. 5), PCR, PLS1 and PLS2 methods have in general used a larger number of components in the case of high multicollinearity compared to low. Surprisingly, the envelope methods (Senv and Xenv) have mostly used a distinctly smaller number of components in both cases of multicollinearity compared to other methods.

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

A clear interaction between the position of relevant predictors and the multicollinearity, which 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 predictor variables. Such performance is also seen the case of prediction error (See Rimal et al. [22], however, the number of components used for optimization of prediction is smaller than in the case of estimation. Even when the relevant components 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 components at position 1, 2, 3, 4. This shows that the envelope methods identify the predictor space relevant to the response differently, from the other methods and with very few numbers of latent components. This is particularly the case when multicollinearity in $\mathbf{x}$ is high.

The following sub-section explores in particular the prediction and estimation errors and the estimated regression coefficient of Simultaneous Envelope and Partial Least Squares for a design having high multicollinearity, and with predictor components at positions 5, 6, 7, 8. Here we will use the design with $n > p$ and two levels of correlation between the responses. These correspond to Design-9 and Design-29 in our simulations.

Fig. 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 only one to two components and the corresponding estimated regression coefficients approximate the true coefficients (Fig. 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, in design 29, the prediction error corresponding to the envelope methods are much smaller. In both of these designs and in prediction and estimation error, Xenv has equally and better in some responses than Senv. This difference needs further exploration in the case where there are more than one true response dimension.

Table 1
Minimum prediction and estimation error for design 9.

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)

In this study, 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.

Fig. 7 also shows in both designs that Senv has large estimation errors 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 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). Here, the variation in the estimation error can increase drastically also for PCR and PLS methods, when number of components more than 10 (not seen in the figure) is included. This is hinted in the estimation error plot (Fig. 7) for PLS2 for 8–10 number of components are included in the model.

In addition to the prediction and estimation error, Fig. 6 gives a closer view of how the average coefficients corresponding to these methods approximate to the true values. In the figure, PLS2 has used seven to eight components to reach the closest approximation to the true coefficients, but with increasing errors after including more components 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 more stable result in such situations. Further, the envelope methods have presented their ability to converge estimates to the true value in just one or two components. However, one should be cautious about determining the optimal number of components using method like cross-validation while working with real data.

Despite having a large variation in prediction and estimation error, the envelope based methods have produced a better result even for the difficult data cases as shown for Design 9.

7. Analysis

A statistical analysis using a Multivariate Analysis of variance (MANOVA) model is performed on both the *error dataset* and the *component dataset* in order to better understand the association between data properties and the estimation methods. Let the corresponding MANOVA models be termed as the *error model* (10) and the *component model* (11) in the following. Here the equations represent a heuristic representation of the MANOVA model and is closer to R-representation than a proper mathematical formulation. In the model, we will consider up-to the third order interaction of simulation parameters (p, gamma, eta, and relpos) and Method as is represented by cube notation. The models are fitted using correspondingly the *error dataset* ($\mathbf{u}$) and the *component dataset* ($\mathbf{v}$).

Error Model:

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

Component Model:

$$\mathbf{v} = (v_j) = \boldsymbol{\mu}_j + (\mathrm{p} + \mathrm{gamma} + \mathrm{eta} + \mathrm{relpos} + \mathrm{Method}s)^3 + \boldsymbol{\varepsilon} \tag{11}$$

where, $\mathbf{u}$ corresponds to the estimation errors in *error dataset* and $\mathbf{v}$ 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. [22]; we have also used Pillai's trace statistic for accessing the result of MANOVA. Fig. 8 plots the Pillai's trace statistics as bars with corresponding F-values as text labels. The leftmost plot corresponds to the *error model* and the rightmost plot corresponds to the *component model*. Here we use the custom R-notation indicating interactions up to order three for the parameters within the brackets.

Error Model: Unlike for the prediction error in Rimal et al. [22]; Method has a smaller effect, while the amount of multicollinearity, controlled by the gamma parameter, has a larger effect in the case of estimation error (Fig. 8). In addition, the position of relevant components and its interaction with the gamma parameters also have substantial effects on the estimation error. This also supports the results seen in the LABEL:Section:exploration. Exploration section where relevant predictors at position 5, 6, 7, 8 combined with high multicollinearity creates a large uninformative variance in the components 1, 2, 3, 4 making the design difficult with regards to estimation. 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 Method does not have a large impact on the estimation error, the *component model* in Fig. 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.

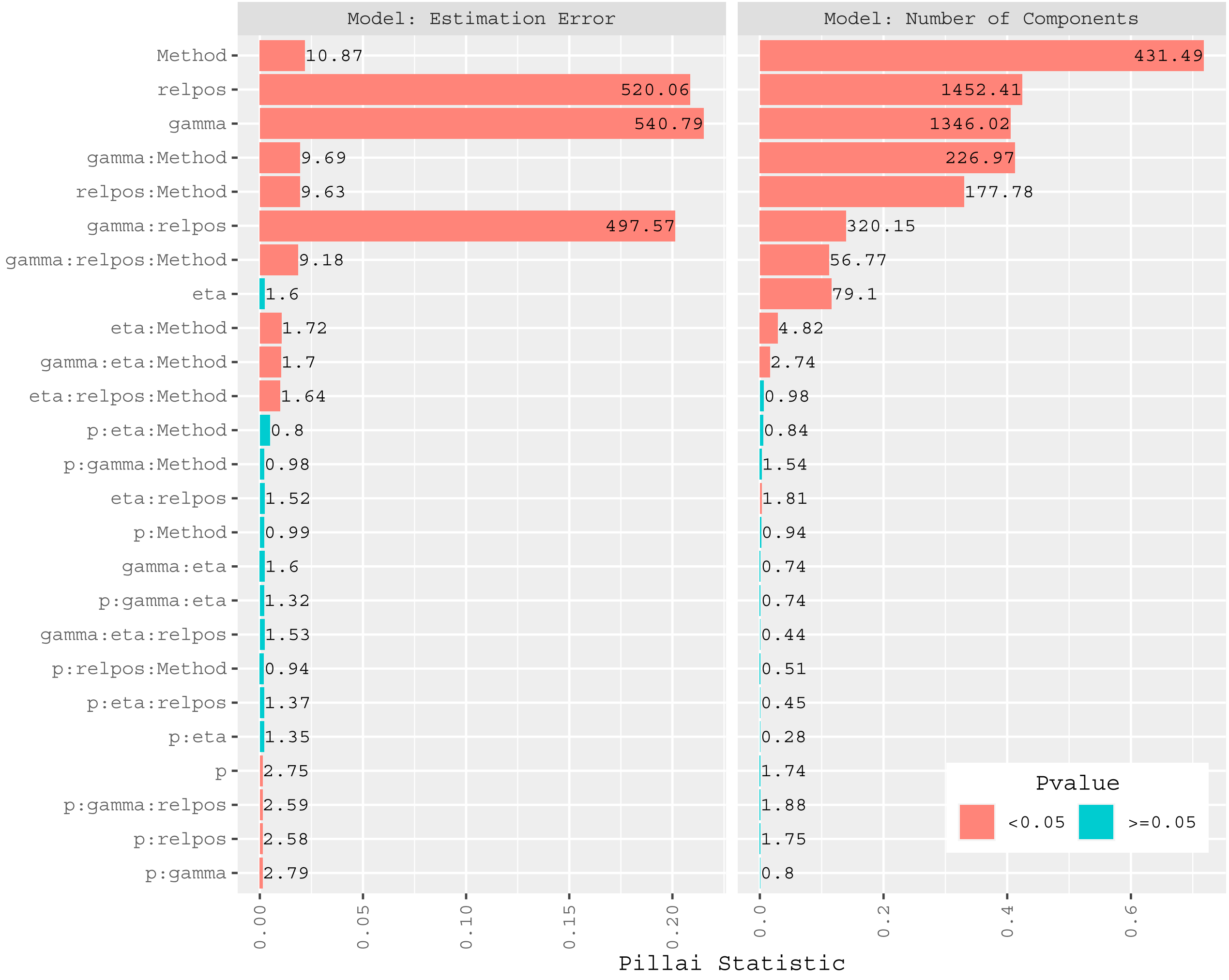


Fig. 8. Pillai Statistic and F-value for the MANOVA model. The bar represents the Pillai Statistic and the text labels are F-value for the corresponding factor.

[22]. 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 Fig. 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. In the case of designs with uncorrelated responses, envelope methods have on average smallest estimation errors. While PCR and PLS2, being somewhat invariant to the effect of this correlation structure, have performed better than the envelope methods in the designs with highly correlated responses.

For all methods, the error in the case of relevant predictors at positions 5, 6, 7, 8 is huge as compared to the case where relevant predictors are at positions 1, 2, 3, 4.

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

Finally, we note that the average estimation error corresponding to envelop methods in the designs with low multicollinearity is smaller than for the other methods.

7.2. Effect analysis of the component model

In the case of the 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 estimation error. Although the envelope methods have comparable 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 9, which we have considered in the previous section, has minimum estimation error for both envelope methods using only one predictor component. 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 Fig. 10. As seen previously, PCR uses, in general, a larger number of components than the other methods.

8. Discussion and conclusion

The overall performance of all 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 the methods. In addition, the high interaction of gamma with the relpos parameter suggests to carefully consider the number of relevant predictor components in the case of highly multicollinear data since this choice may have a large effect on the results. Although the interaction does not have this extent of influence in prediction, one should be careful about interpreting the estimates. In such cases, careful validation of model complexity, preferably using cross-validation or test data is advisable also for estimation purposes.

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. This indicates that the reduction of the response space becomes unstable with high collinearity between the responses for the envelope methods. Since the log likelihood objective function of envelope methods are non convex, the highly correlated responses might produce objective function with multiple maxima [8]. Despite the

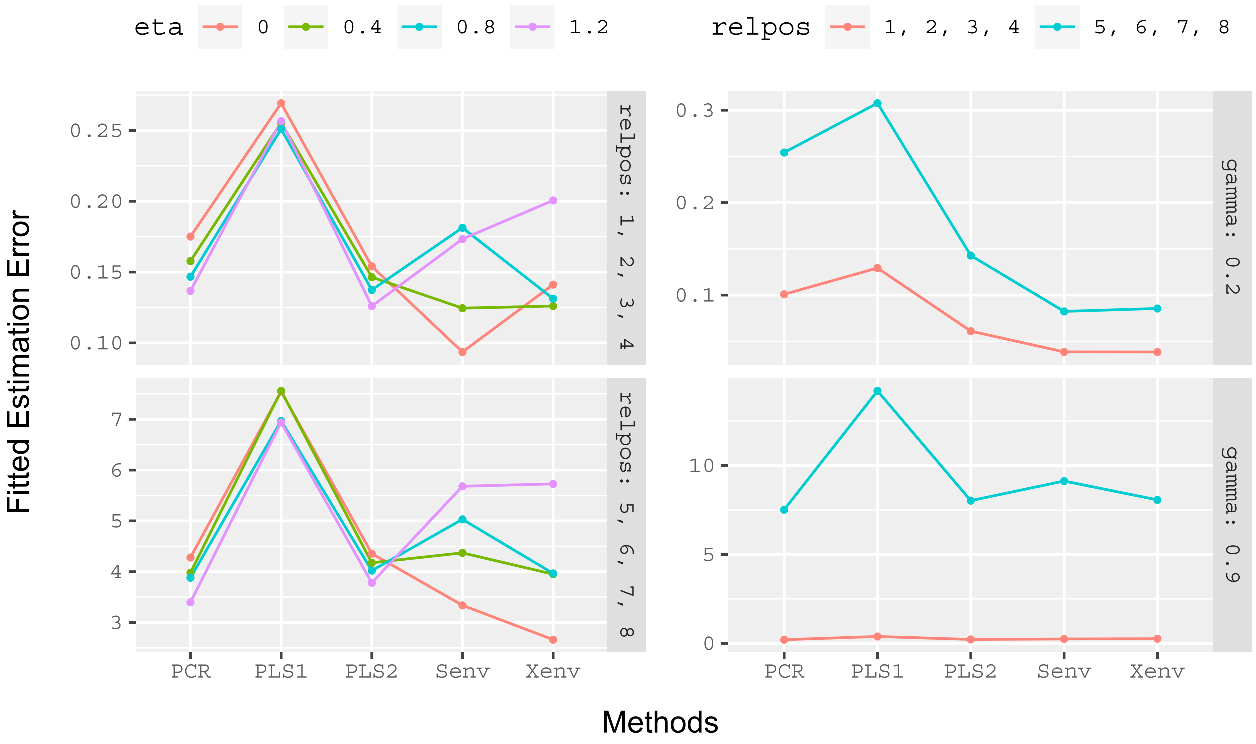


Fig. 9. Effect plot of some interactions of the MANOVA corresponding to fitted *error model*.

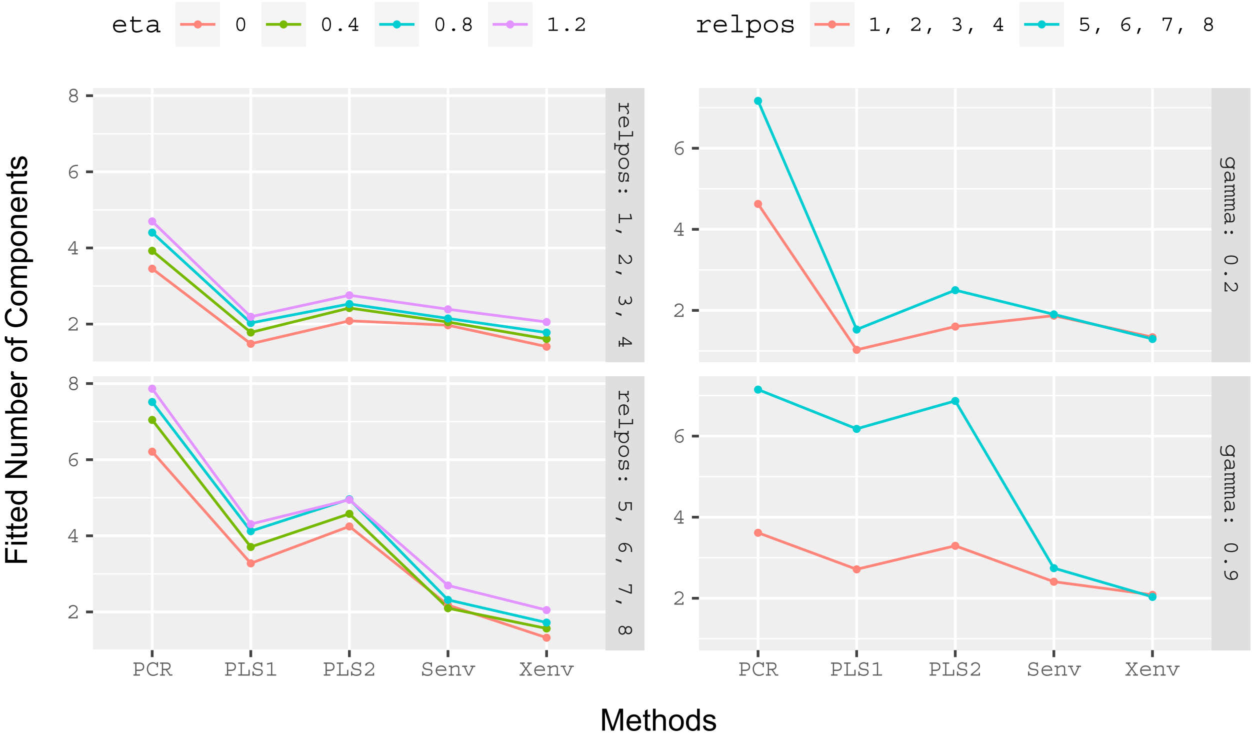


Fig. 10. Effect plots of some interactions of the multivariate linear model corresponding to the *component model*.

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. Since the envelope methods are likelihood based and are asymptotically efficient, with sufficiently large number of samples, the methods can produce smaller prediction and estimation errors than others usual MLE methods.

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 Rimal et al. [22]; has been useful so that we were able to model the data using envelope methods without losing too-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, which models the responses separately, is in general performing poorer than other methods. Unlike in prediction comparison, the performance of the envelope methods is comparable to the others except for the use of the number of components to obtain the minimum estimation error. The envelope methods have used 1–2 components in almost all designs, which is quite impressive. Although non-optimal number of components can lead to large estimation error and so one should be careful in this respect however this can easily be controlled through a method like cross-validation. Both PLS1 and PLS2 use a smaller number of components when the relevant components are at positions 1, 2, 3, 4. However, both methods used 7–8 components for the designs with relevant components at positions 5, 6, 7, 8.

We expect the results from this study may help researchers, working on theory, application and modelling, to understand these methods and their performance on data with varying properties.

The first part of this study [22] on prediction comparison should be considered to obtain a comprehensive view of this comparison. A shiny [2] 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.

CRediT authorship contribution statement

Raju Rimal: Methodology, Software, Validation, Formal analysis, Data curation, Writing - original draft, Writing - review & editing, Visualization. **Trygve Almøy:** Conceptualization, Methodology, Writing - review & editing, Supervision. **Solve Sæbø:** Conceptualization, Methodology, Writing - review & editing, Supervision, Resources.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chemolab.2020.104093.

References

[1] T. Almøy, A simulation study on comparison of prediction methods when only a few components are relevant, Comput. Stat. Data Anal. 21 (1996) 87–107, https://doi.org/10.1016/0167-9473(95)00006-2.

[2] W. Chang, J. Cheng, J. Allaire, Y. Xie, J. McPherson, Shiny: web application framework for R, URL, https://CRAN.R-project.org/package=shiny. r package version 1.2.0, 2018.

[3] R.D. Cook, An Introduction to Envelopes : Dimension Reduction for Efficient Estimation in Multivariate Statistics, 1 ed., John Wiley & Sons, Hoboken, NJ, 2018, 2018.

[4] R.D. Cook, I.S. Helland, Z. Su, Envelopes and partial least squares regression, J. Roy. Stat. Soc. B Stat. Methodol. 75 (2013) 851–877, https://doi.org/10.1111/rssb.12018.

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

[6] R.D. Cook, B. Li, F. Chiaromonte, Envelope models for parsimonious and efficient multivariate linear regression, Stat. Sin. 20 (2010) 927–1010.

[7] R.D. Cook, X. Zhang, Simultaneous envelopes for multivariate linear regression, Technometrics 57 (2015) 11–25, https://doi.org/10.1080/00401706.2013.872700.

[8] R.D. Cook, X. Zhang, Algorithms for envelope estimation, J. Comput. Graph Stat. 25 (2016) 284–300, https://doi.org/10.1080/10618600.2015.1029577, arXiv: 1403.4138.

[9] I.S. Helland, Partial least squares regression and statistical models, Scand. J. Stat. 17 (1990) 97–114, https://doi.org/10.2307/4616159.

[10] I.S. Helland, Model reduction for prediction in regression models, Scand. J. Stat. 27 (2000) 1–20, https://doi.org/10.1111/1467-9469.00174.

[11] I.S. Helland, T. Almøy, Comparison of prediction methods when only a few components are relevant, J. Am. Stat. Assoc. 89 (1994) 583–591, https://doi.org/10.1080/01621459.1994.10476783.

[12] I.S. Helland, S. Saebø, T. Almøy, R. Rimal, Model and estimators for partial least squares regression, J. Chemometr. 32 (2018), e3044, https://doi.org/10.1002/cem.3044.

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

[14] I.T. Jolliffe, Principal Component Analysis, second ed., 2002, https://doi.org/10.2307/1270093 arXiv:arXiv:1011.1669v3.

[15] S. de Jong, SIMPLS: an alternative approach to partial least squares regression, Chemometr. Intell. Lab. Syst. 18 (1993) 251–263, https://doi.org/10.1016/0169-7439(93)85002-X.

[16] F. Lindgren, P. Geladi, S. Wold, The kernel algorithm for pls, J. Chemometr. 7 (1993) 45–59, https://doi.org/10.1002/cem.1180070104, doi: doi:10.1002/cem.1180070104.

[17] B.H. Mevik, R. Wehrens, Theplspackage: principal component and partial least squares regression inr, J. Stat. Software 18 (2007), https://doi.org/10.18637/jss.v018.i02 nil. URL doi: doi:10.18637/jss.v018.i02.

[18] T. Næs, I.S. Helland, Relevant components in regression, Scand. J. Stat. 20 (1993) 239–250.

[19] T. Næs, H. Martens, Comparison of prediction methods for multicollinear data, Commun. Stat. Simulat. Comput. 14 (1985) 545–576, https://doi.org/10.1080/03610918508812458.

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

[21] R. Rimal, T. Almøy, S. Sæbø, A tool for simulating multi-response linear model data, Chemometr. Intell. Lab. Syst. 176 (2018) 1–10, https://doi.org/10.1016/j.chemolab.2018.02.009.

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

[23] S. Sæbø, T. Almøy, I.S. Helland, Simrel - a versatile tool for linear model data simulation based on the concept of a relevant subspace and relevant predictors, Chemometr. Intell. Lab. Syst. 146 (2015) 128–135, https://doi.org/10.1016/j.chemolab.2015.05.012.

[24] H. Wold, Soft modelling by latent variables: the non-linear iterative partial least squares (nipals) approach, J. Appl. Probab. 12 (1975) 117–142, https://doi.org/10.1017/s0021900200047604. URL.