EXPLORATION OF MULTI-RESPONSE MULTIVARIATE METHODS

Utforskning av multi-respons multivariate metoder

DOCTOR OF PHILOSOPHY (PHD) THESIS

RAJU RIMAL

Biostatistics
Dept. of Chemistry, Biotechnology and Food Science
Norwegian University of Life Sciences

Ås, 2019

Norwegian University of Life Sciences

Thesis Number: 2019:76
ISSN: 1894-6402
ISBN: 978-82-575-1636-9

The goal is to turn data into information, and information into insight.
– CARLY FIORINA, FORMER CEO OF HEWLETT–PACKARD

Supervisors:
Professor *Solve Sæbø*
Prorector of Education
Norwegian University of Life Sciences
Ås, Norway

Associate Professor *Trygve Almøy*
Dept. of Chemistry, Biotechnology and Food Science
Norwegian University of Life Sciences
Ås, Norway

Exploration of Multi-Response Multivariate Methods
PHD THESIS, 2019, AUG

WEBSITE:
https://therimalaya.github.com/thesis
E-MAIL:
raju.rimal@nmbu.no

This thesis is prepared with `ArsClassica` LaTeX template with `pandoc` and r-package `bookdown`.

SUMMARY

A linear regression model defines a linear relationship between two or more random variables. The random variables that depend on other random variables are often called response variables and the independent random variables are called predictor variables. In most cases not all variation is relevant for regression, i.e. only a certain amount of the variation in the predictors is relevant and only so for a part of the variation in the response. This leads to a reduction of the linear regression model where one can imagine a subspace of the space spanned by the predictor variables that contains all the relevant information for a subspace of the space spanned by the response variables.

In this thesis we attempt to compare some new methods which are based on the envelope model and some established methods such as principal components regression (PCR) and partial least squares regression (PLS). The comparison tests these methods on their performance of producing minimum prediction and estimation error while modelling data simulated with specifically designed properties. For the simulation we have also created an R-package called `simrel` with a web interface.

A simulation model for a multi-response multivariate linear model, on which the simulation tool is based, is discussed in the first paper. This paper prepares a basic foundation for the simulations with the concept of reduction of regression models. The second paper discusses the similarities of the envelope, PCR and PLS population models. This paper compares the prediction performance of several multivariate methods using a model with a single response.

The final two papers make an extensive investigation evaluating the prediction and estimation performance of established (PCR, PLS1 and PLS2) and newly developed envelope based (Xenv and Senv) methods. Unsurprisingly the study found that not one method dominates in all situations, but their performance depend on the properties of the data they model. However, the envelope based methods have shown remarkable performance in many cases, both in prediction and estimation. The study also recommend researchers to use and evaluate the envelope methods.

SAMMENDRAG

En lineær regresjonsmodell definerer et lineært forhold mellom to eller flere tilfeldige variabler. De tilfeldige variablene som er avhengige av andre tilfeldige variabler, kalles ofte responsvariabler, og de uavhengige tilfeldige variablene kalles prediktorvariabler. I de fleste tilfeller er ikke all variasjon relevant for regresjon, dvs. bare en viss mengde variasjonen i prediktorene er relevante, og bare for en del av variasjonen i responsen. Dette fører til en reduksjon av den lineære regresjonsmodellen der man kan forestille seg et underrom av rommet som spennesut av prediktorvariablene som inneholder all relevant informasjon for et underrom av rommet spent ut av responsvariablene.

I denne avhandlingenprøver vi å sammenligne noen nye metoder som er basert på Envelopemodellen og noen etablerte metoder som principal komponent regresjon (PCR) og partiell minste kvadraters regresjon (PLS). Sammenligningen tester disse metodene på deres ytelse til å produsere minimum prediksjon- og estimeringsfeil, mens modelleringsdata simuleres med spesielt designede egenskaper. For simuleringen har vi også laget en R-pakke kalt `simrel` med et webgrensesnitt.

En simuleringsmodell for multirespons, multivariat lineær modell, som simuleringsverktøyet bygger på, diskuteres i den første artikkelen. Denne artikkelen utarbeider et grunnleggende fundament for simuleringene basert på konseptet om reduksjon av regresjonsmodeller. Den andre artikkelen diskuterer likhetene i Envelope-, PCR- og PLS-populasjonsmodellene. Denne artikkelen sammenligner prediksjonsytelsen til flere multivariate metoder ved bruk av en modell med en enkelt respons.

De to siste artiklene gir en grundig evaluering av prediksjons- og estimeringsegenskapene til etablerte metoder (PCR, PLS1 og PLS2) og nyutviklede envelope-baserte metoder (Xenv og Senv). Ikke uventet fant studien at det ikke finnes en enkelt metode som dominerer i alle situasjoner, men resultatene deres avhenger av egenskapene til dataene de modellerer. Imidlertid har envelope-baserte metoder vist bemerkelsesverdig resultater i mange tilfeller, både når det gjelder prediksjon og estimering. Studien anbefaler også forskere å bruke og evaluere envelope-metodene.

ACKNOWLEDGMENT

First and foremost, I am indebted to my supervisor Solve Sæbø who picked me up from nowhere and brought me into a scientific community by giving me a chance to pursue this degree. His inspiration and encouragement have been an essential element in the course of this journey. I am grateful to my co-supervisor Trygve Almøy for being a mentor, a friend, a colleague, and a guardian and guiding and supporting me throughout this period. He has always been there for me with my frustration and excitement.

I am forever grateful to my father Narayan Prasad Rimal and mother Bhagawati Rimal for their continuous support and encouragement. Their belief in me and push for my education have shined the light in my hard and easy times. I am also thankful to my dear wife Junali Chhetri who has inspired me every step of my life and help me to better understand myself. And of course, a thank goes to my beloved son Nirvan Rimal who has understood my busy time during this study.

I would also like to thank Professor Inge Helland for his insights, suggestions and comments on many mathematical problems on various statistical methods presented in the thesis.

Last, but importantly, my thank goes to the Biostatistics group with whom I have collected beautiful memories. Thanks to all the members of the group from past and present who have always made my stay at NMBU happy, festive and full of joy.

PREFACE

This thesis is a part of Doctor of Philosophy (PhD) study. The first part of the thesis constitute a gentle introduction to the objective of the study and some of its background. This is followed by the summary of individual research paper on which this thesis is based on. The discussion section tries to bind the finding from theses papers. The final chapter will discuss the limitations and future prospect of the study. The second part contains all the papers attached.

An R-package called `simrel` is available as part of the first paper included in this thesis. The package lets users simulate data from a multi-response linear model. The package can be installed from R-package repository CRAN or from GitHub. In addition, a web application that gives users a graphical user interface for the package is also available from GitHub. All the results and the documentations of the research can be reproduced from the codes in GitHub repository with software and packages required installed. In addition, one can use docker image together with the code for reproducing the thesis together with all included papers. All related resources are listed in the final chapter.

CONTENTS

LIST OF RESEARCH PAPERS

LIST OF FIGURES

INTRODUCTION

Rapid development in technology and computational power have resulted in heaps of data. Extracting information from this chaotic heaps of data has become another problem. Many statistical and machine learning tools are devised for this purpose, most of which focus to identify the relationships between different variables. A linear relationship is the most common assumption. This thesis confined itself in the exploration of linear relationships, where a set of independent variables, called predictor variables, affect another set of dependent variables, called response variables. The space spanned by the columns of predictors and responses are termed *predictor space* and *response space*, respectively.

Many projection-based statistical methods such as Principal Components Regression (PCR), Partial Least Squares (PLS) Regression and some variants of Envelopes only consider a subspace of predictor space relevant for defining the linear relationship between the predictors and the response(s). This brings us to the concept of relevant and irrelevant space introduced by Naes and Martens [1985]. The relevant space can be described as the subspace that contains all the required information to define the relationship between the predictors and the response in a model. The irrelevant space, on the other hand, does not contain any information regarding this relationship.

Latent components corresponding to predictor variables, which we will refer to as "predictor components", are linear combinations of the predictor variables. Naes and Martens [1985] and later Helland [1990], Næs and Helland [1993] and Helland and Almøy [1994] have defined a set of predictor components as irrelevant components if they have no correlation with the response variables and the relevant part. Using only a subset of the latent components for modeling, is often termed as "dimension reduction". Methods like PCR, PLS and many other variants of PLS has leveraged this concept and are serving as prime tools in many disciplines, most notably in chemometrics.

Relatively new methods based on the concept of "envelopes" introduced by Cook et al. [2007], more specifically envelope in predictor variable

(Xenv), have also used this concept of dimension reduction. In addition, envelope in response variable (Yenv) and simultaneous envelope in predictor and response (Senv) have extended the concept of relevant and irrelevant space to the response space as well, which they referred to as material and immaterial part. These methods are discussed in Background section.

Despite having similar underlying population model, these methods estimate the model parameters differently. Model parameters are the unknowns, which help to define a complex relationships between the variables. Regression coefficient vector (β) in (2) is an example of a model parameter. All methods use data to estimate these parameters. So, the properties of a dataset affect the estimation and consequently the prediction performance of the methods. Evaluation of these methods is essential to understand how they interact with various properties of the data. This thesis will explore some of these methods and assess their estimative and predictive strength and weaknesses through both simulated and real datasets.

This exploration adds a reference for researchers to motivate them for using different methods based on the properties of the data they are working on. This study is exploratory in nature where we assess and compare different multi-response multivariate methods, but most importantly study their interaction with the properties of the data. The properties include the correlation between predictor variables, the position of principal components of predictor variables (predictor components) that are relevant for certain principal components of the response variables (response components), the amount of correlation between the response variables and the number of predictor variables. The effect of the correlation structure of the response matrix is less explored and it is expected to shed some light on how similar and how different the methods are in terms of modelling this structure. In order to simulate data with these properties varying at different levels, we have created an R-package called `simrel`, which is an extension of the previous version introduced by Sæbø et al. [2015] to incorporate multiple responses.

BACKGROUND

This section discusses the relevant topics that have been used in the included papers.

Multivariate Linear Regression Model

The joint normal distribution of a random variable-vector $\mathbf{y}$ of m response variables with mean of $\boldsymbol{\mu}_y$ and another random variable-vector $\mathbf{x}$ of p predictor variables with mean $\boldsymbol{\mu}_x$ as,

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

where $\boldsymbol{\Sigma}_{xx}$ and $\boldsymbol{\Sigma}_{yy}$ are the variance-covariance matrices of $\mathbf{x}$ and $\mathbf{y}$, respectively, and $\boldsymbol{\Sigma}_{xy} = \boldsymbol{\Sigma}_{yx}^t$ is the covariances between them.

A model that linearly relates $\mathbf{x}$ and $\mathbf{y}$ through regression coefficient vector $\boldsymbol{\beta}$ is often written as,

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

where $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{y|x})$

We can write the regression coefficient $\boldsymbol{\beta} = \boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}$ in terms of the covariance matrices. A complete simulation of this model requires to specify $1/2(p+m)(p+m+1)$ unknowns.

With a transformation defined as $\mathbf{z} = \mathbf{R}\mathbf{x}$ and $\mathbf{w} = \mathbf{Q}\mathbf{y}$ with $\mathbf{R}_{p\times p}$ and $\mathbf{Q}_{m\times m}$ as random orthogonal rotation matrices, model (1) can be rewritten as,

$$\begin{aligned}\begin{bmatrix}\mathbf{w}\\ \mathbf{z}\end{bmatrix} &\sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = N\left(\begin{bmatrix}\boldsymbol{\mu}_w\\ \boldsymbol{\mu}_z\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{ww} & \boldsymbol{\Sigma}_{wz}\\ \boldsymbol{\Sigma}_{zw} & \boldsymbol{\Sigma}_{zz}\end{bmatrix}\right)\\ &= N\left(\begin{bmatrix}\mathbf{Q}\boldsymbol{\mu}_y\\ \mathbf{R}\boldsymbol{\mu}_x\end{bmatrix}, \begin{bmatrix}\mathbf{Q}\boldsymbol{\Sigma}_{yy}\mathbf{Q}^t & \mathbf{Q}\boldsymbol{\Sigma}_{yx}\mathbf{R}^t\\ \mathbf{R}\boldsymbol{\Sigma}_{xy}\mathbf{Q}^t & \mathbf{R}\boldsymbol{\Sigma}_{xx}\mathbf{R}^t\end{bmatrix}\right)\end{aligned} \tag{3}$$

Since both $\mathbf{Q}$ and $\mathbf{R}$ are orthonormal matrices, i.e., $\mathbf{Q}^t\mathbf{Q} = \mathbf{I}_m$ and $\mathbf{R}^t\mathbf{R} = \mathbf{I}_p$, the inverse transformation can be defined as,

$$\begin{aligned} \boldsymbol{\Sigma}_{yy} &= \mathbf{Q}^t\boldsymbol{\Sigma}_{ww}\mathbf{Q} \quad \boldsymbol{\Sigma}_{yx} = \mathbf{Q}^t\boldsymbol{\Sigma}_{wz}\mathbf{R} \\ \boldsymbol{\Sigma}_{xy} &= \mathbf{R}^t\boldsymbol{\Sigma}_{zw}\mathbf{Q} \quad \boldsymbol{\Sigma}_{xx} = \mathbf{R}^t\boldsymbol{\Sigma}_{zz}\mathbf{R} \end{aligned} \tag{4}$$

Here, $\boldsymbol{\Sigma}_{zz}$ and $\boldsymbol{\Sigma}_{ww}$ are diagonal matrices of eigenvalues corresponding to predictors and responses respectively. Following the concept of relevant components $\boldsymbol{\Sigma}_{wz} = \boldsymbol{\Sigma}_{zw}^t$ has non-zero elements for relevant components. With some random orthogonal rotation matrices $\mathbf{R}$ and $\mathbf{Q}$, which can be easily generated, the unknowns required for simulation may drastically decrease. Following the idea from Sæbø et al. [2015], Paper I uses exponential decay of eigenvalues, as in (5), that fills the diagonals of $\boldsymbol{\Sigma}_{zz}$ and $\boldsymbol{\Sigma}_{ww}$. Here the decay factor γ controls the multicollinearity such that a higher value of gamma corresponds to high multicollinearity.

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

A thorough discussion on the reparameterization of a linear model to simulate data by the concept of "relevant components" can be found in Paper I. The following subsection discusses the concept of relevant components in brief.

Relevant Space and Relevant Components

In the model (1), not all information in $\mathbf{x}$ is relevant for $\mathbf{y}$ and not all variation in $\mathbf{y}$ is explainable or non-redundant. We can refer to the space "with information" as relevant (informative) space and the rest as irrelevant (uninformative) space. Naes and Martens [1985] introduced the definition of relevant space as the decomposition of the predictor space into two orthogonal subspaces: the relevant and the irrelevant space. Additionally, a set of predictor components defined as irrelevant components do not have any correlation with the response and the relevant part of the data. The relevant components, on the other hand, contains all the required information to explain the variation in the response $\mathbf{y}$. Multivariate methods such as Principal Components Regression (PCR) and Partial Least Squares (PLS) Regression uses the eigenvectors to span the relevant and irrelevant spaces. Here, we refer the eigenvectors that span the relevant

space as *relevant eigenvectors*. The concept was further discussed and developed by Helland [1990], Næs and Helland [1993] and Helland and Almøy [1994]. However, all these studies have discussed the separation of relevant and irrelevant space only in the predictor space.

More recently, various estimators [Cook et al., 2010, 2013, Cook and Zhang, 2015b] based on a so-called "envelope" [Cook et al., 2007] have used and extended the concepts of the separation of relevant and irrelevant spaces to the response space as well. The relevant and irrelevant spaces are referred to as material and immaterial spaces in their literature (Figure 1). The envelope methods use "envelope", a linear combination of relevant eigenvectors [Cook, 2018], to span the relevant space.

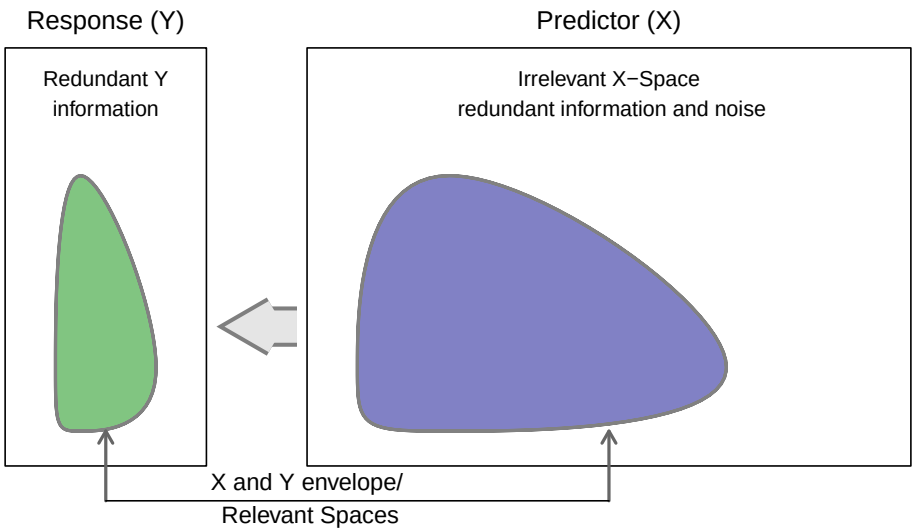


Figure 1: A heuristic illustration of relevant and irrelevant spaces in a response space and a predictor space

To elaborate on the concept of relevant components and how they interact with other properties and influence the prediction of methods, let us consider an example. Assume a single response model with 10 predictor variables where the information contained in these 10 predictors can be completely explained by four principal components of Σ_{xx}, the variance-

covariance matrix of the predictor ($\mathbf{x}$). These four components are the relevant components. Consider two cases:

CASE 1 (FIGURE 2, LEFT): The position of these relevant components are 1, 2, 3 and 4. The eigenvalues of Σ_{xx} decay slowly, i.e. low multicollinearity. Here, the relevant components from 1 to 4 have large variation, so that, most methods easily extract the information and fit the model quite accurately.

CASE 2 (FIGURE 2, RIGHT): The position of the relevant components are at 5, 6, 7 and 8. The eigenvalues of Σ_{xx} decay rapidly, i.e. high multicollinearity. Here the relevant components from 5 to 8 have small variation, so that, it is difficult for most methods to extract the information and fit the model.

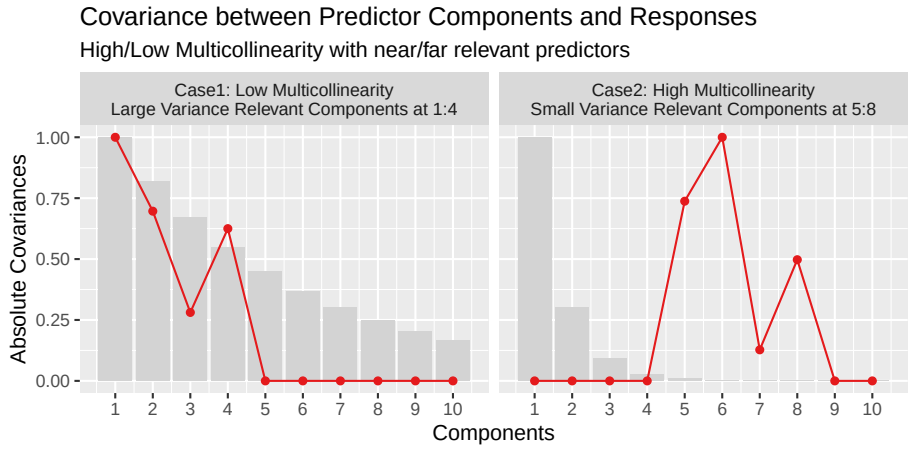


Figure 2: Relevant components at two different set of positions and two different levels of multicollinearity. The points represents the correlation of predictor components and the response variable. The grey bars are the eigenvalues of $\mathbf{\Sigma}_{xx}$.

Further, PCR and PLS regression are used with the data simulated from these two cases. Also, leave-one-out cross-validation validates their prediction performance, and the root mean squares error of prediction measures their prediction error (Figure 3).

Different methods target these cases differently. For example, PCR tries to capture maximum variation in $\mathbf{x}$ through principal components, so it

starts reducing its prediction error only after including the relevant components. For this method, in the first case, prediction error starts decreasing from the first component on, and stabilize after the fourth component while in the second case, prediction error only starts decreasing after the fifth component. This method requires all four relevant components to get the minimum prediction error. Partial Least Square Regression (PLS), on the other hand, is motivated to maximize the covariance between the predictors and the response. We can see a significant decline of prediction error after the first relevant components is included but it uses fewer components to get the minimum prediction error than PCR in both cases. Helland and Almøy [1994] has shown a similar result and shown that the relevant components with small variation make the prediction difficult.

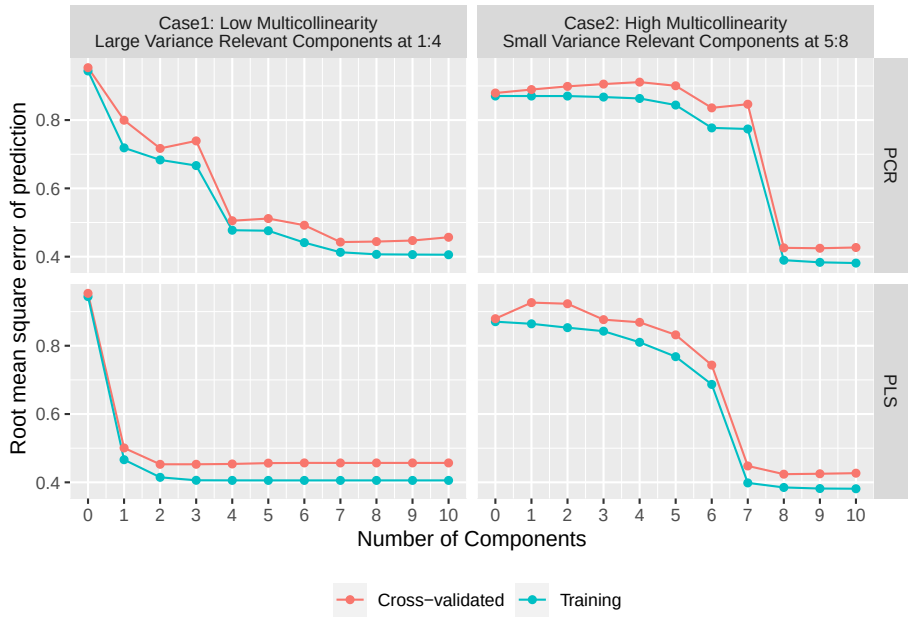


Figure 3: Root mean square error of cross-validation from PCR and PLSR

The concept of relevant components can also be extended to the response such that a subspace contains the information relevant for a model. The concept is implemented in the simultaneous envelope [Cook and Zhang, 2015b] and the response envelope [Cook et al., 2010] methods.

Simulation

Random variables are the basic components of a complex model and a stochastic simulation. These random variables can be generated on a computer by sampling and manipulating uniform random variables $U(0,1)$ which requires random numbers. Although computers can not generate truly random numbers, it can, however, generate pseudo-random numbers. These numbers appear as random numbers but they are completely deterministic. Since they are deterministic, any experiment performed using these numbers can be repeated exactly [Jones et al., 2014]. We can use these uniform random variables to create other random variables that follow a certain distribution. Standard Normal Distribution is a common one and is used in many statistical simulations including the tool discussed in paper I. Given that we can simulate a standard normal variable $\mathbf{z}$, one can obtain any normal distribution with arbitrary mean μ and variance σ^2 as $\mu + \sigma Z$. Here, we can control the parameters μ and σ.

Simulation refers to generating data from a known underlying population structure. Controlling the properties of the population is vital in the simulation. This enables researchers and users to use data for comparison of methods, assessing new methodology, testing theory and evaluating algorithms. Such data can also be used for educational purposes.

All the research studies in this thesis have used an R-package called `simrel` for simulating multi-response linear model data (paper I). The simulation tool is general purpose in nature and has a limited number of parameters that controls the essential properties of the population. It is flexible and enables users to simulate data with a wide range of properties. Some of these properties include the level of correlation between the predictors (`gamma`) and responses (`eta`) through exponential decay factor as in (5). The position of the relevant components (`relpos`), the number of predictor variables (`p`) and the number of response variables (`m`) can also be controlled during the simulation.

Estimation and Prediction

Measures such as mean and standard deviation for a population are usually referred to as parameters of the population. A model as in (2), which expresses the relationship between $\mathbf{x}$ and $\mathbf{y}$ in the population, uses param-

eters such as the error variance and regression coefficients. Usually, due to the lack of known population distribution, the values of these parameters are calculated using a sample collected from the population. The process of determining the value of certain parameters is called estimation. The estimated parameter values from any two samples are different. A method for estimation is considered better if the expected squared difference between the estimated and true value is small and has small variance. The goodness of estimation method depends on the nature of the data. Estimation error with true and estimated regression coefficient β and $\widehat{\beta}$ respectively, can be defined as in (6).

$$\text{Estimation Error} = E\left[\left(\beta - \hat{\beta}\right)^t \left(\beta - \hat{\beta}\right)\right] \tag{6}$$

A fitted or trained model is mostly used for prediction. Prediction refers to determining the value of the response for a new set of predictors, which were not used to train the model. Most studies under "data science" field are targeted for better prediction. Most comparisons in this thesis evaluate the prediction performance of the multivariate methods using the prediction error measured as in (7).

$$\text{Prediction Error} = E\left[\left(\beta - \hat{\beta}\right)^t \Sigma_{xx} \left(\beta - \hat{\beta}\right)\right] + \Sigma_{y|x} \tag{7}$$

From (6) and (7), we can see that the prediction errors are influenced by the covariance of the predictors directly, while estimation error is not. In the case of multicollinear predictors, estimation error can be huge, while due to the scaling of the covariation of predictors, the prediction error can still be small. A good estimation can give a proper and trustworthy idea about the relation between certain predictor variation with a certain response variable. This is important in policymaking, academic researches and to understand the relationships when developing new models. Prediction, on the other hand, is widely used from weather forecasting, economic forecasting, prediction in production and sales, and many more.

Multivariate Methods

Various multivariate methods such as ordinary least squares (OLS), principal components regression (PCR), partial least squares (PLS) regression and envelope methods are used for comparative studies included in this

thesis. All of these methods except OLS use the concept of relevant space and the reduction of the regression model. Here we will refer PLS2, which models all the response variables together, as PLS and PLS1, which models each responses separately, as PLS1.

Methods based on Envelope Model

Three different methods based on envelopes are also included for comparison. Cook et al. [2007] defined envelope as the smallest subspace that includes the span of true regression coefficients and developed various estimators based on the concept of the envelope through various subsequent papers. Response envelope (Yenv) [Cook et al., 2010] performs dimension reduction only in the response space while Predictor envelope (Xenv) [Cook et al., 2013] performs dimension reduction only in the predictor space. The simultaneous envelope (Senv) [Cook and Zhang, 2015b] performs dimension reduction on both predictor and response space simultaneously. If all the possible components (latent dimension) are included in these methods, the results are equivalent to OLS regression. The comparisons of these envelope methods together with PCR and PLS in the third and fourth paper have shown encouraging results for envelope methods in both easy and difficult cases.

PLS and its derivatives

Since the PLS method has been both popular and productive in fields like chemometrics, its development has progressed quickly over time through the formulation of various derivatives. CPLS and CPPLS are among them which combines PLS and canonical correlation analysis (CCA) and give a joint framework for classification and regression [Indahl et al., 2009]. Paper-I has made some basic comparison of these methods for their predictive ability. More recently, Helland et al. [2012]] introduced the Bayes PLS method. The method only works with a single response model and has shown promising results compared to other methods in Paper-II.

Wentzell and Montoto [2003] has assembled many comparisons made on PCR and PLS where they conclude that PLS has not shown a clear advantage over PCR over predictive ability in most studies, but uses fewer components than PCR. Many studies are available comparing PCR, PLS

and their derivatives. However, there are not any studies to date which have made any empirical comparisons of the newly developed *envelope* based methods using real and simulated data with these more established methods.

Details on each of these methods can be obtained from the corresponding references.

Experimental Design

In all the post hoc comparisons, simulation parameters are considered as independent variables (factors), and the prediction- and estimation errors are considered as outcome variables (responses). Factorial Design is implemented as an experimental design which allowed us to compare all possible combination of different factor levels. For example, the factorial design used throughout the third and fourth paper, shown in Figure 4, has four factors: a) Number of predictor variables (p) with two levels, b) level of multicollinearity (gamma) with two levels, where higher value represents a higher level of multicollinearity, c) position index of relevant predictor components (relpos) and d) the level of collinearity in response (eta), with four levels where higher value represents a higher correlation between the response variables. The combination of these factors has created 32 unique designs which are then used for simulating data with those particular properties. Such data, with all possible combination of these properties, have made both thorough and rigorous comparison possible.

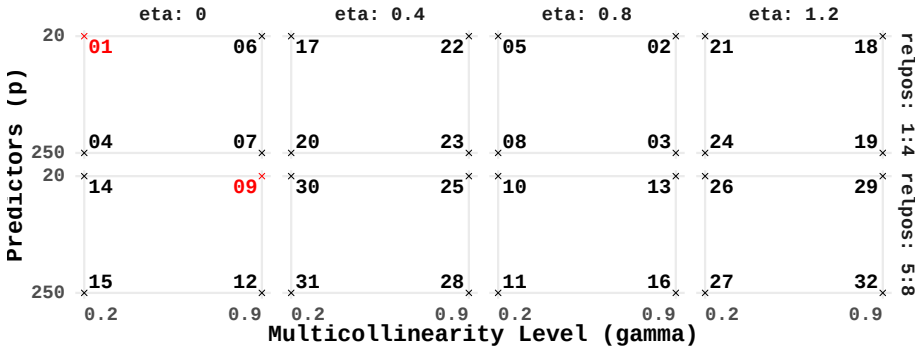


Figure 4: An example of a factorial design used in the third and fourth paper.

Let us dig a little deeper to understand how these simulation parameters are tied with the properties of the simulated data. As an example, let us take Design 1 and Design 9 of Figure 4 where data simulated with Design 1 have low multicollinearity and the position index of relevant components are 1, 2, 3, 4, while Design 9 have high multicollinearity and the position index of relevant components are 5, 6, 7, 8. With other factors or properties of the data being the same for both, the difference in these two designs help us to analyse the interaction between the multicollinearity in the data and the position of relevant components on, for instance, prediction performance of the methods.

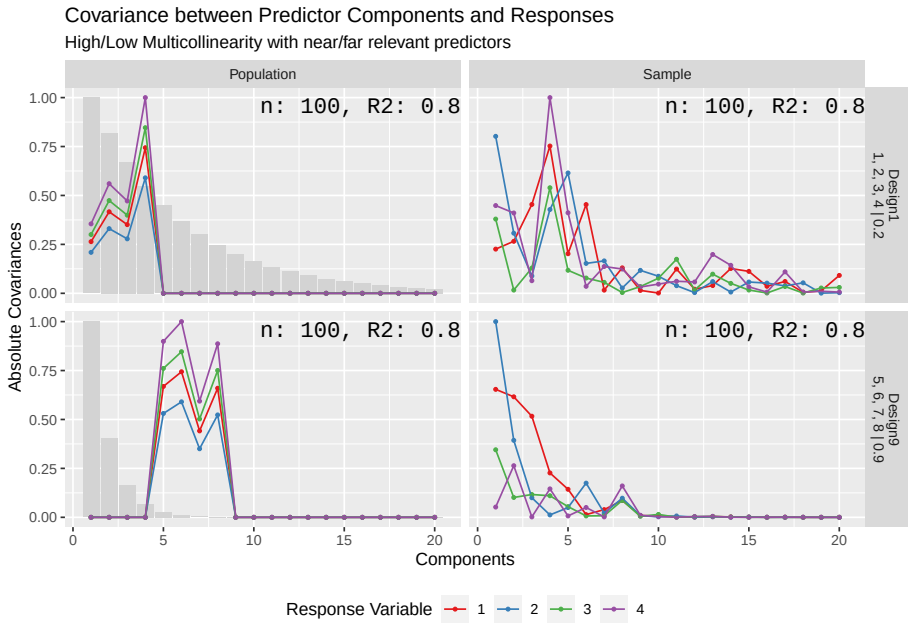


Figure 5: *Design 1*: Relevant components have large variation, *Design 9*: irrelevant components have large variation and relevant components have small variation.

Figure 5 (top-row) shows the scaled covariance between the predictor components and the response variables for Design 1. Here the relevant components with larger variation (due to low multicollinearity) simulate data that are easier to model by most methods. Figure 5 (bottom-row) for Design 9 shows that the relevant components at position 5, 6, 7, 8 have small variation and irrelevant components at position 1, 2, 3, 4 have large variation. This design simulates data that are difficult to model by

most methods. The population covariances in the figure give clear and distinct relationship, while the sample covariances give a somewhat rough approximation of the population.

Analysis of Variance

The analysis in these studies have used various exploratory plots of prediction error, estimation error and the number of components used by different methods. Also, visualizations from principal components analysis (PCA) have been used on these errors. Besides, a more formal analysis is made using analysis of variance (ANOVA). ANOVA allowed us not only to understand the effect of various properties of data controlled by the simulation parameters but also analyses the effect of the interaction of these properties with the methods. The third and fourth paper use multivariate analysis of variance (MANOVA) to analyze the effect on four response variables.

MANOVA is the multivariate counterpart of the ANOVA where various test statistic are used, such as Wilks' Lambda, Lawley-Hotelling trace, Pillai trace and Roy's largest root. All of these methods use the within ($\mathbf{E}$) and between ($\mathbf{H}$) sum of squares and the cross products matrices. All four test statistic are nearly equivalent for large sample size [Johnson and Wichern, 2018]. In our studies, Pillai trace is used, which is defined as,

$$\text{Pillai statistic} = \text{tr}\left[(E+H)^{-1}H\right] = \sum_{i=1}^{m} \frac{\nu_i}{1+\nu_i} \tag{8}$$

where, ν_i represents the eigenvalues corresponding to $\mathbf{E^{-1}H}$.

PAPER SUMMARY

Paper 1: A Tool for Simulating Multi Response Linear Model Data

As an extension of Sæbø et al. [2015] to simulate linear model data with multiple response variables, this paper discusses the simulation model, the strategy for simulation, and compares some multivariate methods using simulated data. Additionally, it includes an R-package called `simrel` that is built based on the mathematical formulation discussed in the paper.

The simulation of the linear model discussed here is based on the concept of the relevant components. A subspace of the predictor space, which is relevant for a subspace of response space, is the basis of the simulation tool. These subspaces are assumed to be spanned by a subset of respective latent components. The simulation strategy started with identifying the covariance between these components that satisfy the user's condition for the data, i.e. the simulation parameters. A covariance structure of the latent space is then created which is rotated by an arbitrary orthogonal rotation matrix to obtain the population covariance structure of the simulated data. Data is then sampled from a normal distribution with the constructed covariance structure. The tool also provides mathematically computed properties of the data such as true regression coefficient, minimum model error, coefficient of determination and the predictor variables relevant for a given response.

In addition to the mathematical formulation for simulation, the study compares some multivariate methods including OLS, PCR, PLS and Envelope using two simulation examples. It has included some derivatives of PLS such as PLS1, PLS2, CPLS and CPPLS and some methods based on envelope estimation such as Xenv, Yenv and Senv. The first example has three relevant response components rotated into five response variables. Additionally, four simulation designs were constructed using factorial design with low and high multicollinearity interacting with low and high noise levels. The simultaneous envelope (Senv) method has achieved the smallest prediction error with a smaller number of components in the dataset

with low noise level (high coefficient of determination), while canonical PLS (CPLS and CPPLS) have shown better performance in the dataset with a higher level of noise. All the methods are found robust for the multicollinearity problem. The second example compares PLS1 and PLS2 where, on most occasions, the latter dominates the earlier with regard to minimum prediction error. Further, the paper has also introduced the shiny [Chang et al., 2018] web application designed for easier access to the simulation tool.

Paper 2: Model and Estimators for PLS Regression

Comparison of methods requires us to understand the modelling approach of the corresponding methods. This paper formulates five different ways to present a PLS model [Helland, 1990] and shows how they are equivalent. Additionally, it argues that the concept of relevant components for reduction of the regression model is the simplest way for it. My contribution to the paper was to compare the performance of PCR, PLS, Bayes PLS and Envelope (Xenv) methods using both simulated and real data.

The comparison was based on simulated data with 32 unique properties through a factorial design of simulation parameters. The parameters include medium and high levels of coefficient of determination, medium and high levels of multicollinearity, four different position index of relevant predictor components and two different n/p ratios, 0.3 and 0.8. The study is based on a single response model.

The study found some interesting results for the envelope and Bayes PLS methods. Since the Envelope (Xenv) method is based on maximum likelihood, the designs with n/p ratio equals to 0.8 destroyed its prediction while the method has fine prediction when the ratio was 0.3. Bayes PLS has shown remarkable prediction performance in most design, however, both methods had convergence problem in many situations.

Despite having the best performance, Bayes PLS has time-consuming computation and failed to converge for some cases. For practical purpose, the study recommends the ordinary PLS algorithm as a good option for prediction purpose.

Paper 3: Comparison of Multi Response Prediction Methods

Since prediction has been an essential component in data science, understanding how the prediction methods interact with different properties of data is important. This paper, together with the next, makes a comprehensive comparison using simulated data with specifically designed properties through various simulation parameters. The experimental design in Figure 4, discussed in the previous section, has been used in both of these comparisons. Besides, for the prediction comparison, two real data examples have also been used in the study. These two papers try to give an understanding of the interaction between methods and data properties in multi-response cases and also assess the performance of the envelope methods (Xenv and Senv) using both simulation and data from the field of chemometrics. Further, these studies not only use prediction and estimation error for assessment but also the number of components used to get the minimum error. Here only methods based on relevant space such as PCR, PLS (PLS1 and PLS2) and Envelopes (Xenv and Senv) are considered for comparison.

Since envelope methods are unable to fit a model with $p > n$, principal components of the predictor matrix were used to reduce the number of predictors. The number of components that explains the minimum of 97.5% of the variation in $\mathbf{x}$ are chosen. The regression coefficients were later transformed back using the respective eigenvectors. Since the envelope methods reduce the dimension as part of its fitting process, this detour in $p > n$ cases does not give them extra benefit which we have tested for $n > p$ cases using with and without principal components. This paper also illustrates the use of principal components for implementing envelope methods in data with wide $(p > n)$ predictor matrix which is common in fields like chemometrics and bioinformatics.

The minimum prediction error and the number of components to get that error are considered as observed responses in the study. The simulation parameters used in the experimental design are considered as factor variables for further statistical analysis. Multivariate Analysis of Variance (MANOVA) is used for proper statistical analysis with third-order interaction of these factors. The effect of different levels of the factors and their interactions are used for minute comparison.

Envelope methods in the study have produced a small prediction error using fewer components than other methods. The effect of correlation

between the response variables is small for all methods, however, envelope methods are more sensitive to this correlation. All methods are robust for handling multicollinearity, but PCR and PLS methods struggle more when the relevant predictor components have small variance and irrelevant components have a large variance.

Example with real data shows PCR and PLS have the smallest prediction error, but the number of components used by these is higher than the envelope methods. Envelope methods in these examples have obtained prediction error closer to the minimum obtained by PCR and PLS, but using a smaller number of components.

Paper 4: Comparison of Multi Response Estimation Methods

In many disciplines, the correct and stable estimation is just as an important primary objective as the prediction. This paper extends the analysis from Paper 3 to analyze the estimation aspect of the methods. The same experimental design and simulated data are used for this assessment as well.

The study found that overall performance highly depends on the nature of the data since simulation parameters, such as multicollinearity level and position of relevant predictors significantly interact with the methods. Since both envelope methods have smaller prediction and estimation error and have used fewer number of components, low multicollinearity with independent response variables are in favour of these methods. Higher correlations between the responses have given a larger estimation error for envelope methods. For these methods, choosing the wrong number of components can result in large estimation error, so the study also suggests using validation for estimation purpose. Both prediction and estimation error from PCR are more stable than other methods, while as PLS1 method models each response separately, the performance in general is poorer than other methods.

DISCUSSIONS & CONCLUSIONS

Simulated data are used in many scientific studies and teaching purposes. Assessing the properties of methods or algorithms is essential and usual in the scientific community. Since scientists often spend a lot of time developing a simulation model, paper-I attempts to present a simple, versatile and general-purpose tool for simulating such data only using few parameters. This attempt of adding a tool in scientists' toolbox aims at making the laborious work of researchers simpler and less time-consuming. Although not discussed much in the paper, the tool can also be useful for teaching purposes. Using the tool, educators can easily simulate data based on their context and need.

Most of our comparisons are on the methods that are based on the concept of relevant spaces. The study in paper-II helped us to understand the similarities and differences between these methods. My contribution to the second part of the paper was to use the simulation tool discussed in paper-I to compare these methods empirically. The Bayes PLS method has shown the best performance in these simulation results, and its performance on real data was satisfactory. This pointed us to explore the methods comprehensively. However, due to the time-consuming computation and as the Bayes PLS method has not yet been developed to work with multiple responses, we planned to use only the envelope methods, PCR and PLS for further exploration.

The further exploration continued on the multi-response setting for evaluating and comparing PCR, PLS and two envelope methods (Xenv and Senv) for their performance on prediction and estimation. These methods are capable of modelling multi-response models and are based on the concept of relevant space and dimension reduction.

Prediction and estimation both have many aspects to be discussed, we have divided the comparison study into two papers: Paper-III and Paper-IV. Since both papers use the same simulated data based on the same experimental design, it became easier to make comparisons of prediction and estimation for individual methods.

Since multicollinearity highly interacts with the position of the principal components, these factors highly influence both estimation and prediction. These factors were used as simulation parameters in addition to a factor that controls the correlation between the response variables. The study on the response correlation and its interaction with these methods and other simulation parameters are limited. This studies' attempts to fill up the gap have made this thesis novel and useful.

In the last two papers, Envelope methods have shown fine performance, specifically in the simulation examples. The PCR method has shown good performance if an optimal number of component is used. The performance is also stable, even with non-optimal number of components. Both PLS1 and PLS2 have stable and better performance, particularly when relevant components are at the initial position (i.e. with large variation). The fine performance of envelope methods is achieved using a smaller number of components, which shows its remarkable strength in dimension reduction. An optimal number of components is crucial for the Envelope methods than for the PCR and PLS methods, as the estimation error rapidly increases with an increasing number of non-optimal components.

In general, the study encourages researchers for using newly developed methods such as the envelope. This kind of comparisons in chemometrics data is relatively new for both chemometrics fields and the envelope methods. This thesis also hopes to be a useful reference for other researchers.

Since Envelope methods have dimension reduction in response, it can be useful when many responses can be explained by fewer response components. Not a single method is superior for all kinds of data, and using methods correctly requires identifying the properties of data. More sophisticated assessment and comparison can be possible through the tool `simrel`. Researchers are encouraged to leverage the tool for their study and experiments. We would like to request the developer of the envelope to reach different fields and spread the envelope in a more simple and less mathematical form of communication.

LIMITATIONS & FUTURE PERSPECTIVES

Although the studies in the thesis are all comparisons of methods, it is important to make those comparisons to evaluate the methods and to understand their interaction with various properties that can exist in real data. This provides an example assessment for method developers and gives a clear understanding of the methods under comparisons for these specific cases to other researchers.

The study mostly covers the comparisons through simulated data and some real data, but it also provides a direction for further exploration of these methods and other methods. Ridge, Lasso and other methods could have been used for comparison, but since they are not explicitly based on the concept of relevant components, we have discarded them from these comparisons at this point. Although we did some basic comparison by including them, they require a separate and a more comprehensive study.

These studies are highly based on simulated data and somewhat on real data, it could also have been extended to the comparison of their mathematical formulation. This has been done, to some extent, in the second paper for a single response case but the simultaneous envelope and multi-response case need a separate study.

In the current state, the simulation tool assumes that the predictor components relevant for one response component are not relevant for others. This can be further studied and can be extended to simulate a more general data structure. Additionally, due to the rise in the popularity of machine learning methods, a similar comparative study of statistical and machine learning methods is also recommended as a future perspective of this study.

TOOLS AND RESOURCES

R-PACKAGE:
https://github.com/simulatr/simrel

SHINY APPLICATION:
https://github.com/simulatr/AppSimulatr

THESIS GITHUB REPOSITORY:
https://github.com/therimalaya/Thesis

PAPER 1:
https://github.com/therimalaya/simrel-m

PAPER 2:
https://github.com/therimalaya/model-comparison-paper

PAPER 3:
https://github.com/therimalaya/03-prediction-comparison

PAPER 4:
https://github.com/therimalaya/04-estimation-comparison

REFERENCES

Magne Aldrin. Multivariate prediction using softly shrunk reduced-rank regression. *American Statistician*, 54(1):29–34, 2000. ISSN 15372731. doi: 10.1080/00031305.2000.10474504.

Trygve Almøy. A simulation study on comparison of prediction methods when only a few components are relevant. *Computational Statistics & Data Analysis*, 21(1):87–107, jan 1996. doi: 10.1016/0167-9473(95)00006-2.

T. W. Anderson, I. Olkin, and L. G. Underhill. Generation of Random Orthogonal Matrices. *SIAM Journal on Scientific and Statistical Computing*, 8 (2):625–629, 1987. ISSN 0196-5204. doi: 10.1137/0908055.

Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie, and Jonathan McPherson. *shiny: Web Application Framework for R*, 2018. URL https://CRAN.R-project.org/package=shiny. R package version 1.2.0.

R. D. Cook, I. S. Helland, and Z. Su. Envelopes and partial least squares regression. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 75(5):851–877, 2013. ISSN 13697412. doi: 10.1111/rssb.12018.

R. Dennis Cook. *An introduction to envelopes : dimension reduction for efficient estimation in multivariate statistics*. Hoboken, NJ : John Wiley & Sons, 2018., 1 edition, 2018. ISBN 9781119422952.

R. Dennis Cook and Zhihua Su. Scaled envelopes: Scale-invariant and efficient estimation in multivariate linear regression. *Biometrika*, 100(4): 939–954, 2013. ISSN 00063444. doi: 10.1093/biomet/ast026.

R. Dennis Cook and Xin Zhang. Foundations for Envelope Models and Methods. *Journal of the American Statistical Association*, 110(510):599–611, 2015a. ISSN 1537274X. doi: 10.1080/01621459.2014.983235.

R. Dennis Cook and Xin Zhang. Simultaneous envelopes for multivariate linear regression. *Technometrics*, 57(1):11–25, 2015b. ISSN 15372723. doi: 10.1080/00401706.2013.872700.

R. Dennis Cook and Xin Zhang. Algorithms for Envelope Estimation. *Journal of Computational and Graphical Statistics*, 25(1):284–300, 2016. ISSN 15372715. doi: 10.1080/10618600.2015.1029577.

R. Dennis Cook, Bing Li, and Francesca Chiaromonte. Dimension reduction in regression without matrix inversion. *Biometrika*, 94(3):569–584, aug 2007. ISSN 0006-3444. doi: 10.1093/biomet/asm038.

R Dennis Cook, Bing Li, and Francesca Chiaromonte. Envelope Models for Parsimonious and Efficient Multivariate Linear Regression. *Statistica Sinica*, 20(3):927–1010, 2010. ISSN 10170405.

R. Dennis Cook, Zhihua Su, and Yi Yang. envlp: A MATLAB Toolbox for Computing Envelope Estimators in Multivariate Analysis. *Journal of Statistical Software*, 62(8):??–??, 2015. ISSN 1548-7660. doi: 10.18637/jss.v062.i08.

R. Dennis Cook, Liliana Forzani, and Zhihua Su. A note on fast envelope estimation. *Journal of Multivariate Analysis*, 150:42–54, 2016. ISSN 10957243. doi: 10.1016/j.jmva.2016.05.006.

Sijmen de Jong. SIMPLS: An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 18(3): 251–263, mar 1993. ISSN 01697439. doi: 10.1016/0169-7439(93)85002-X.

D Gamerman and H F Lopes. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference, Second Edition*, volume 1. Taylor & Francis, 2006. ISBN 9781584885870.

Lars Erik Gangsei, Trygve Almøy, and Solve Sæbø. Theoretical evaluation of prediction error in linear regression with a bivariate response variable containing missing data. *Communications in Statistics - Theory and Methods*, 0926(just-accepted):1–9, 2016. ISSN 0361-0926. doi: 10.1080/03610926.2016.1222434.

Gangsei L. E., Almøy T., and Sæbø S. Linear regression with bivariate response variable containing missing data. An empirical Bayes strategy to

increase prediction precision. *Communications in Statistics – Simulation and Computation*, 2016.

Gene H Golub, Charles F Van Loan, and C F V Loan. *Matrix computations*, volume 3. JHU Press, 2012. ISBN 0801854148. doi: 10.1063/1.3060478.

Richard M Heiberger. Algorithm AS 127: Generation of Random Orthogonal Matrices. *Applied Statistics*, 27(2):199, 1978. ISSN 00359254. doi: 10.2307/2346957.

Inge S. Helland. Partial least squares regression and statistical models. *Scandinavian Journal of Statistics*, 17(2):97–114, 1990. ISSN 0303-6898. doi: 10.2307/4616159.

Inge S. Helland. Model Reduction for Prediction in Regression Models. *Scandinavian Journal of Statistics*, 27(1):1–20, mar 2000. ISSN 0303-6898. doi: 10.1111/1467-9469.00174.

Inge S. Helland and Trygve Almøy. Comparison of prediction methods when only a few components are relevant. *Journal of the American Statistical Association*, 89(426):583–591, 1994. ISSN 1537274X. doi: 10.1080/01621459.1994.10476783.

Inge S. Helland, Solve Saebø, and Ha Kon Tjelmeland. Near Optimal Prediction from Relevant Components. *Scandinavian Journal of Statistics*, 39(4):695–713, mar 2012. ISSN 03036898. doi: 10.1111/j.1467-9469.2011.00770.x.

Inge Svein Helland, Solve Saebø, Trygve Almøy, Raju Rimal, Solve Sæbø, Trygve Almøy, and Raju Rimal. Model and estimators for partial least squares regression. *Journal of Chemometrics*, 32(9):e3044, sep 2018. ISSN 08869383. doi: 10.1002/cem.3044.

Ulf Indahl. A twist to partial least squares regression. *Journal of Chemometrics*, 19(1):32–44, 2005. ISSN 08869383. doi: 10.1002/cem.904.

Ulf G. Indahl, Kristian Hovde Liland, and Tormod Næs. Canonical partial least squares-a unified PLS approach to classification and regression problems. *Journal of Chemometrics*, 23(9):495–504, 2009. ISSN 08869383. doi: 10.1002/cem.1243.

R.A. Johnson and D.W. Wichern. *Applied Multivariate Statistical Analysis (Classic Version)*. Pearson Modern Classics for Advanced Statistics Series. Pearson Education Canada, 2018. ISBN 9780134995397. URL https://books.google.no/books?id=QBqlswEACAAJ.

I T Jolliffe. *Principal Component Analysis, Second Edition*. 2002. ISBN 0387954422. doi: 10.2307/1270093.

Owen Jones, Robert Maillardet, and Andrew Robinson. *Introduction to scientific programming and simulation using R*. Chapman and Hall/CRC, 2014.

Henk A.L. Kiers and Age K. Smilde. A comparison of various methods for multivariate regression with highly collinear variables. *Statistical Methods and Applications*, 2007. ISSN 16182510. doi: 10.1007/s10260-006-0025-5.

Øyvind Langsrud. Rotation tests. *Statistics and Computing*, 15(1):53–60, 2005. ISSN 09603174. doi: 10.1007/s11222-005-4789-5.

Minji Lee and Zhihua Su. *Renvlp: Computing Envelope Estimators*, 2018. URL https://CRAN.R-project.org/package=Renvlp. R package version 2.5.

Bjørn-Helge Mevik, Ron Wehrens, and Kristian Hovde Liland. *pls: Partial Least Squares and Principal Component Regression*, 2018. URL https://CRAN.R-project.org/package=pls. R package version 2.7-0.

Tormod Næs and Inge S Helland. Relevant components in regression. *Scandinavian Journal of Statistics*, 20(3):239–250, 1993.

Tormod Naes and Harald Martens. Comparison of prediction methods for multicollinear data. *Communications in Statistics - Simulation and Computation*, 14(3):545–576, jan 1985. ISSN 0361-0918. doi: 10.1080/03610918508812458.

Tormod Næs, Oliver Tomic, Nils Kristian Afseth, Vegard Segtnan, and Ingrid Måge. Multi-block regression based on combinations of orthogonalisation, pls-regression and canonical correlation analysis. *Chemometrics and Intelligent Laboratory Systems*, 124:32–42, 2013.

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/.

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

Raju Rimal, Trygve Almøy, and Solve Sæbø. A tool for simulating multi-response linear model data. *Chemometrics and Intelligent Laboratory Systems*, 176:1 – 10, 2018a. ISSN 0169-7439. doi: https://doi.org/10.1016/j.chemolab.2018.02.009. URL http://www.sciencedirect.com/science/article/pii/S0169743917304823.

Raju Rimal, Trygve Almøy, and Solve Sæbø. A tool for simulating multi-response linear model data. *Chemometrics and Intelligent Laboratory Systems*, 176:1–10, may 2018b. ISSN 18733239. doi: 10.1016/j.chemolab.2018.02.009.

Raju Rimal, Trygve Almøy, and Solve Sæbø. Comparison of multi-response prediction methods. *Chemometrics and Intelligent Laboratory Systems*, 190:10 – 21, 2019. ISSN 0169-7439. doi: https://doi.org/10.1016/j.chemolab.2019.05.004. URL http://www.sciencedirect.com/science/article/pii/S016974391930187X.

B D Ripley. *Stochastic Simulation*, volume 2009. John Wiley & Sons, 1987. ISBN 0471818844. doi: 10.1002/9780470316726.

Solve Sæbø, Trygve Almøy, Arnar Flatberg, Are H. Aastveit, and Harald Martens. LPLS-regression: a method for prediction and classification under the influence of background information on predictor variables. *Chemometrics and Intelligent Laboratory Systems*, 91(2):121–132, 2008. ISSN 01697439. doi: 10.1016/j.chemolab.2007.10.006.

Solve Sæbø, Trygve Almøy, and Inge S. Helland. 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, 2015. ISSN 18733239. doi: 10.1016/j.chemolab.2015.05.012.

Peter D. Wentzell and Lorenzo Vega Montoto. Comparison of principal components regression and partial least squares regression through

generic simulations of complex mixtures. *Chemometrics and Intelligent Laboratory Systems*, 65(2):257 – 279, 2003. ISSN 0169-7439. doi: https://doi.org/10.1016/S0169-7439(02)00138-7. URL http://www.sciencedirect.com/science/article/pii/S0169743902001387.

LIST OF RESEARCH PAPERS

A TOOL FOR SIMULATING MULTI RESPONSE LINEAR MODEL DATA

Chemometrics and Intelligent Laboratory Systems 176 (2018) 1–10

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

A tool for simulating multi-response linear model data

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

ARTICLE INFO

Keywords:
Simrel package in r
Data simulation
Linear model
Multivariate

ABSTRACT

Data science is generating enormous amounts of data, and new and advanced analytical methods are constantly being developed to cope with the challenge of extracting information from such "big-data". Researchers often use simulated data to assess and document the properties of these new methods, and in this paper we present an extension to the R-package simrel, which is a versatile and transparent tool for simulating linear model data with an extensive range of adjustable properties. The method is based on the concept of relevant components, and is equivalent to the newly developed envelope model. It is a multi-response extension of R-package simrel which is available in R-package repository CRAN, and as simrel the new approach is essentially based on random rotations of latent relevant components to obtain a predictor matrix **X**, but in addition we introduce random rotations of latent components spanning a response space in order to obtain a multivariate response matrix **Y**. The properties of the linear relation between **X** and **Y** are defined by a small set of input parameters which allow versatile and adjustable simulations. Sub-space rotations also allow for generating data suitable for testing variable selection methods in multi-response settings. The method is implemented as an update to the R-package simrel.

1. Introduction

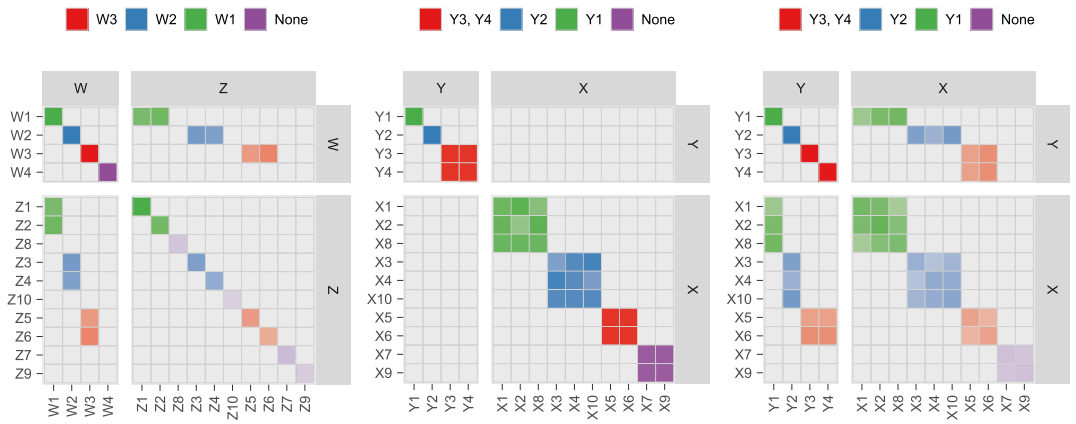
Technological advancement has opened a door for complex and sophisticated scientific experiments that were not possible before. Due to this change, enormous amounts of raw data are generated which contain massive information but is difficult to excavate. Finding information and performing scientific research on these raw data has now become another problem. In order to tackle this situation new methods are being developed. However, before implementing any method, it is essential to test its performance and explore its properties. Often, researchers use simulated data for the purpose which itself is a time-consuming process. The main focus of this paper is to present a simulation method, along with an extension to the r-package called simrel, that is versatile in nature and yet simple to use.

The simulation method we are presenting here is based on the principle of relevant space for prediction [13] which assumes that there exists a y-relevant subspace in the complete space of predictor variables that is spanned by a subset of eigenvectors of these predictor variables. Our extension to this principle is to introduce a subspace in **y** (material space) which contains the information that predictor space is relevant for. The concept of response reduction to the material space in response variable was introduced by Cook et al. [6]. Our r-package based on this principle lets the user specify various population properties such as; which latent components in **x** are relevant for a latent subspace of the responses **y** and the collinearity structure of **x**. This enables the possibility to construct data for evaluating estimation methods and methods developed for variable selection.

Among several publications on simulation, Johnson [16]; Ripley [17] and Gamerman and Lopes [9] have exhaustively discussed the topic. In particular, methods based on covariance structure has been discussed by Arteaga and Ferrer [2]; Arteaga and Ferrer [3] and Camacho [4], following approaches to find simulated data satisfying the desired correlation structure. In addition, many publications have implemented simulated data in order to investigate new estimation methods and prediction strategies [see:8, 5, 14]. However, most of the simulations in these studies were developed to address their specific problem. A systematic tool for simulating linear model data with single response, which could serve as a general tool for all such comparisons, was presented in Sæbø et al. [19] and as the r-package simrel. This paper extends simrel in order to simulate linear model data with multivariate response. The github repository of the package at http://github.com/simulatr/simrel has rich documentation with many examples and cases along with detailed descriptions of simulation parameters. In the following two sections, the discussion encircle the mathematical framework behind. In

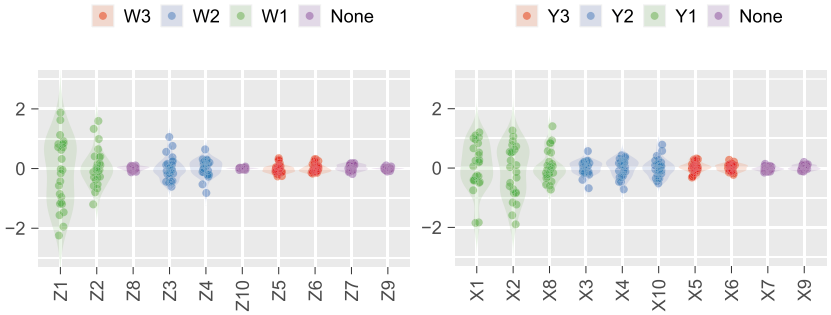
* Corresponding author.
E-mail addresses: raju.rimal@nmbu.no (R. Rimal), trygve.almoy@nmbu.no (T. Almøy), solve.sabo@nmbu.no (S. Sæbø).

https://doi.org/10.1016/j.chemolab.2018.02.009
Received 26 July 2017; Received in revised form 14 February 2018; Accepted 19 February 2018
Available online 23 February 2018



(a) Relevant Position (b) Rotation Matrix (c) Relevant Predictors

Fig. 1. Simulation of predictor and response variables after orthogonal transformation of predictor and response components by rotation matrices Q and R shown as the upper left and the lower right block matrices in (b).



(a) Before Rotation (b) After Rotation

Fig. 2. Simulated data before and after rotation.

Table 1
Parameter setting of simulated data for comparison of estimation methods.

Decay of eigenvalues (γ)	Coef. of Determination (ρ^2_{wj})
Design1 0.2	0.8, 0.8, 0.4
Design2 0.8	0.8, 0.8, 0.4
Design3 0.2	0.4, 0.4, 0.4
Design4 0.8	0.4, 0.4, 0.4

addition, in section 4 and 5? we have also discussed the input parameters needed for simrel function in brief. In section 4, an implementation is presented as a case example and the final section introduces the shiny web application for this tool.

2. Statistical model

In this section we describe the model and the model parameterization which is assumed throughout this paper. We assume:

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

where, $\mathbf{y}$ is a response vector with m response variables $y_1, y_2, \ldots y_m$ with mean vector $\boldsymbol{\mu}_y$, and $\mathbf{x}$ is vector of p predictor variables with mean vector $\boldsymbol{\mu}_x$. Further,

$\boldsymbol{\Sigma}_{yy}(m \times m)$is the variance-covariance matrix of $\mathbf{y}$
$\boldsymbol{\Sigma}_{xx}(p \times p)$is the variance-covariance matrix of variables $\mathbf{x}$
$\boldsymbol{\Sigma}_{yx}(m \times p)$is the matrix of covariance between $\mathbf{x}$ and $\mathbf{y}$

Standard theory in multivariate statistics may be used to show that $\mathbf{y}$ conditioned on $\mathbf{x}$ corresponds to the linear model,

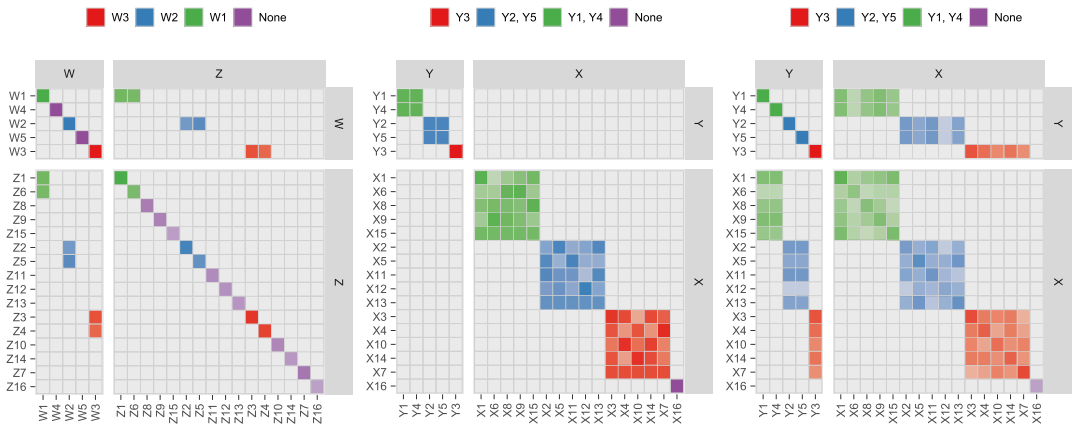
$$\mathbf{y} = \boldsymbol{\mu}_y + \boldsymbol{\beta}^t(\mathbf{x} - \boldsymbol{\mu}_x) + \boldsymbol{\epsilon} \tag{2}$$

where, $\boldsymbol{\beta}^t$ is a $(m \times p)$ matrix of regression coefficients, and $\boldsymbol{\epsilon}$ is an error term such that $\boldsymbol{\epsilon} \sim N(0, \boldsymbol{\Sigma}_{y|x})$. The properties of the linear model (2) can be expressed in terms of covariance matrices in (1).

Regression Coefficients The matrix of regression coefficients is given by

$$\boldsymbol{\beta} = \boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}$$

Coefficient of Determination Since, a matrix of coefficient-of-determination represents the proportion of variation explained by the predictors, we can write this matrix by its elements as,



(a) Relevant Position (b) Rotation Matrix (c) Relevant Predictors

Fig. 3. Simulation of predictor and response variables for design one after orthogonal transformation of predictor and response components by rotation matrices Q and R shown as the upper left and the lower right block matrices in (b). Here (a) is the covariance structure of the latent space, which is rotated by the block diagonal rotation matrix in (b) resulting the covariance structure of simulated data in (c).

Table 2
Minimum average prediction error (number of components corresponding to minimum prediction error, minimum prediction error) (For *Y*env, the number of response components is given).

Model	Design: 1	Design: 2	Design: 3	Design: 4
CPLS	(3, 3.24)	(4, 3.22)	(3, 4.09)	(3, 4.05)
CPPLS	(3, 3.21)	(3, 3.17)	(3, 4.11)	(3, 4.04)
OLS	(1, 3.60)	(1, 3.58)	(1, 4.57)	(1, 4.50)
PCR	(7, 3.28)	(6, 3.19)	(6, 4.08)	(6, 4.04)
PLS1	(2, 3.32)	(5, 3.20)	(1, 4.16)	(5, 4.07)
PLS2	(5, 3.29)	(6, 3.19)	(3, 4.11)	(6, 4.06)
Senv	(4, 3.17)	(5, 3.14)	(3, 4.35)	(5, 4.28)
Xenv	(5, 3.23)	(6, 3.20)	(5, 4.10)	(6, 4.11)
Yenv	(3, 3.24)	(3, 3.23)	(3, 4.29)	(3, 4.24)

$$\left(\rho_y^2\right)_{jj'} = \frac{\sigma_{xy_j}^t \Sigma_{xx}^{-1} \sigma_{xy_{j'}}}{\sqrt{\sigma_{y_j}^2 \sigma_{y_{j'}}^2}} \forall j, j' = 1 \ldots m$$

where, σ_{xy_j}, $\sigma_{xy_{j'}}$ are covariances between $\mathbf{x}$ and y_j, $y_{j'}$ respectively. Also, $\sigma_{y_j}^2$ and $\sigma_{y_{j'}}^2$ are unconditional variances of y_j and $y_{j'}$.Here the numerator is equivalent to the covariance of fitted $\mathbf{y}$ in sample space. if $j = j'$, it corresponds to a population version of the mean sum of squares of regression. The denominator gives the total unconditional variation in $\mathbf{y}$. The diagonal elements of this matrix is the proportion of variation in a response $y_j, j = 1, \ldots m$ explained by the predictors.

Conditional variance The conditional variance-covariance matrix of $\mathbf{y}$ given $\mathbf{x}$ is,

$$\boldsymbol{\Sigma}_{y|x} = \boldsymbol{\Sigma}_{yy} - \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}.$$

The diagonal elements of this matrix equals the minimum least squared error of prediction $[\mathrm{E}(y - \widehat{y})^2]$ for each of the response variables.

Let us define a transformation of $\mathbf{x}$ and $\mathbf{y}$ as, $\mathbf{z} = \mathbf{R}\mathbf{x}$ and $\mathbf{w} = \mathbf{Q}\mathbf{y}$. Here, $\mathbf{R}_{p\times p}$ and $\mathbf{Q}_{m\times m}$ are rotation matrices that rotate $\mathbf{x}$ and $\mathbf{y}$ to yield $\mathbf{z}$ and $\mathbf{w}$, respectively. The model (1) can be re-expressed in terms of these transformed variables as:

$$\begin{aligned}\begin{bmatrix}\mathbf{w}\\ \mathbf{z}\end{bmatrix} &\sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = N\left(\begin{bmatrix}\boldsymbol{\mu}_w\\ \boldsymbol{\mu}_z\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{ww} & \boldsymbol{\Sigma}_{wz}\\ \boldsymbol{\Sigma}_{zw} & \boldsymbol{\Sigma}_{zz}\end{bmatrix}\right)\\ &= N\left(\begin{bmatrix}\boldsymbol{Q}\boldsymbol{\mu}_y\\ \boldsymbol{R}\boldsymbol{\mu}_x\end{bmatrix}, \begin{bmatrix}\boldsymbol{Q}\boldsymbol{\Sigma}_{yy}\boldsymbol{Q}^t & \boldsymbol{Q}\boldsymbol{\Sigma}_{yx}\mathbf{R}^t\\ \boldsymbol{R}\boldsymbol{\Sigma}_{xy}\boldsymbol{Q}^t & \boldsymbol{R}\boldsymbol{\Sigma}_{xx}\mathbf{R}^t\end{bmatrix}\right)\end{aligned} \tag{3}$$

In addition, a linear model relating $\mathbf{w}$ conditioned on $\mathbf{z}$ can be written as,

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

where $\boldsymbol{\alpha}$ is the regression coefficient vector for the transformed model and $\boldsymbol{\tau} \sim N(0, \boldsymbol{\Sigma}_{w|z})$. Further, if both $\mathbf{Q}$ and $\mathbf{R}$ are orthonormal matrices, i.e., $\mathbf{Q}^t\mathbf{Q} = \mathbf{I}_m$ and $\mathbf{R}^t\mathbf{R} = \mathbf{I}_p$, the inverse transformation can be defined as,

$$\begin{aligned}\boldsymbol{\Sigma}_{yy} &= \mathbf{Q}^t\boldsymbol{\Sigma}_{ww}\mathbf{Q} \quad \boldsymbol{\Sigma}_{yx} = \mathbf{Q}^t\boldsymbol{\Sigma}_{wz}\mathbf{R}\\ \boldsymbol{\Sigma}_{xy} &= \mathbf{R}^t\boldsymbol{\Sigma}_{zw}\mathbf{Q} \quad \boldsymbol{\Sigma}_{xx} = \mathbf{R}^t\boldsymbol{\Sigma}_{zz}\mathbf{R}\end{aligned} \tag{5}$$

From this, we can find a direct connection between different population properties of (2) and (4).

Regression Coefficients:

$$\boldsymbol{\alpha} = \boldsymbol{\Sigma}_{wz}\boldsymbol{\Sigma}_{zz}^{-1} = \boldsymbol{Q}\boldsymbol{\Sigma}_{yz}\mathbf{R}^t[\boldsymbol{R}\boldsymbol{\Sigma}_{xx}\mathbf{R}^t]^{-1} = \mathbf{Q}\left[\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\right]\mathbf{R}^t = \mathbf{Q}\boldsymbol{\beta}\mathbf{R}^t$$

Conditional Variance Further, the conditional variance-covariance matrix of $\mathbf{w}$ given $\mathbf{z}$ is,

$$\begin{aligned}\boldsymbol{\Sigma}_{w|z} &= \boldsymbol{\Sigma}_{ww} - \boldsymbol{\Sigma}_{wz}\boldsymbol{\Sigma}_{zz}^{-1}\boldsymbol{\Sigma}_{zw}\\ &= \boldsymbol{Q}\boldsymbol{\Sigma}_{yy}\mathbf{Q}^t - \boldsymbol{Q}\boldsymbol{\Sigma}_{yx}\mathbf{R}^t[\boldsymbol{R}\boldsymbol{\Sigma}_{xx}\mathbf{R}^t]^{-1}\boldsymbol{R}\boldsymbol{\Sigma}_{xy}\mathbf{Q}^t\\ &= \boldsymbol{Q}\boldsymbol{\Sigma}_{yy}\mathbf{Q}^t - \boldsymbol{Q}\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}\mathbf{Q}^t\\ &= \mathbf{Q}\left[\boldsymbol{\Sigma}_{yy} - \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}\right]\mathbf{Q}^t = \mathbf{Q}\boldsymbol{\Sigma}_{y|x}\mathbf{Q}^t\end{aligned}$$

Coefficient of Determination The coefficient-of-determination matrix corresponding to $\mathbf{w}$ can be written as,

$$\begin{aligned}\left(\rho_w^2\right)_{jj'} &= \boldsymbol{\Sigma}_{ww}^{-1/2}\boldsymbol{\Sigma}_{wz}\boldsymbol{\Sigma}_{zz}^{-1}\boldsymbol{\Sigma}_{zw}\boldsymbol{\Sigma}_{ww}^{-1/2}\\ &= \frac{\sigma_{zw_j}^t \Sigma_{zz}^{-1} \sigma_{zw_{j'}}}{\sqrt{\sigma_{w_j}^2 \sigma_{w_{j'}}^2}} \forall j, j' = 1 \ldots m\end{aligned}$$

where, σ_{zw_j} and $\sigma_{zw_{j'}}$ are covariances of $\mathbf{z}$ with w_j and $w_{j'}$, respectively. Also, $\sigma_{w_j}^2$ and $\sigma_{w_{j'}}^2$ are unconditional variances of w_j and $w_{j'}$. For simplicity,

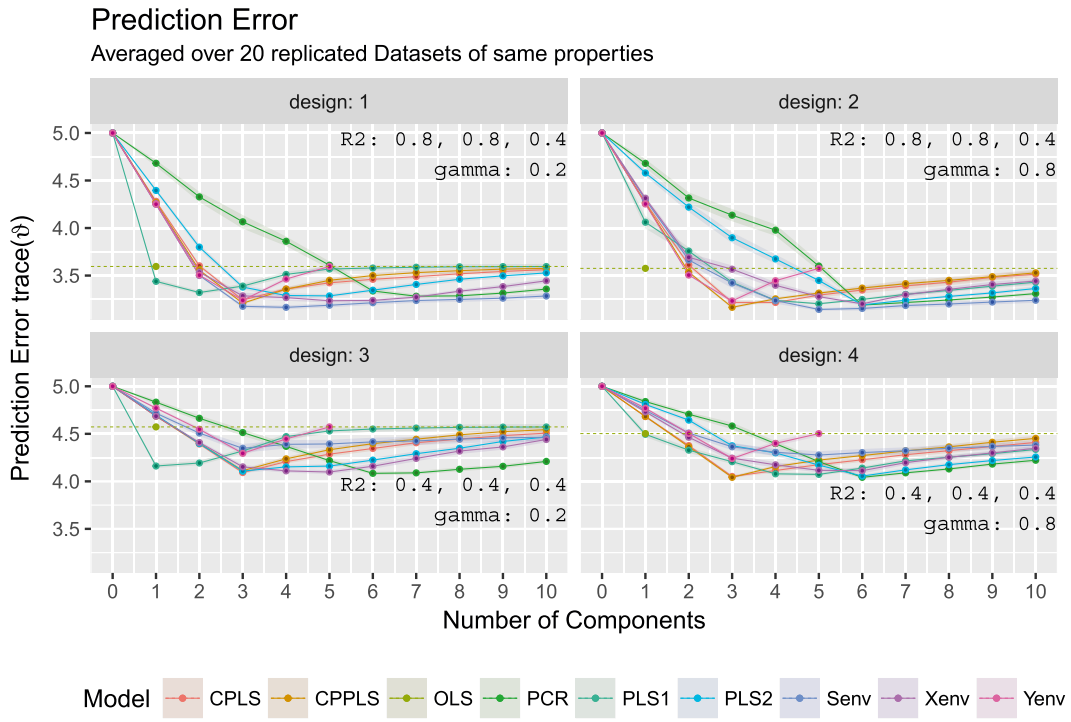


Fig. 4. Minimum of average prediction error.

Table 3
Simulation Design of second example.

η : 0.1	η : 0.8	Parameter	Value
Single Informative Response Component			
Design 1	*Design 2*	relpos	2, 3, 5, 7
		q	1000
		R2	0.8
Two Informative Response Components			
Design 3	*Design 4*	relpos	2; 3
		q	500; 500
		R2	0.6; 0.6

we will denote $\sigma_{z_i w_j}$ by σ_{ij}.

Since the rotation matrices give a direct connection between the covariance of (1) and (3), a straight forward relationship can be worked out between the terms in the above given matrix and their counterpart covariance matrices of the **xy**-space.

From the eigenvalue decomposition principle, if $\mathbf{\Sigma}_{xx} = \mathbf{R\Lambda R}^t$ and $\mathbf{\Sigma}_{yy} = \mathbf{Q\Omega Q}^t$ then **z** and **w** can be interpreted as principal components of **x** and **y** respectively. In this paper, these principal components will be termed as *predictor components* and *response components* respectively. Here, $\mathbf{\Lambda}$ and $\mathbf{\Omega}$ are diagonal matrices of eigenvalues of $\mathbf{\Sigma}_{xx}$ and $\mathbf{\Sigma}_{yy}$, respectively.

3. Relevant components

Consider a single response linear model with p predictors.

$$y = \mu_y + \boldsymbol{\beta}^t(\mathbf{x} - \mu_x) + \varepsilon$$

where, $\varepsilon \sim N(0, \sigma^2)$ and **x** is a vector of random predictors. Following the concept of relevant space and irrelevant space which is discussed extensively in Helland and Almøy [13], Helland [12], Helland et al. [14], Cook et al. [5], and Sæbø et al. [19], we can assume that there exists a subspace of the full predictor space which is relevant for y. An orthogonal space to this space does not contain any information about y and is considered as irrelevant. Here, the $y-$ relevant subspace of **x** is spanned by a subset of the principal components defined by the eigenvectors of the covariance matrix of **x**, i.e. $\mathbf{\Sigma}_{xx}$.

This concept can be extended to m responses so that the subspace of **x** is relevant for a subspace of **y**. This corresponds to the concept of simultaneous envelopes [8] where relevant (material) and irrelevant (immaterial) space were discussed for both response and predictor variables.

3.1. Model parameterization

In order to construct a fully specified and unrestricted covariance matrix of **z** and **w** for the model in equation (3), we need to identify $1/2(p+m)(p+m+1)$ unknown parameters. For the purpose of simulation, we implement some assumptions to re-parameterize and simplify the model. This enables us to construct a wide range of model properties from only few key parameters.

Parameterization of $\mathbf{\Sigma}_{zz}$ If we let the rotation matrix **R** correspond to the eigenvectors of $\mathbf{\Sigma}_{xx}$, then **z** becomes the set of principal components of **x**. In that case $\mathbf{\Sigma}_{zz}$ is a diagonal matrix with eigenvalues $\lambda_1, \ldots, \lambda_p$. Further, we adopt the same parametric representation as Sæbø et al. [19] for these eigenvalues:

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

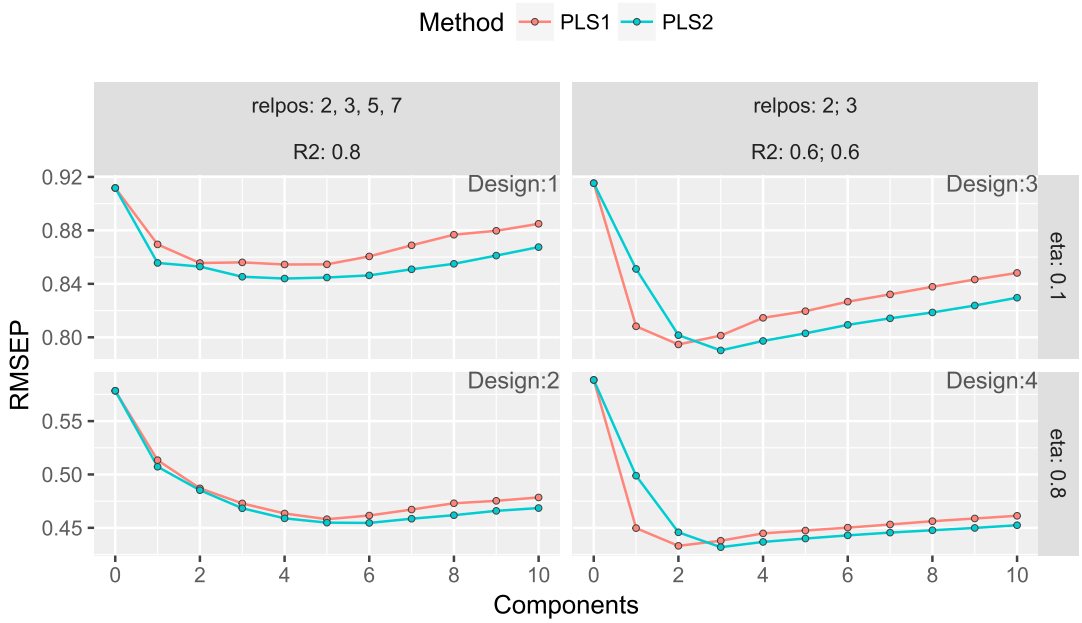


Fig. 5. Root mean square of error of prediction of test observation averaged over all response variables.

Here, as γ increases, the decline of eigenvalues becomes steeper, hence the parameter γ controls the level of multicollinearity in $\mathbf{x}$. We can write $\Sigma_{zz} = \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_p)$.

Parameterization of Σ_{ww} In similar manner, a parametric representation of eigenvalues corresponding to Σ_{ww} is adopted as,

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

Here, the decline of eigenvalues becomes steeper as η increases from zero. At $\eta = 0$, all w will have equal variance 1. Hence we can write $\Sigma_{ww} = \text{diag}(\kappa_1, \ldots, \kappa_m)$.

Parameterization of Σ_{zw} After parameterization of $\mathbf{\Sigma}_{zz}$ and $\mathbf{\Sigma}_{ww}$, we are left with $m \times p$ number of unknowns corresponding to $\mathbf{\Sigma}_{zw}$. Some of the elements of $\mathbf{\Sigma}_{zw}$ may be equal to zero, which implies that the given z is irrelevant for the given variable w. The non-zero elements define which of the z that are relevant for $\mathbf{w}$. We typically refer to the indices of these z variables as the positions of relevant components. In order to re-parameterize this covariance matrix, it is necessary to discuss the position of relevant components in detail.

3.1.1. Position of relevant components

Let k_1 components be relevant for w_1, k_2 components be relevant for w_2 and so on. Let the positions of these components be given by the index sets $\mathscr{P}_1, \mathscr{P}_2, \ldots, \mathscr{P}_m$ respectively. Further, the covariance between w_j and z_i is non-zero only if z_i is relevant for w_j. If σ_{ij} is the covariance between w_j and z_i then $\sigma_{ij} \neq 0$ if $i \in \mathscr{P}_j$ where $i = 1, \ldots, p$ and $j = 1, \ldots, m$ and $\sigma_{ij} = 0$ otherwise.

In addition, the true regression coefficients α for w_j (4) is given by:

$$\boldsymbol{\alpha}_j = \Lambda^{-1}\sigma_{ij} = \sum_{i \in \mathscr{P}_j} \frac{\sigma_{ij}}{\lambda_i}, \qquad j = 1, 2, \ldots m$$

The positions of the relevant components have heavy impact on prediction. Helland and Almøy [13] have shown that if the relevant components have large eigenvalues (variances), which here implies small index values in $\mathscr{P}_j$, prediction of $\mathbf{y}$ from $\mathbf{x}$ is relatively easy and if the eigenvalues (variances) of relevant components are small, the prediction becomes difficult, given that the coefficient of determination and other model parameters are held constant. For example, if the first and second components, z_1 and z_2, are relevant for w_1 and fifth and sixth components, z_5 and z_6, are relevant for w_2, it is relatively easier to predict w_1 than w_2, other properties being similar. This might be so, because the first and second principal components have larger variances than the fifth and sixth components.

Although the covariance matrix may depend on few relevant components, we can not choose these covariances freely since we also need to satisfy following two conditions:

- The covariance matrices Σ_{zz}, Σ_{ww} and Σ must be positive definite
- The covariance σ_{ij} must satisfy user defined coefficient of determination

We have the relation,

$$\boldsymbol{\rho}_w^2 = \mathbf{\Sigma}_{ww}^{-1/2}\mathbf{\Sigma}_{zw}^t\mathbf{\Sigma}_{zz}^{-1}\mathbf{\Sigma}_{zw}\mathbf{\Sigma}_{ww}^{-1/2}$$
$$= \frac{\sigma_{ij}^t\Lambda^{-1}\sigma_{ij'}}{\sqrt{\sigma_j^2\sigma_{j'}^2}} \forall j, j' = 1 \ldots m$$

Applying our assumptions that, $\mathbf{\Sigma_{ww}} = \text{diag}(\kappa_1, \ldots, \kappa_m)$ (7) and $\mathbf{\Sigma}_{zz} = \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_p)$ (6), we obtain,

$$\rho_w^2 = \Sigma_{ww}^{-1/2}\Sigma_{zw}^t\Lambda^{-1}\Sigma_{zw}\Sigma_{ww}^{-1/2} = \begin{bmatrix} \sum_{i=1}^{p} \frac{\sigma_{i1}^2}{\lambda_i\kappa_1} & \cdots & \sum_{i=1}^{p} \frac{\sigma_{i1}\sigma_{im}}{\lambda_i\sqrt{\kappa_1\kappa_m}} \\ \vdots & \ddots & \vdots \\ \sum_{i=1}^{p} \frac{\sigma_{i1}\sigma_{im}}{\lambda_i\sqrt{\kappa_1\kappa_m}} & \cdots & \sum_{i=1}^{p} \frac{\sigma_{im}^2}{\lambda_i\kappa_m} \end{bmatrix}$$

Furthermore, we assume that there are no overlapping relevant components for any two w, i.e, $\mathscr{P}_j \cap \mathscr{P}_{j^*} = \varnothing$ or $\sigma_{ij}\sigma_{ij^*} = 0$ for $j \neq j^*$. The additional unknown parameters in the diagonal of $\boldsymbol{\rho}_w^2$ should agree with

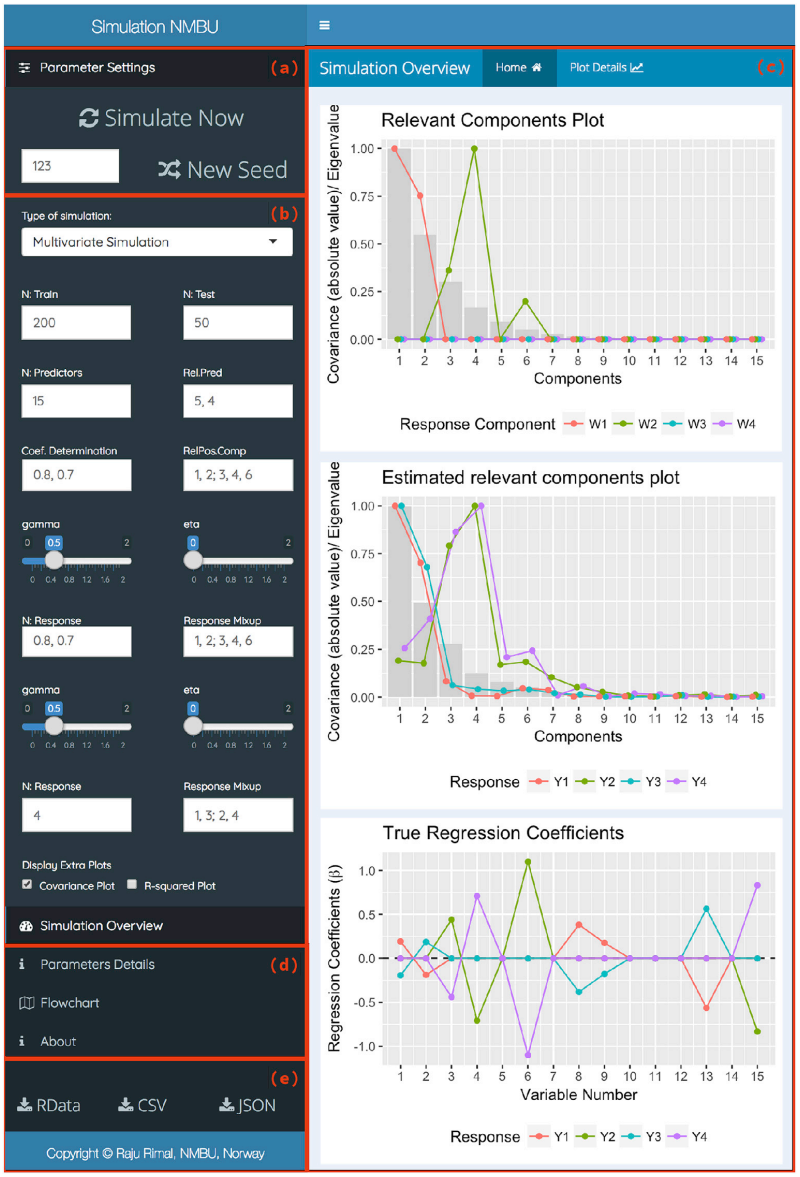


Fig. 6. Web interface of shiny application of simrel: (**a**) Buttons to trigger simulation, (**b**) Parameters for simulation, (**c**) Visualization of the true properties of simulated data (regression coefficients, true and estimated covariance between response and predictors components) (**d**) Additional analysis (**e**) Download option of simulated data.

user specified coefficients of determination for $\mathbf{w}$. i.e, $\rho^2_{w_j}$ is,

$$\rho^2_{w_j} = \sum_{i=1}^{p} \frac{\sigma^2_{ij}}{\lambda_i \kappa_j}$$

Here, only the relevant components have non-zero covariances with $\boldsymbol{w}_j$, so,

$$\rho^2_{w_j} = \sum_{i \in \mathscr{P}_j} \frac{\sigma^2_{ij}}{\lambda_i \kappa_j}$$

For some user defined $\rho^2_{w_j}$ the σ^2_{ij} is determined as follows,

1. Sample k_j values from a uniform distribution $\mathscr{U}(-1, 1)$ distribution. Let them be denoted $\mathscr{S}_{\mathscr{P}_1}, \ldots, \mathscr{S}_{\mathscr{P}_{k_j}}$.
2. Define,

$$\sigma_{ij} = \text{Sign}(\mathscr{S}_i) \sqrt{\frac{\rho^2_{w_j} \left|\mathscr{S}_i\right|}{\sum_{k \in \mathscr{P}_j} \left|\mathscr{S}_k\right|} \lambda_i \kappa_j}$$

for $i \in \mathscr{P}_j$ and $j = 1, \ldots m$

This means that the covariances between the predictor components and the response components are sampled randomly, but with restriction

```
simrel(
  n      = 200, # Number of training observations
  ntest  = 50, # Number of test observations
  p      = 15, # Number of predictor variables
  q      = c(5, 4), # Number of relevant predictors
  relpos = list(c(1, 2), c(3, 4, 6)),
         # Relevant predictor components
  R2     = c(0.8, 0.7), # Rsq for each response component
  m      = 4, # Number of response variables
  gamma  = 0.6, # Decay factor of eigenvalues of predictors
  eta    = 0, # Decay factor of eigenvalues of responses
  ypos   = list(c(1, 3), c(2, 4)),
         # Combination of response components on rotation
  type   = "multivariate"
)
```

that the requested $\rho^2_{w_j}$ values are satisfied. This also implies that the regression coefficients $\boldsymbol{\alpha}$ in (4) and $\boldsymbol{\beta}$ in (2) are sampled randomly under the same restriction.

3.1.2. Data simulation

From the above given parameterizations and the user defined choices of model parameters, a fully defined and known covariance matrix $\boldsymbol{\Sigma}$ of $(\mathbf{w}, \mathbf{z})$ is given. For the simulation of a single observation of $(\mathbf{w}, \mathbf{z})$ let us define $\mathbf{g} = \boldsymbol{\Sigma}^{-1/2}\mathbf{u}$ such that $\text{cov}(\mathbf{g}) = \boldsymbol{\Sigma}$. Here $\boldsymbol{\Sigma}^{-1/2}$ is obtained from Choleskey decomposition of $\boldsymbol{\Sigma}$, and $\mathbf{u}$ is simulated from independent standard normal distribution.

Similarly, in order to simulate n observations, we define $\underset{n\times(m+p)}{\mathbf{G}} = \mathbf{U}\boldsymbol{\Sigma}^{-1/2}$. Here the first m columns of $\mathbf{G}$ will serve as $\mathbf{W}$ and remaining p columns will serve as $\mathbf{Z}$. Further, each row of $\mathbf{G}$ will be a vector sampled independently from the joint normal distribution of $(\mathbf{w}, \mathbf{z})$. Finally, these simulated matrices $\mathbf{W}$ and $\mathbf{Z}$ are orthogonally rotated in order to obtain $\mathbf{Y}$ and $\mathbf{X}$, respectively. In the following section we discuss these rotation matrices in more detail.

3.2. Rotation of predictor space

Initially, let us consider an example where a regression model with $p = 10$ predictors $(\mathbf{x})$ and $m = 4$ responses $(\mathbf{y})$. Let's assume that only three response components $(w_1, w_2$ and $w_3)$ are needed to describe all four response variables. Further, let the index sets $\mathscr{P}_1 = \{1, 2\}, \mathscr{P}_2 = \{3, 4\}$ and $\mathscr{P}_3 = \{5, 6\}$ define the positions of the predictor components of $\mathbf{x}$ that are relevant for w_1, w_2 and w_3, respectively. Let $\mathscr{S}_1, \mathscr{S}_2$ and $\mathscr{S}_3$ be the orthogonal spaces spanned by each set of predictor components. These spaces together span $\mathscr{S}_k = \mathscr{S}_1 \oplus \mathscr{S}_2 \oplus \mathscr{S}_3$, which is the minimum relevant space and equivalent to the x-envelope as discussed by Cook et al. [5].

Moreover, let $q_1 = 3, q_2 = 3$ and $q_3 = 2$ be the number of predictor variables we want to have relevant for w_1, w_2 and w_3 respectively. Then $q_1 = 3$ predictors may be obtained by rotating the predictor components in $\mathscr{P}_1$ along with one more irrelevant component. Similarly, $q_2 = 3$ predictors, relevant for w_2, can be obtained by rotating predictor components in $\mathscr{P}_2$ along with one more irrelevant component and finally, $q_3 = 2$ predictors, relevant for w_3, can be obtained by rotating the components in $\mathscr{P}_3$ without any additional irrelevant component. Let the space spanned by the q_1, q_2 and q_3 number of predictors be $\mathscr{S}_{q_1}, \mathscr{S}_{q_2}$ and $\mathscr{S}_{q_3}$. Together they span a space $\mathscr{S}_q = \mathscr{S}_{q_1} \oplus \mathscr{S}_{q_2} \oplus \mathscr{S}_{q_3}$. This space is bigger than $\mathscr{S}_k$ since in the process two irrelevant components were included in the rotations. Here, $\mathscr{S}_k$ is orthogonal to $\mathscr{S}_{p-k}$ and $\mathscr{S}_q$ is orthogonal to $\mathscr{S}_{p-q}$. Generally speaking, here we are splitting the complete variable space $\mathscr{S}_p$ into two orthogonal spaces – $\mathscr{S}_k$ relevant for $\mathbf{w}$ and $\mathscr{S}_{p-k}$ irrelevant for $\mathbf{w}$.

In the previous section, we discussed about the construction of a covariance matrix for the latent structure. Fig. 1(a) shows a similar structure resembling the example here. The three colors represent the relevance with the three latent response components $(w_1, w_2$ and $w_3)$. Here we can see that z_1 and z_2 (first and second predictor components of $\mathbf{x}$) have non-zero covariance with w_1 (first latent component of response $\mathbf{y}$). In the similar manner other non-zero covariances are self-explanatory.

In order to simulate predictor variables $(\mathbf{x})$, we construct matrix $\mathbf{R}$ which then is used for orthogonal rotation of the predictor components $\mathbf{z}$. This defines a new basis for the same space as is spanned by the predictor components. In principle, there are many possible options for defining a rotation matrix. Among them, the eigenvector matrix of $\boldsymbol{\Sigma}_{xx}$ can be a candidate. However, in this reverse engineering approach both rotation matrices $\mathbf{R}$ and $\mathbf{Q}$ along with the covariance matrices $\boldsymbol{\Sigma}_{xx}$ are unknown. So, we are free to choose any $\mathbf{R}$ that satisfies the properties of a real valued rotation matrix, i.e $\mathbf{R}^{-1} = \mathbf{R}^t$ and $\det(\mathbf{R}) = \pm 1$ so that $\mathbf{R}$ is orthonormal. Here the rotation matrix $\mathbf{R}$ should be block diagonal as in Fig. 1(b) in order to rotate spaces $\mathscr{S}_1, \mathscr{S}_2 \ldots$ separately. Fig. 2(a) shows the simulated predictor components $\mathbf{z}$ that we are following in our example where we can see that the components z_1 and z_2 (relevant for w_1) is getting rotated together with an irrelevant component z_8. The resultant predictors (Fig. 2(b)) x_1, x_2 and x_8 will hence also be relevant for w_1. In the figure, we can see that components z_7, z_8, z_9 and z_{10} are not relevant for any responses before rotation, however, the x_8, x_{10} predictors become relevant after rotation keeping x_7 and x_9 still irrelevant.

Among several methods [1,11] for generating random orthogonal matrix, in this paper we are using orthogonal matrix $\mathscr{E}$ obtained from QR-decomposition of a matrix filled with standard normal variates. The rotation here can be a) restricted and b) unrestricted. The latter rotates all components $\mathbf{z}$ together and makes all predictor variables somewhat relevant for all response components. However, the former performs a block-wise rotation so that it rotates certain selected predictor

components together. This gives control for specifying certain predictors as relevant for selected responses, which was discussed in our example above. This also allows us to simulate irrelevant predictors such as x_7 and x_9 which can be detected during variable selection procedures.

3.3. Rotation of response space

The previous example has four response variables with only three informative components w_1, w_2 and w_3. During the rotation procedure, the response space is also rotated along with the predictor space. Fig. 1 shows that the informative response component w_3 is rotated together with the uninformative response component w_4 so that the predictors which were relevant for w_3 will be relevant for response variables y_3 and y_4. Similarly, response components w_1 and w_2 are rotated separately so that predictors relevant for w_1 and w_2 will only be relevant for y_1 and y_2 respectively, which we can see in Fig. 2. Although the response components have exclusive set of relevant predictors, the rotation of the response space has the potential of creating several response variables that depend on the same relevant predictor space. In the r-package *simrel*, the combining of the response components is specified by a parameter ypos.

4. Implementation

This section demonstrates an application of multi-response extension of simrel with two examples in order to compare different estimation methods on the basis of prediction error. These example are simply a demonstration of the use of simrel package rather than an extensive comparison of methods.

4.1. Example 1

For the comparison, we have considered four well established estimation methods.

a) Ordinary Least Squares (OLS),
b) Principal Component Regression (PCR),
c) Partial Least Squares predicting individual response variable separately (PLS1) and
d) Partial Least Squares predicting all response variables together (PLS2).

We have also considered four relatively new estimation methods in multi-response regression:

a) Canonically Powered Partial Least Squares regression (CPPLS) [15],
b) Canonical Partial Least Squares regression (CPLS) [15],
c) Envelope estimation in predictor space (Xenv) [6],
d) Envelope estimation in response space (Yenv) [7] and
e) Simultaneous estimation of x- and y-envelope (Senv) [8].

From the possible combinations of two levels of coefficient of determination (ρ^2) and two levels of γ (6) (the factor that controls the multicollinearity in predictor variables), four simulation designs (design 1–4) were prepared. Replicating each design 20 times, 80 datasets with five response variables ($m = 5$) and 16 predictor variables ($p = 16$) were simulated using the method discussed in this paper. It was also assumed that three response components (w_1, w_2 and w_3) completely describe the variation present in five response variables ($y_1 \ldots y_5$). Here, in this example we have assumed that all w's have equal variance, i.e. $\Sigma_{ww} = \mathbf{I}_m$, that is, $\eta = 0$ in (7). The four designs are presented in Table 1. All datasets contained 100 sampled observations and out of 16 predictor variables, three disjoint sets of five predictor variables each are relevant for response components w_1, w_2 and w_3. Although the simulation method is well equipped to simulate data with $p \gg n$, for incorporating envelope estimation methods, which are based on maximization of likelihood, we have chosen a $n > p$ situation in the example. Further, predictor components z_1 and z_6 were relevant for response component w_1, predictor components z_2 and z_5 were relevant for response component w_2 and predictor component z_3 and z_4 were relevant for response component w_3. In addition, following the discussion about rotation of response space (section 3.3), w_1 was rotated together with w_4 and w_2 was rotated together with w_5. Fig. 3 visualizes the covariance structure and relationship between the response and predictor variables for the first design.

For each method, we can write an expected squared prediction error as,

$$\underset{m\times m}{\vartheta} = \mathrm{E}\left[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \boldsymbol{\Sigma}_{xx} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right] + \boldsymbol{\Sigma}_{y|x} \tag{8}$$

where, $\widehat{\boldsymbol{\beta}}$ is an estimate of the true regression coefficient $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}_{xx}$ is the true covariance structure of the predictor variables obtained from simrel. Also, $\boldsymbol{\Sigma}_{y|x}$ is the true minimum error of the model. Here $\widehat{\boldsymbol{\beta}}$ varies across different estimation methods while the remaining terms are the same for each dataset design. The expression in (8) is estimated from 20 replicated calibration sets. Further, an overall prediction error of all responses is measured by the trace of ϑ (8).

The minimum prediction error (measured as discussed above) for nine estimation methods averaged over 20 replications of four designs are shown in Table 2. The table also gives the number of predictor components (response components in case of Yenv), a method has used in order to obtain the minimum of average prediction error.

Table 2 shows that the simultaneous envelope has prediction error of 3.17 and 3.14 in design 1 (with 4 components) and design 2 (with 5 components), respectively, which is smaller than other methods. However, the method was not able to show the same performance in design 3 and design 4. The PCR model has the smallest prediction error (4.08) from 6 components in design 3 and Canonically Powered PLS has minimum prediction error (4.04) from 3 components in design 4. In design 3, we can also see that the Canonical PLS method has second best performance with only three components. The number of components vary across different replicated dataset, but the component corresponding to minimum prediction error is discussed here. A detailed picture of prediction error for each estimation method obtained for each additional component is shown in Fig. 4. Although designs 2 and 4 have higher levels of multicollinearity, the performance of the estimation methods is indifferent to its effect. Since all methods, except OLS, are based on shrinking of estimates, they are less influenced by the multicollinearity problem.

The analysis presented in Fig. 4 has addressed some questions such as how methods work when there exist a true reduced dimension in response space, but also raised other questions like why they perform differently. For example, what is the reason for the decreasing relative performance of the simultaneous envelope method as the ρ^2 values are reduced? Does this depend on the dimensions and shape of the **y** envelopes? Since the example is merely intended as a demonstration of how simrel can be used in scientific study, a more elaborative studies would be necessary to answer such questions, but for this purpose simrel would be a powerful tool.

4.2. Example 2

In this second example, wide matrices with 100 observations and 1000 predictor variables were simulated. Since wide matrices are common in various fields such as genomics, spectroscopy and chemometrics, we set up this second example to compare two variants of partial least square regression – PLS1 and PLS2. While estimating regression coefficients PLS1 uses each response variable separately, while PLS2 uses them all simultaneously. A simulation design was constructed as in Table 3. With each design, 20 replicated datasets were simulated having five response variables and a moderate level of multicollinearity within the predictor variables ($\gamma = 0.5$).

The comparison were based on the prediction error measured by root mean square error of prediction (RMSEP). In order to approximate the error to theoretically computed error, 1000 extra test samples were drawn from the same distribution as the training samples during simulation.

One to ten components were used to fit the simulated data models. The prediction error was recorded for each response variable and each additional component. The first and second design in Table 3 has one informative response component for which four predictor components are relevant at positions 2, 3, 5 and 7, and the coefficient of determination is 0.8. Since the informative response component is rotated together with four uninformative response components, the information is shared among all five response variables after rotation.

The third and fourth design has two informative response components. The first response component has one relevant predictor component at position 2 and a coefficient of determination of 0.6. Similarly, the second response component has one relevant predictor component at position 3 and also here the coefficient of determination is 0.6.

In addition to having one and two response component models, two levels of variance structure of the response components is considered and defined by η parameters with values 0.1 and 0.8 respectively. In the first and third design, all response components vary in similar manner ($\eta = 0.1$), while in the second and fourth design the informative response components have higher variance ($\eta = 0.8$) than the uninformative ones as the eigenvalues of Σ_{ww} drop faster in this case.

Fig. 5 shows the average prediction error of test observations modelled by PLS1 and PLS2 for all four designs. The prediction errors are averaged over all 20 replicated datasets.

In general, PLS2 dominates PLS1 with regard to minimum error achieved for these simulated designs. The difference is largest for the designs with $\eta = 0.1$ in which case the response are moderately correlated and prediction appears to be more difficult than for $\eta = 0.8$. The effect of number of relevant response and predictor components appears to have less influence on the results than the covariance structure of Σ_{yy}. This small example of the use of simrel indicates that a more elaborate comparison study should be done on PLS1 and PLS2 in this respect.

5. Web interface

In order to give an alternative interface for simrel, we have created a shiny app which allows users to provide the simulation parameters through different input fields. Fig. 6 shows a screenshot of the application. The application contains three main sections through which the user can interact with this simulation approach. A random seed can be selected using section Fig. 6 (a) so that a particular set of data can be re-simulated if needed. Fig. 6 (b) has all the input panels where the user-dependent parameters for simulation can be entered. Here the user also has the option to simulate univariate, bivariate or multivariate response data. In addition, a simulated R-object comprising the simulated data can be downloaded in Rdata format (section (e) in Fig. 6). The object holds the simulated data along with other properties such as coefficient of determination for each response, true regression coefficients and rotation matrices. Users can also download simulated data in JSON and CSV format.

All simrel parameters can be entered using a simple user interface where vector elements are separated with comma (,) and list elements are separated with semicolon (;). For instance, the relevant position discussed in the implementation (section 4) of this paper can be entered as 1, 6; 2, 5; 3, 4 which is equivalent to R syntax list (c(1, 6), c (2, 5), c(3, 4)). An R expression equivalent to the input parameters as shown in Fig. 5(b) can be written as,

With the parameters for simulation in the screenshot (Fig. 6) 200 training sets (n) and 50 test sets (ntest) will be simulated with 15 predictor variables (p) and 4 response variables (m). The 4 response variables will have a true latent dimension of two, which is spanned by the relevant *response components*. The first response component is rotated together with the third (irrelevant) response component and the second response component is rotated together with the fourth (irrelevant) response component as set in ypos. Out of 15 predictors, 5 will be relevant for the first response component and 4 will be relevant for the second response component, as set by q. The 5 predictor variables, that are relevant for the first response component, span the same space as the predictor components at position 1 and 2. Similarly, the 4 predictor variables that are relevant for the second response component, span the same space as the predictor components at position 3, 4 and 6 (relpos). The coefficient of determination for the first and second response components are 0.8 and 0.7, respectively (R2). The eigenvalues of the predictor components decay exponentially by the factor of 0.6 (gamma), whereas the eigenvalues of response components are constant (but can be set to exponential decay) (eta).

The application not only allows users to simulate data, but also gives some insight into simulated data properties. Section (c) in Fig. 6 contains three plots – a) true regression coefficients b) relevant components and c) estimated relevant components. In the first plot (Fig. 6(c) top) we can see that predictor variables (1, 2, 8, 9 and 13) are relevant for the first and third response variables (red and blue line) by their non-zero coefficients, whereas predictor variables (3, 4, 6 and 15) are relevant for the second and fourth response variables (purple and green line). The second plot (Fig. 6(c) middle) shows the covariances between the response components and the predictor components along with the corresponding eigenvalues in the background (bar plot). In the plot the absolute value of the covariances after scaling with the largest covariance are shown. As in our parameter setting, the plot shows that the first (red line) and second (green line) predictor components have non-zero covariance with the first and third response components, and the fourth and sixth predictor components have non-zero covariance with the second response component. The third plot (Fig. 6(c) bottom) is the estimated covariances between the predictor components and the response variables, for the simulated data. Since the first and third response components are rotated together, in the plot, the covariance between the predictor components and the first and third response variables (red and blue line) are following similar patterns as the theoretical (6(c) middle). This also suggests that the predictor components which were relevant for the first response component, becomes relevant for the first and third response variables after rotation.

Along with these main sections, section (d) in the same figure contains additional analysis performed on the simulated data such as its estimation with different methods. This section is intended for educational purposes to show how changing the data properties influences the performances of different estimation and prediction methods. Beside this application, for Rstudio users, a gadget will be available after installing the r-package. This gadget provides an interface enabling users to input simulation parameters and access some of the properties.

Many scientific studies [8,14,18] are using simulated data in order to compare their findings with others or assess its properties. In many of these situations, a user-friendly and versatile simulation tool like simrel can play an important role. Gangsei et al. [10] and Sæbø et al. [19] are some examples where the univariate and bivariate form of simrel have been used for such purposes.

6. Conclusion

Whether comparing methods or assessing and understanding the properties of any method, tool or procedure; simulated data allows for controlled tests for researchers. However, researchers spend enormous amount of time creating such simulation tools so that they can obtain a particular nature of data. We believe that this tool along with the R-package and the easy-to-use shiny web interface will become an assistive tool for researchers in this respect.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.chemolab.2018.02.009.

References

[1] T.W. Anderson, I. Olkin, L.G. Underhill, Generation of random orthogonal matrices, SIAM J. Sci. Stat. Comput. 8 (4) (1987) 625–629.

[2] F. Arteaga, A. Ferrer, How to simulate normal data sets with the desired correlation structure, Chemometr. Intell. Lab. Syst. 101 (1) (2010) 38–42.

[3] F. Arteaga, A. Ferrer, Building covariance matrices with the desired structure, Chemometr. Intell. Lab. Syst. 127 (2013) 80–88.

[4] J. Camacho, On the generation of random multivariate data, Chemometr. Intell. Lab. Syst. 160 (2017) 40–51.

[5] R. Cook, I. Helland, Z. Su, Envelopes and partial least squares regression, J. Roy. Stat. Soc. B 75 (5) (2013) 851–877.

[6] R.D. Cook, B. Li, F. Chiaromonte, Envelope models for parsimonious and efficient multivariate linear regression, Stat. Sin. (2010) 927–960.

[7] R.D. Cook, X. Zhang, Foundations for envelope models and methods, J. Am. Stat. Assoc. 110 (510) (2015) 599–611.

[8] R.D. Cook, X. Zhang, Simultaneous envelopes for multivariate linear regression, Technometrics 57 (1) (2015) 11–25.

[9] D. Gamerman, H.F. Lopes, Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference, CRC Press, 2006.

[10] L.E. Gangsei, T. Almøy, S. Sæbø, Theoretical evaluation of prediction error in linear regression with a bivariate response variable containing missing data, Commun. Stat. Theor. Meth. 0 (0) (2016) 1–9. https://doi.org/10.1080/03610926.2016.1222434.

[11] R. M. Heiberger, Algorithm as 127: generation of random orthogonal matrices, J. Roy. Stat. Soc. C Appl. Stat. 27 (2) (1978) 199–206.

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

[13] I.S. Helland, T. Almøy, Comparison of prediction methods when only a few components are relevant, J. Am. Stat. Assoc. 89 (426) (1994) 583–591.

[14] I.S. Helland, S. Saebø, H. Tjelmeland, et al., Near optimal prediction from relevant components, Scand. J. Stat. 39 (4) (2012) 695–713.

[15] U.G. Indahl, K.H. Liland, T. Næs, Canonical partial least squares–a unified pls approach to classification and regression problems, J. Chemometr. 23 (9) (2009) 495–504.

[16] M.E. Johnson, Multivariate Statistical Simulation: a Guide to Selecting and Generating Continuous Multivariate Distributions, John Wiley & Sons, 2013.

[17] B.D. Ripley, Stochastic Simulation, vol. 316, John Wiley & Sons, 2009.

[18] S. Sæbø, T. Almøy, A. Flatberg, A.H. Aastveit, H. Martens, Lpls-regression: a method for prediction and classification under the influence of background information on predictor variables, Chemometr. Intell. Lab. Syst. 91 (2) (2008) 121–132.

[19] 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, Chemometrics and Intelligent Laboratory Systems, 2015.

MODEL AND ESTIMATORS FOR PARTIAL LEAST SQUARES REGRESSION

Received: 20 November 2017 | Revised: 29 March 2018 | Accepted: 8 April 2018
DOI: 10.1002/cem.3044

RESEARCH ARTICLE

WILEY CHEMOMETRICS

Model and estimators for partial least squares regression

Inge Svein Helland[1] | Solve Sæbø[2] | Trygve Almøy[2] | Raju Rimal[2]

[1]Department of Mathematics, University of Oslo, Oslo NO-0315, Norway

[2]Norwegian University of Life Sciences, Ås 1430, Norway

Correspondence
Inge Svein Helland, Department of Mathematics, University of Oslo, P.O. Box 1053, Oslo NO-0316, Norway.
Email: ingeh@math.uio.no

Abstract

Partial least squares (PLS) regression has been a very popular method for prediction. The method can in a natural way be connected to a statistical model, which now has been extended and further developed in terms of an envelope model. Concentrating on the univariate case, several estimators of the regression vector in this model are defined, including the ordinary PLS estimator, the maximum likelihood envelope estimator, and a recently proposed Bayes PLS estimator. These are compared with respect to prediction error by systematic simulations. The simulations indicate that Bayes PLS performs well compared with the other methods.

KEYWORDS

Bayes PLS estimator, envelope model, partial least squares, partial least squares model, simulation

1 | INTRODUCTION

Supervised learning from multivariate data is a central problem area in applied statistics and also in chemometrics. Specifically, let our task be to predict a single variable y from a p-dimensional variable $\boldsymbol{x}$, having data on n units. From a statistical point of view, a large number of learning methods are discussed in Hastie et al,[1] mainly under the ordinary multiple regression model. In chemometrics, partial least squares (PLS) regression is the dominating method.

Partial least squares regression has had a vigorous development in the chemometric literature since it was proposed by Wold et al[2] and Martens and Næs.[3] The method has been extended in several directions, and its applications have been expanded to an increasing number of fields, for instance, genomic data.[4] Both these issues have been discussed in detail in a recent paper by Mehmood and Ahmed,[5] where a wealth of further references may be found.

Sometimes, the issue is prediction, but very often, one also see interpretations of scoring, loading, and correlation plots; see, for instance, Martens and Martens.[6] Such plots are not unfamiliar to statisticians in principal component connections, but they are much more used by the chemometric society, and many scientists find them informative. They are plots of the sample variants of the latent variables and parameters defined by (3), (4), and (5) below and, thus, involve consistent estimates of these quantities when $n \to \infty$ and probably also in the more general case $p/n \to 0$.

In the beginning, the PLS method was to some extent neglected or turned down by statisticians (an exception among others was Frank and Friedman[7]; see also Helland[8,9]), but it is now included as a tool among other biased regression methods by applied statisticians. For a general discussion paper with contributions both from mathematical statisticians and chemometricians, see Sundberg.[10]

Indeed, there was a difference in culture between chemometricians and statisticians then, and this difference still exists to a large extent. A statement by Munck et al[11] illustrates this, as seen from one side: "If chemometrics in its historical development had been limited to follow current scientific (and statistical) theories there would have been minimal progress in its wide applications today."

Recently, the difference in culture was discussed in some detail by Martens.[12] On the one hand, Martens makes the point that the field of Chemometrics has a lot to learn from other disciplines—mathematics, statistics, and computer science.

Among other things, he says that it will not be enough to have efficient "black box" algorithms. On the other hand, he accuses statisticians in general for a predilection for "macho mathematics," concluding in part that Chemometrics need more statistics but not more statisticians. In other parts of the paper, he talks about bridging the gap between the 2 disciplines, an effort that we whole heartedly support.

This difference in culture may in part be related to the concepts of creativity and rigor, qualities which to some extent may be called complementary. One could say that one culture puts more emphasis on creativity, the other on rigor. Of course, this is a huge simplification. First, there is a lot of creativity among statisticians, also mathematical statisticians. Secondly, one should emphasize that precise thinking also should influence practice. A case of point is the following: Chung and Keleş[13] recently proved that the PLS regression vector is inconsistent when $p/n \to k > 0$ under a wide set of conditions. This result is probably not too well known among chemometricians; some may have a tendency to put much confidence in PLS regression when $p \sim n$ or $p > n$. It is to be emphasized that the inconsistency result in Chung and Keleş[13] is only concerned with estimation of the regression vector. The mathematical properties of PLS as a "prediction" method when $p > n$ are largely unknown, from a statistical point of view. There is much positive empirical evidence among applied researchers on these properties, but statisticians have only started to attack this problem, since it from an analytic point of view is very difficult. In particular, see the very recent paper by Cook and Forzani,[14] where asymptotic expansions allowing both n and p to be large are developed for PLS prediction with 1 component.

It is true that chemometricians have had a leading edge in the development of PLS and of certain multivariate methods, in particular, with respect to visualization etc, and they still are ahead of statisticians in this sense.

Accepting this, an important general question is what mathematical statisticians can contribute with in this development. There are relatively few papers by mathematical statisticians investigating statistical properties of the PLS regression method itself. There are however several investigations on the shrinkage properties of PLS; see Krämer[15] and references there, and also Foschi[16] with references. Garthwaite[17] offered a simple interpretation of PLS. Stone and Brooks[18] and Naik and Tsai[19] discuss different generalizations of PLS; in the latter paper also, consistency of PLS is proved. In Stoica and Söderström,[20] an asymptotic formulae related to PLS is derived. Chun and Keleş[21] extends consistency to the case $p/n \to 0$, introduces a sparse PLS algorithm, and compares methods by simulation. In Krämer and Sugiyama,[22] the degrees of freedom of PLS regression is discussed, and this concept is used in model selection. See also references in this last paper.

In Helland and Almøy,[23] several predictors in the random x regression model were compared asymptotically as $n \to \infty$, including principal component regression (PCR) and sample PLS regression (see the next section). The conclusion was that PCR is best for very large irrelevant eigenvalues (excluded from the prediction equation), whereas PLS regression tends to be best for intermediate irrelevant eigenvalues. Because the difference is extremely small for small irrelevant eigenvalues, and because very large irrelevant eigenvalues seldom occur in practice (and if they do, they should be included in the prediction equation), it was concluded that PLS regression is the method of choice in many cases. An additional argument for PLS over PCR is that PLS involves only choosing the number of components, whereas PCR also entails deciding which of the components should be included in the prediction.

As already mentioned, Cook and Forzani[14] give an asympotic expansion of the prediction error in PLS regression, which also is informative when $p \to \infty$, but mainly limited to 1 component. Results with several components are also announced.

A vital aspect in the history of statistics is the interplay between model and estimators. Once a model is formulated, one can in principle think of several estimators in this model. A statistician will talk about a "hard" model in terms of probability distributions—at least in terms of a model equation and a statement of correlation between terms in this model. This is a concept that has had and has a great success in a number of disciplines and is at the very core of statistics as a science. Our goal in the present paper is to show that this concept can be applied—and is useful—also in connection to PLS. Specifically, our purposes are to

- stress that PLS as an algorithm can be connected to a unique statistical model (known since 1990);
- formulate 5 different ways to present this model (known in the statistical literature since 2013);
- argue that the simplest way to present the model is through the concept of relevant components—a reduction of the random x regression model;
- review briefly some statistical investigations related to PLS;
- ask if the PLS algorithm may be improved by modifying the weights;
- argue that once the model is presented, the comparison of different estimators in the model is relevant;
- present a systematic tool (`simrel`) for comparing estimators in the model with relevant components;
- present the maximum likelihood estimator in the model;

- present a Bayes estimator connected to the model;
- and compare the PLS algorithm, the maximum likelihood estimator, and the Bayes estimator in a systematic simulation study, mainly with near collinear data.

Thus, in the PLS model, one can certainly discuss other estimators than the usual PLS regression estimator, which can be seen as originating by replacing population (co)variances in the model by sample (co)variances. Two examples are the maximum likelihood estimator of Cook et al,[24] see also Cook et al[25,26] and Cook and Zhang,[27] and the Bayesian estimator of Helland et al.[28] By simulation, both these estimators have performed well compared with PLS regression under certain conditions, but they have their disadvantages. The maximum likelihood estimator cannot be used in the case when the data matrix has rank less than p, and the Bayesian estimator requires heavy computations, in particular, when p is large.

To compare estimators, we make vital use of the recently developed simulation package `simrel`; see Sæø et al.[29] It is very important to have such a tool in an area where it is difficult to obtain results by purely analytical means.

We emphasize that this paper is based upon reduction of the *random* $\boldsymbol{x}$ regression model. When considering latent variables from PLS, and when considering near collinearity in the observed $\boldsymbol{x}$-variables, it is natural to treat these $\boldsymbol{x}$-variables as random. It is our philosophy that this is also the best way to look upon model reduction. On the other hand, in the context of prediction, one could argue that one should condition upon the $\boldsymbol{x}$-variables and consider them as fixed. A prominent paper on PLS regression, taking fixed $\boldsymbol{x}$-variables in the basic model, is Krämer and Sugiyama,[22] where further references can be found.

In recent years, there has been a rapidly growing statistical literature on the envelope model—a model generalizing the PLS model. In addition to the maximum likelihood estimation paper mentioned above, the most important papers seem to be Cook and Zhang,[30] where simultaneous reduction in the $\boldsymbol{x}$- and $\boldsymbol{y}$-space is proposed, and Cook and Zhang,[31] where extensions to other regression methods than linear regression are discussed. More references can be found in these papers.

Model reduction in regression models is discussed in general from the point of view of rotations in the x-space in Helland.[32]

The plan or this paper is as follows: In Section 2, we formulate the model in 5 different ways, which can be shown to be equivalent. In Section 3, we define 4 different estimators in the model, including the recent Bayes PLS estimator of Helland et al.[28] In Section 4, we ask the question if the ordinary PLS estimator with m components can be improved by forcing the weight vector at step $m + 1$ to vanish; the answer turns out to be negative. In Section 5, we describe the simulations done for comparison of estimators with respect to prediction error, and in Section 6, we give the results of the simulations. In Section 7, we illustrate the methods on a real data set. Finally, Section 8 is a discussion section.

2 | THE MODEL: SEVERAL FORMULATIONS

Take as a point of departure the linear model

$$y = \mu_y + \boldsymbol{\beta}'(\boldsymbol{x} - \boldsymbol{\mu}_x) + \epsilon, \tag{1}$$

where $\boldsymbol{\beta}$ and $\boldsymbol{x}$ are p-dimensional and the random predictor $\boldsymbol{x}$ has mean μ_x and covariance matrix Σ_{xx}, for simplicity, assumed nonsingular here (this can be relaxed to assuming $\boldsymbol{\beta} \in \text{span}(\Sigma_{xx})$ in the case where this matrix is singular; see Cook et al,[24] and also C below). Independently, ϵ is distributed with mean 0 and variance σ^2. When doing prediction from this model for near collinear data, a model reduction may be called for. Throughout this paper, a definite m-dimensional model reduction, which may be formalized in several equivalent ways, will be used. When this model holds, we say that we have an envelope model or a PLS model of dimension m or that there are m relevant components for prediction in the model.

A. Given a subspace $\mathcal{S}$ of R^p, let $\boldsymbol{P}_{\mathcal{S}}$ be the projection upon $\mathcal{S}$, and let $\boldsymbol{Q}_{\mathcal{S}}$ be the projection orthogonal to $\mathcal{S}$. For simplicity, discuss the case where $\mu_x = \boldsymbol{0}$. Let now $\mathcal{S}$ be the smallest space such that (1) $\boldsymbol{Q}_{\mathcal{S}}\boldsymbol{x}$ is uncorrelated with $\boldsymbol{P}_{\mathcal{S}}\boldsymbol{x}$ and (2) y is uncorrelated with $\boldsymbol{Q}_{\mathcal{S}}\boldsymbol{x}$ given $\boldsymbol{P}_{\mathcal{S}}\boldsymbol{x}$. In this case, we may say that $\boldsymbol{Q}_{\mathcal{S}}\boldsymbol{x}$ contains no linear information about y, neither directly nor through $\boldsymbol{P}_{\mathcal{S}}\boldsymbol{x}$. Consider a reduction of the data to $\boldsymbol{P}_{\mathcal{S}}\boldsymbol{x}$.

B. Here is an algebraic characterization, which turns out to be equivalent. For a matrix $\boldsymbol{M}$, define $\boldsymbol{M}\mathcal{S}$ as the space of vectors $\boldsymbol{Mz}$, as $\boldsymbol{z}$ runs through $\mathcal{S}$, and let $\mathcal{S}^{\perp}$ be the space perpendicular to $\mathcal{S}$. Let now $\mathcal{S}$ be the smallest space in R^p such that (1) both $\Sigma_{xx}\mathcal{S} \subseteq \mathcal{S}$ and $\Sigma_{xx}\mathcal{S}^{\perp} \subseteq \mathcal{S}^{\perp}$ and (2) $\text{span}(\boldsymbol{\beta}) \subseteq \mathcal{S}$. In this case, we say that $\mathcal{S}$ is the Σ_{xx}-envelope of $\text{span}(\boldsymbol{\beta})$. It can be shown Cook et al[33] that the envelope always exists as the smallest space with the stated properties.

C. The regression vector $\boldsymbol{\beta}$ can always be expanded in terms of the eigenvectors $\boldsymbol{d}_i$ of Σ_{xx}:

$$\boldsymbol{\beta} = \sum_{i=1}^{p} \gamma_i \boldsymbol{d}_i. \tag{2}$$

In general, when there are coinciding eigenvalues in Σ_{xx}, this expansion is not unique. However, assume that this sum can be reduced to exactly m nonzero terms: $\boldsymbol{\beta} = \sum_{i=1}^{m} \gamma_i \boldsymbol{d}_i$, where the $\boldsymbol{d}_i$ correspond to different eigenvalues of Σ_{xx}. We then say that there are m relevant components for prediction in the model. This reduction can be imagined to take place by 2 mechanisms: (1) Some of the γ_i's are really zero, and (2) there are coinciding eigenvalues in Σ_{xx}. Then, one can rotate such that it is enough with 1 eigenvector for each eigenspace in the sum. In this approach, it is important that we only know that there are m nonzero terms in the sum, not which terms that are nonzero. For a closer discussion of this, see Næs and Helland[34] and Helland and Almøy.[23]

D. Consider the population version of the well-known PLS algorithm: Take $\boldsymbol{e}_0 = \boldsymbol{x} - \mu_x, f_0 = y - \mu_y$, and for $a = 1, 2, \ldots, m$ compute successively:

$$\boldsymbol{w}_a = \mathrm{cov}(\boldsymbol{e}_{a-1}, f_{a-1}), \quad t_a = \boldsymbol{w}_a' \boldsymbol{e}_{a-1}, \tag{3}$$

$$\boldsymbol{p}_a = \mathrm{cov}(\boldsymbol{e}_{a-1}, t_a)/\mathrm{var}(t_a), \; q_a = \mathrm{cov}(f_{a-1}, t_a)/\mathrm{var}(t_a), \tag{4}$$

$$\boldsymbol{e}_a = \boldsymbol{e}_{a-1} - \boldsymbol{p}_a t_a, \quad f_a = f_{a-1} - q_a t_a.$$

It can be proved[9] and is important in this connection that under the reduced model C, this algorithm stops automatically after m steps when $m < p$: It stops because $\boldsymbol{w}_{m+1} = \mathrm{cov}(\boldsymbol{e}_m, f_m) = 0$. After those m steps, we get the representations

$$\boldsymbol{x} = \boldsymbol{\mu}_x + \boldsymbol{p}_1 t_1 + \ldots + \boldsymbol{p}_m t_m + \boldsymbol{e}_m, \; y = \mu_y + q_1 t_1 + \ldots + q_m t_m + f_m \tag{5}$$

with the corresponding PLS population prediction

$$y_{m,PLS} = \mu_y + q_1 t_1 + \ldots + q_m t_m = \mu_y + \boldsymbol{\beta}_{m,PLS}'(\boldsymbol{x} - \boldsymbol{\mu}_x).$$

Theorem 1. (Helland[9] and Cook et al[24])

(a) The 2 conditions A and B on the space S are equivalent.
(b) The models formulated by C and D are equivalent.
(c) When there are m relevant components for prediction, the envelope space S has dimension m, and S can be taken as span $(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_m) = span\ (\boldsymbol{d}_1, \cdots, \boldsymbol{d}_m)$.
(d) When the envelope space has dimension m, there are m relevant components for prediction.
(e) In this case, we have $\boldsymbol{\beta}_{m,PLS} = \boldsymbol{\beta}$.

Proof. (a) is proved in Cook et al,[24, Proposition 1] and (b) in Helland.[9, Theorem 2] Finally, (c)-(e) and the equivalence with E below are contained in Cook et al.[24, Proposition 5] □

In this sense, all the model formulations (A-D) are equivalent; they describe the same reduced model. In Helland[9] and Cook et al,[24] a fifth equivalent formulation in terms of a Krylov sequence is also given:

E. S is also spanned by the vectors $\boldsymbol{\sigma}_{xy}, \Sigma_{xx}\boldsymbol{\sigma}_{xy}, \cdots, \Sigma_{xx}^{m-1}\sigma_{xy}$, and m is the smallest integer such that $\boldsymbol{\beta} = \Sigma_{xx}^{-1}\boldsymbol{\sigma}_{xy}$ belongs to S.

The simplest way to express the model reduction implied by PLS seems to be C. In analogy with the exivalence between A and B, this can also be expressed as a reduction of the x vector. Consider again the centered case $\mu_x = \boldsymbol{0}$. For details, see Næs.[34]

C'. Let $\boldsymbol{R}$ be a nonrandom pxm matrix of rank m. Normalize such that $\boldsymbol{R}'\boldsymbol{R} = \boldsymbol{I}$. There are m relevant components $\boldsymbol{R}'\boldsymbol{x}$ for predicting y if and only if $\boldsymbol{R}$ can be found such that (a) $\boldsymbol{\beta} \in \mathrm{span}(\boldsymbol{R})$ and (b) $\mathrm{span}(\boldsymbol{R})$ is spanned by eigenvectors of Σ_{xx}.

Being a reduced model that can be motivated in so many different ways, it is definitively of interest to find a good estimator of the regression vector $\boldsymbol{\beta}$ under this model.

3 | ESTIMATORS IN THE PLS/ENVELOPE MODEL

Now that the PLS model is introduced, we will start to look at estimators of the parameters in this model, in particular, estimators of $\boldsymbol{\beta}$, which will give prediction. Of special interest is estimators that perform well in the case of near collinear data. Some estimators are already known from the literature.

- **a.** The ordinary PLS estimator can be introduced as follows: With data $(\boldsymbol{X},\boldsymbol{y})$, take initial values $\boldsymbol{E}_0=\boldsymbol{X}-\bar{\boldsymbol{x}}\boldsymbol{1}'$ and $\boldsymbol{f}_0=\boldsymbol{y}-\bar{y}\boldsymbol{1}$. Run the population PLS algorithm for A steps with population (co)variances replaced by sample (co)variances. Ordinarily, A is found by cross-validation or by similar means. Note that from D in Section 2, the m-step PLS model is characterized by $\boldsymbol{w}_{m+1}=\text{cov}(\boldsymbol{e}_m,f_m)=\boldsymbol{0}$. Theoretically, when $A=m$, we cannot expect the sample weights $\widehat{\boldsymbol{w}}_{m+1}$ to be zero. However, since any continuous function of the sample covariances and variances is consistent for the same function of the population covariances and variances, since $\widehat{\boldsymbol{w}}_{m+1}$ through the PLS algorithm is such a function and since $\boldsymbol{w}_{m+1}=\boldsymbol{0}$, we will have $\lim_{n\to\infty}\widehat{\boldsymbol{w}}_{m+1}=\boldsymbol{0}$ almost surely.
- **b.** The sparse regression SPLS of Chun and Keleş[21]: This requires 2 effective tuning parameters, and it also aims at variable selection. Sparse partial least squares (SPLS) seems to be better than ordinary PLS in certain cases, also when variable selection is not an issue.
- **c.** When $\boldsymbol{S}=(\boldsymbol{X}-\bar{\boldsymbol{x}}\boldsymbol{1}')'(\boldsymbol{X}-\bar{\boldsymbol{x}}\boldsymbol{1}')$ has rank p, which specifically requires $n>p$, the maximum likelihood estimator of $\boldsymbol{\beta}$ under the multinormal envelope model was given in Cook et al.[24] This estimator is of course very useful, but it cannot be used for small n. Modifications of the maximum likelihood estimator, which cover also this case, were recently indicated in Cook et al.[25] That paper also gives a MATLAB toolbox for maximum likelihood estimation in the envelope model and in several generalizations of this model. A faster algorithm for maximum likelihood estimation is discussed in Cook and Zhang.[27] Even faster algorithms with modifications to small sample size $n<p$ are recently described in Cook and Zhang,[35] and an R-package was recently described by Cook et al.[26]
- **d.** Under a specific rotation-invariant prior, the Bayes estimator of $\boldsymbol{\beta}$ under the model with m relevant components was given in Helland et al.[28] This estimator was shown to be close to the best equivariant estimator, but it requires heavy computation.
  The estimation was performed by a Markov Chain Monte Carlo approach. Specifically, for given m, and for observed centered data $\boldsymbol{y}$ and $\boldsymbol{X}$, the likelihood function is proportional to

$$f(\boldsymbol{y},\boldsymbol{X}|\boldsymbol{\nu},\boldsymbol{\gamma},\boldsymbol{D},\sigma^2)\propto(\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\left(\boldsymbol{y}-\boldsymbol{X}\sum_{i=1}^{m}\gamma_i\boldsymbol{d}_i\right)'\left(\boldsymbol{y}-\boldsymbol{X}\sum_{i=1}^{m}\gamma_i\boldsymbol{d}_i\right)\right)\\ \times\left(\prod_{i=1}^{p}\nu_i\right)^{-n/2}\prod_{j=1}^{n}\exp\left(-\frac{1}{2}\boldsymbol{x}_j'\left(\sum_{i=1}^{p}\frac{1}{\nu_i}\boldsymbol{d}_i\boldsymbol{d}_i'\right)\boldsymbol{x}_j\right), \tag{6}$$

  where $\boldsymbol{\nu}=[\nu_1,\ldots,\nu_p]$ and $\boldsymbol{D}=[\boldsymbol{d}_1,\ldots,\boldsymbol{d}_p]$ are the eigenvalues and the eigenvectors of the $\boldsymbol{x}$-covariance matrix Σ_{xx} and $\boldsymbol{\gamma}=[\gamma_1,\ldots,\gamma_m]$ are regression parameters of the PLS model.
  As argued in Helland et al,[28] a near optimal equivariant regressor is found as the Bayesian estimator under rotation-invariant prior for $\boldsymbol{d}_1,\ldots,\boldsymbol{d}_p$ and prior $\pi(\boldsymbol{\gamma})=\prod_i 1/\gamma_i^{1-\epsilon}$, where $1/\epsilon$ is a large uneven integer. Slightly modified scale priors are also chosen for $\boldsymbol{\nu}$ as $\pi(\boldsymbol{\nu})=\prod_i 1/\nu_i\exp(-\epsilon_\nu/2\nu_i)$ and for σ^2 as $\pi(\sigma^2)=1/\sigma^2\exp(-\epsilon_\sigma/2\sigma^2)$. Here, ϵ_ν and ϵ_σ are some small numbers chosen to ensure properness of the posterior distribution. Estimation of model parameters may be done by means of Markov chain Monte Carlo methods. As shown in Helland et al,[28] the marginal posterior distributions for σ^2 and ν_i (for $i=1,\ldots,p$) are, for the given prior distributions, all inverse gamma distributions. Furthermore, the marginal posterior distributions for γ_i (for $i\in 1,\ldots,m$) are approximately normally distributed. There is no closed form posterior distribution for $\boldsymbol{D}$, hence a random walk step with a Metropolis-Hastings acceptance step is necessary for the sampling from the posterior distributions of the parameters. R-code for the Bayes estimator is available at http://www.github.com/solvsa/BayesPLS, and further details on the Markov chain Monte Carlo implementation may be found in the supplementary documentation to Helland et al.[28]

By simulation, both the maximum likelihood estimator c and the Bayes estimator d were shown to perform well compared to the PLS estimator a. These 2 estimators assume a multinormal distribution of the data in their derivation, but the estimators themselves are valid under more general assumptions. Both the chemometric tradition and the envelope model of Cook et al[24,33] demand no detailed distributional assumptions.

4 | CAN A BETTER ESTIMATOR BE FOUND BY SIMPLE MEANS?

The m step PLS model is characterized by the constraint $\boldsymbol{w}_{m+1} = \text{cov}(\boldsymbol{e}_m, f_m) = \boldsymbol{0}$. However, in the sample PLS algorithm, $\widehat{\boldsymbol{w}}_{m+1}$ is a continuous random variable if the data are continuous. Hence, almost surely, $\widehat{\boldsymbol{w}}_{m+1} \neq \boldsymbol{w}_{m+1} = \boldsymbol{0}$. This means that the estimator of the vector of PLS parameter falls outside the corresponding parameter space. On the other hand, by standard statistical theory, the maximum likelihood estimator and any Bayes estimator are always in the parameter space.

In this section, we ask the question whether we can improve the PLS algorithm in some way such that $\widehat{\boldsymbol{w}}_{m+1} = \boldsymbol{0}$ for the improved algorithm. That is, we seek modified weights $\widehat{\boldsymbol{w}}_1, \cdots, \widehat{\boldsymbol{w}}_m$ such that $\widehat{\boldsymbol{w}}_{m+1} = \boldsymbol{0}$ in the modified algorithm. Unfortunately, the answer to this question is no. This programme is only possible when $\boldsymbol{S}$ is invertible, and then it by necessity leads to the least squares solution. Let $\widehat{\boldsymbol{W}}_A = (\widehat{\boldsymbol{w}}_1, \cdots, \widehat{\boldsymbol{w}}_A)$ for any A.

First, we need some properties of the ordinary PLS algorithm.

Proposition 2. *At each step, the PLS weights satisfy*

$$\widehat{\boldsymbol{w}}_{A+1} = \boldsymbol{s} - \boldsymbol{S}\widehat{\boldsymbol{W}}_A(\widehat{\boldsymbol{W}}_A'\boldsymbol{S}\widehat{\boldsymbol{W}}_A)^{-1}\widehat{\boldsymbol{W}}_A'\boldsymbol{s}, \tag{7}$$

and the A step regression vector is

$$\widehat{\boldsymbol{\beta}}_A = \widehat{\boldsymbol{W}}_A(\widehat{\boldsymbol{W}}_A'\boldsymbol{S}\widehat{\boldsymbol{W}}_A)^{-1}\widehat{\boldsymbol{W}}_A'\boldsymbol{s}. \tag{8}$$

Proof. These relations were proved in Helland,[8] see equations (3.3) and (3.7) there, and were also used in Cook et al.[24] □

Now, fix m. To find an algorithm such that $\widehat{\boldsymbol{w}}_{m+1} = 0$, we will have to modify the weights $\widehat{\boldsymbol{w}}_1, \cdots, \widehat{\boldsymbol{w}}_m$.

Definition 1. For the purpose of this section, call a restricted PLS prediction any prediction method based on an estimator of $\boldsymbol{\beta}$ of the form (8) for $A = m$ such that

1.) $\widehat{\boldsymbol{W}}_m = (\widehat{\boldsymbol{w}}_1, \cdots, \widehat{\boldsymbol{w}}_m)$ is modified with respect to PLS in some way.
2.) Equation 7 holds for $A = m$ and gives $\widehat{\boldsymbol{w}}_{m+1} = \boldsymbol{0}$.

Theorem 3. *An RPLS prediction method exists if and only if $\boldsymbol{S}$ is invertible and $\boldsymbol{S}^{-1}\boldsymbol{s} \in span\,\widehat{\boldsymbol{W}}_m$. In that case, $\widehat{\boldsymbol{\beta}}$ is equal to the least squares estimator $\boldsymbol{S}^{-1}\boldsymbol{s}$.*

Proof. Assume that (7) holds for $A = m$ and $\widehat{\boldsymbol{w}}_{m+1} = \boldsymbol{0}$. Then, $\boldsymbol{s} = \boldsymbol{S}\widehat{\boldsymbol{W}}_m(\widehat{\boldsymbol{W}}_m'\boldsymbol{S}\widehat{\boldsymbol{W}}_m)^{-1}\widehat{\boldsymbol{W}}_m'\boldsymbol{s}$. This is possible for general $\boldsymbol{s}$ only if $\boldsymbol{S}$ is nonsingular, and then it is equivalent to $\boldsymbol{R}\sqrt{\boldsymbol{S}}^{-1}\boldsymbol{s} = \sqrt{\boldsymbol{S}}^{-1}\boldsymbol{s}$ with $\boldsymbol{R} = \boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'$, where $\boldsymbol{A} = \sqrt{\boldsymbol{S}}\widehat{\boldsymbol{W}}_m$. Since $\boldsymbol{R}$ is the projector upon span($\boldsymbol{A}$), this is again equivalent to $\sqrt{\boldsymbol{S}}^{-1}\boldsymbol{s} \in \text{span}\,(\sqrt{\boldsymbol{S}}\widehat{\boldsymbol{W}}_m)$, or $\boldsymbol{S}^{-1}\boldsymbol{s} \in \text{span}\,(\widehat{\boldsymbol{W}}_m)$. Then, putting $\boldsymbol{s} = \boldsymbol{S}\widehat{\boldsymbol{W}}_m\boldsymbol{q}$ in (8) for some $\boldsymbol{q}$ gives $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{W}}_m\boldsymbol{q} = \boldsymbol{S}^{-1}\boldsymbol{s}$. □

Thus, Theorem 3 shows clearly that it is not possible to modify the PLS weights in a nontrivial way such that the modified estimator belongs to the parameter space.

5 | DATA SIMULATIONS FOR COMPARISON OF ESTIMATORS

A comparative study of the prediction performances of the regular PLS algorithm, the maximum likelihood envelope method, the Bayes PLS, and the method of ordinary least squares (OLS) was performed on data simulated from the random regression model (1) and a real dataset measuring various properties and near infrared (NIR) spectra of diesel fuels. This and the following section will focus on simulation study in detail. In the study, we consider envelope method for predictor reduction[24] and use R-code discussed in Cook et al.[26] A detailed description of the simulation procedure can be found in Sæbø et al[29] with the accompanying R-package `simrel`, but key features of the approach are presented next. The simulation set up is best explained from reexpressing model (1) in the Gaussian case as

$$\begin{bmatrix} y \\ \mathbf{x} \end{bmatrix} \sim \mathcal{N}\left(\boldsymbol{\mu}_{yx}, \Sigma_{yx}\right) = \mathcal{N}\left(\begin{bmatrix} \mu_y \\ \boldsymbol{\mu}_x \end{bmatrix}, \begin{bmatrix} \sigma_y^2 & \boldsymbol{\sigma}_{xy}^t \\ \boldsymbol{\sigma}_{xy} & \Sigma_{xx} \end{bmatrix}\right), \tag{9}$$

where $\boldsymbol{\sigma}_{xy}$ is a vector holding the covariances between the predictors ($\mathbf{x}$) and the response (y). The vector of regression coefficients $\boldsymbol{\beta}$ is by standard theory given as $\boldsymbol{\beta} = \Sigma_{xx}^{-1}\boldsymbol{\sigma}_{yx}$, which in turn can be expressed in terms of the eigenvalues $\nu_1, \ldots, \nu_p$ and the eigenvectors $\boldsymbol{d}_1, \ldots, \boldsymbol{d}_p$ of Σ_{xx}:

$$\boldsymbol{\beta} = \sum_{i=1}^{p} \frac{\boldsymbol{d}_i' \boldsymbol{\sigma}_{yx}}{\nu_i} \boldsymbol{d}_i = \sum_{i=1}^{p} \gamma_i \boldsymbol{d}_i, \tag{10}$$

as given in Equation 2. In `simrel`, the following simplifying assumptions are made:

- ▷ It is assumed that $\nu_i = e^{-\eta(i-1)}$ for $i = 1, \ldots, p$, implying $\nu_1 = 1$ (which we may assume without loss of generality) and that all subsequent eigenvalues are decaying according to the size of the parameter η. A large η gives a rapid decrease in eigenvalues, implying high level of multicollinearity in $\boldsymbol{x}$.
- ▷ It is assumed that $m \leq p$ eigenvectors are relevant for y, which means that Equation 10 (potentially) reduces to

$$\boldsymbol{\beta} = \sum_{i \in \mathcal{P}} \gamma_i \boldsymbol{d}_i, \tag{11}$$

  where m-vector $\mathcal{P}$ is the set of indices of the relevant components (relpos) for which $\gamma_i = \boldsymbol{d}_i{}' \boldsymbol{\sigma}_{xy}/\nu_i \neq 0$. Hence, the envelope or the relevant space has dimension m (see Theorem 1).
- ▷ Without loss of generality, it is further assumed that $\sigma_y = 1$, $\mu_y = 0$ and $\mu_x = \mathbf{0}$.

In `simrel`, the actual values of $\boldsymbol{\sigma}_{xy}$ were set to satisfy a prespecified value of the population coefficient of determination ρ^2. It may be shown that under the above assumptions, $\rho^2 = \boldsymbol{\sigma}_{xy}' \Sigma_{xx}^{-1} \boldsymbol{\sigma}_{xy}$. This completes the specification of the parameters used in `simrel`, and in the present comparison study, a design for the simulated data sets in terms of these parameters were as defined in Table 1.

From the possible combination of the above parameters, 32 calibration sets were simulated with 5 replications of each, ie, there were 160 calibration sets (`datasets`) altogether.

6 | SYSTEMATIC COMPARISONS

A systematic comparison of the methods across the simulation designs was made on the basis of their ability to predict test samples. Since the distribution of the simulated variables is fully known, the expected mean squared error of prediction (MSEP) based on some $\hat{\boldsymbol{\beta}}$ estimated from a calibration set may be found as

$$E_x\left[E_y(y - \hat{y})^2\right] = \left[\sigma^2 + E\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^t \Sigma_{xx} \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\right] \frac{n+1}{n} \tag{12}$$

in the model. The expectation on the right-hand side of the above expression is estimated for each method and for each design as an average over the 5 replicated calibration sets. To study the effects of `p`, ρ^2, `relpos`($\mathcal{P}$), `Method`, and (η) along with their interactions, we first retrieved the minimum MSEP for each method across 1 to 10 components (assumed numbers of relevant components). In Figure 1, interaction plots for these data properties are displayed.

The effect of the third-order interaction between p, ρ^2 and `Methods`, which we see in Figure 1 (left), shows that the maximum likelihood–based estimation methods, in our case, the envelope and the OLS, perform poorly on data sets with large number of variables and low ρ^2. Still, the performance of the envelope is better than OLS also in situations where $p = 40$ and $n = 50$, representing here $p \sim n$. The interaction plots suggest that the Bayes PLS and ordinary PLS estimation methods are better and more stable on average than the two other methods.

Similarly, the effect of third-order interaction between `relpos`, η, and `Method` in Figure 1 (right) shows that OLS method gives higher prediction error than other methods, but the effect of `relpos` is small but notable for the envelope method. Again, Bayes PLS and ordinary PLS are best.

TABLE 1 Parameters used for simulating calibration sets

Number of training samples	$\boldsymbol{n}$	**50**
Number of predictor variables	p	15 and 40
Population coefficient of determination	ρ^2	0.5 and 0.9
Position of relevant components	$\mathcal{P}$	▷ 1, 2 ▷ 1, 3 ▷ 2, 3 and ▷ 1, 2, 3
Decay factor of eigenvalues of Σ_{xx}	η	0.5 and 0.9

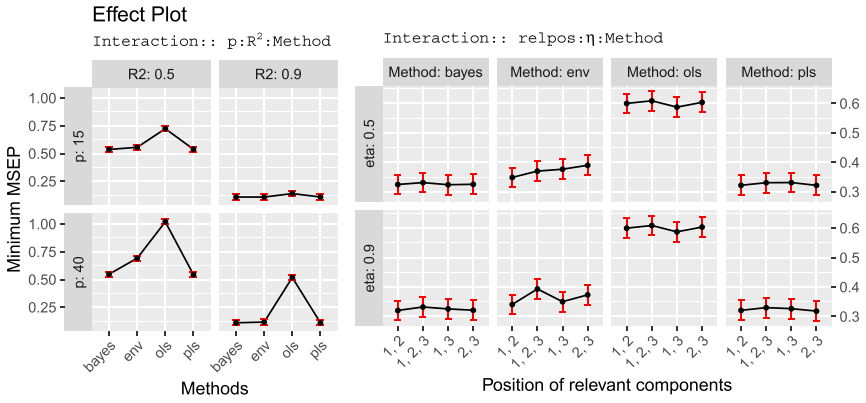


FIGURE 1 Third-order interaction effects. MSEP, mean squared error of prediction; ENV, envelope; OLS, ordinary least squares; PLS, partial least squares

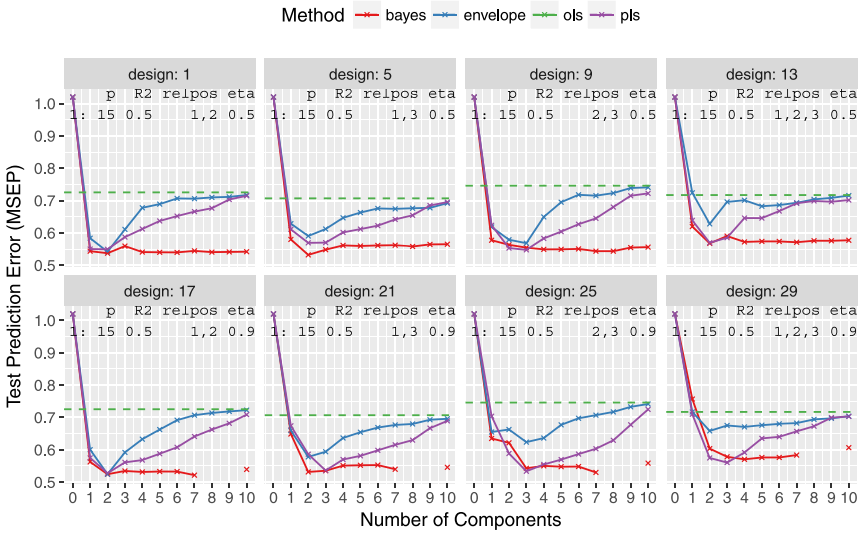


FIGURE 2 Average prediction error for designs with 15 predictor variables where coefficient of determination is 0.5. MSEP, mean squared error of prediction

The prediction error plots below are organized into 4 groups: (*a*) $p = 15$, $\rho^2 = 0.5$; (*b*) $p = 15$, $\rho^2 = 0.9$; (*c*) $p = 40$, $\rho^2 = 0.5$; and (*d*) $p = 40$, $\rho^2 = 0.9$. The OLS prediction error is shown by a straight dotted line.

In group (*a*), with small number of variables ($p \ll n$) and noisy data ($\rho^2 = 0.5$), Figure 2 shows that all the estimation methods performed better than OLS for all designs in this group, Bayes PLS being best in nearly all cases. Some convergence problems with Bayes PLS when eigenvalues decrease rapidly can be ignored since the minimum MSEP is already obtained from fewer components.

Having few variables rich with information ($\rho^2 = 0.9$), the designs in group (*b*) (Figure 3) leads to easy prediction with low prediction error in general for all methods. All the methods including OLS have small MSEPs, but the other methods are still dominant. In most of the situations, Bayes PLS has reached minimum error with only 1 component. In this group, the performance of envelope is better than regular PLS, and the minimum error for envelope is also achieved with fewer components.

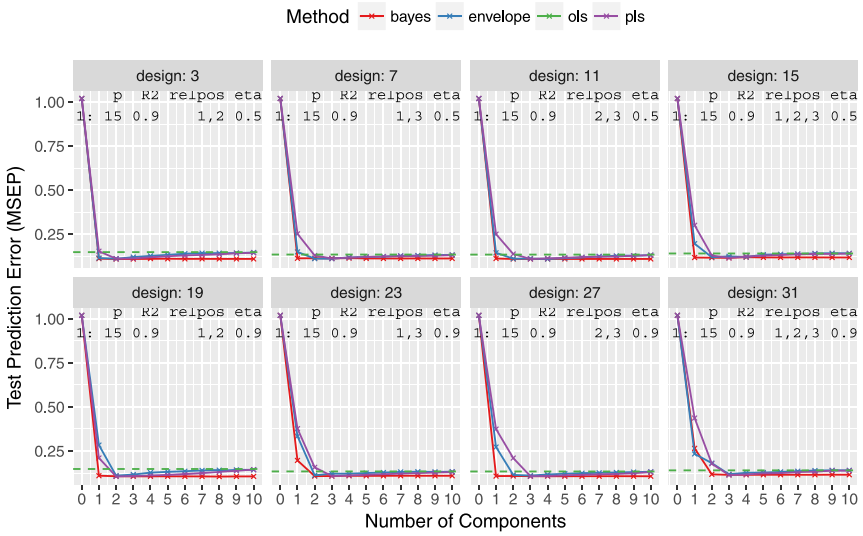


FIGURE 3 Average prediction error for designs with 15 predictor variables where coefficient of determination is 0.9. MSEP, mean squared error of prediction

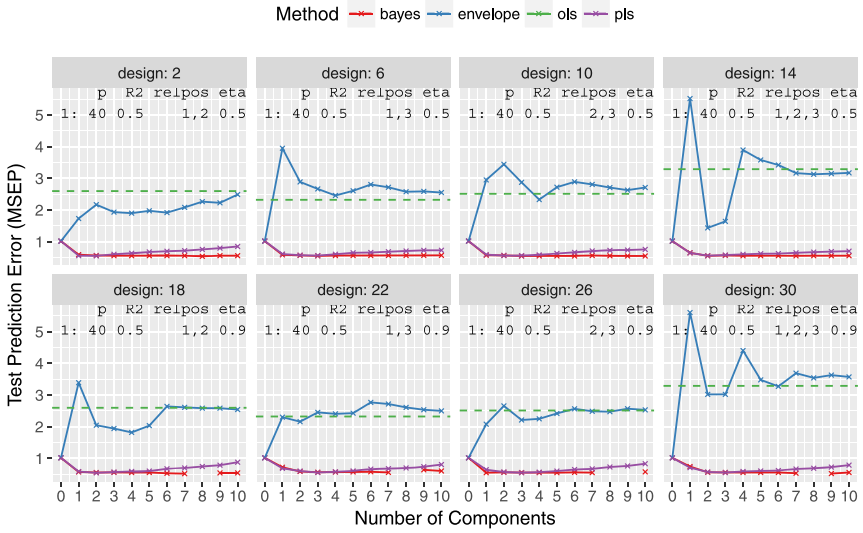


FIGURE 4 Average prediction error for designs with 40 predictor variables where coefficient of determination is 0.5. MSEP, mean squared error of prediction

Low information content combined with many predictor variables characterize the designs in group (*c*), and prediction is in general difficult for these designs. In Figure 4, the methods based on maximum likelihood estimation performed poorly and often poorer than an average guess. Bayes PLS and regular PLS performed well, as in the previous designs.

With 40 predictors ($p \sim n$) and rich information (high ρ^2) (designs in group *d*), Figure 5 shows that in most of the situations (except in design 16), the envelope method has nearly attained true minimum error (0.1) and has outperformed OLS. However, its prediction error is still larger than Bayes PLS and PLS. Bayes PLS and PLS methods are highly stable

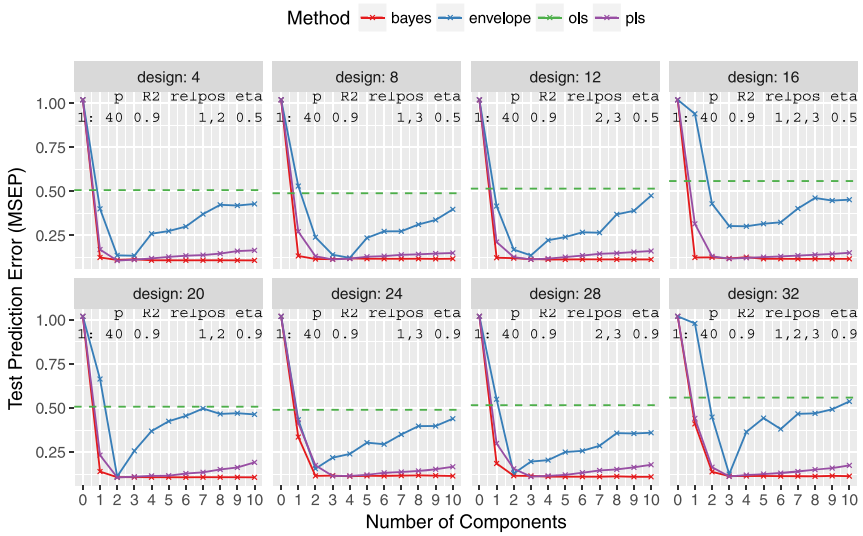


FIGURE 5 Average prediction error for designs with 40 predictor variables where coefficient of determination is 0.9. MSEP, mean squared error of prediction

and are closer to true minimum error. Further, Bayes PLS is able to obtain its minimum prediction error with only a small number of components.

In general, ordinary PLS is very stable in all situations. It is extensible (lots of variants has been developed after its introduction), easy, and less time consuming to fit than Bayes PLS and the envelope method. If the issue is to get closer prediction from squeezing information as much as possible, Bayes PLS will be a good alternative. Its performance with varying number of components is stable and better in all designs studied here. The envelope method performed better than OLS, and the performance increased for informative data ($\rho^2 = 0.9$). However, it has an increased error with additional components in many situations.

Correlation between estimated and true regression coefficients (β) along with the mean square error of estimation is presented for 4 designs in Figure 6. In case of ordinary PLS and the envelope method, the correlation for design 1 from group (a) and design 3 from group (b), both having 15 predictors, is high for small components. However, for design 2 from group (c) and design 4 from group (d), envelope methods exhibit sudden decrease in the correlation with corresponding increase in estimation error. The impressive prediction performance of Bayes PLS is also seen from the high correlation of estimated coefficients and true coefficients. In addition, the average mean square error of regression for this methods is also small compared with others for all the components.

Although having low prediction error in case of envelope estimation method, the coefficient estimates are highly unstable for different components, which we can see from its variation in correlation with true coefficients (Figure 6, top). Bayes PLS and regular PLS estimates are more stable over different replicates and for different components (Figure 6, bottom) especially when $p \sim n$. This stability agrees with the low prediction error we have discussed before.

7 | COMPARISON OF ESTIMATORS USING NIR SPECTRA OF DIESEL FUELS

Let us consider an example using a real dataset. In this example, we have used data from http://www.eigenvector.com/data/SWRI/, which consists of NIR spectra of diesel fuels with different properties measured such as Catane Number. Since the variables in NIR spectra are highly correlated, we have selected a subset of every 10th variable as predictors and the property Catane Number as response. After removing missing observations, the first 150 observations were used as calibration set, and the rest 231 were used as validation set.

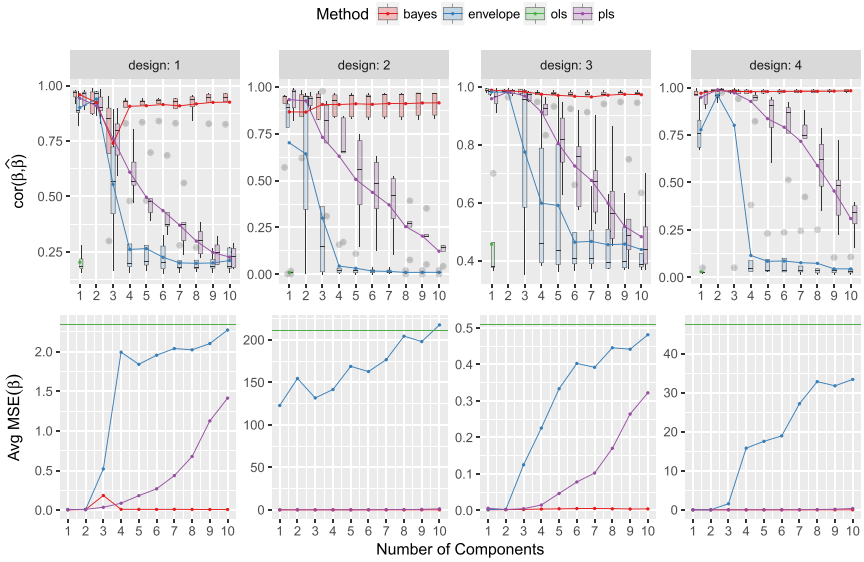


FIGURE 6 Correlation between true and estimated beta coefficient and beta estimation error. Box plots on the plots in first row show the variation in the correlation for each estimator and number of components used

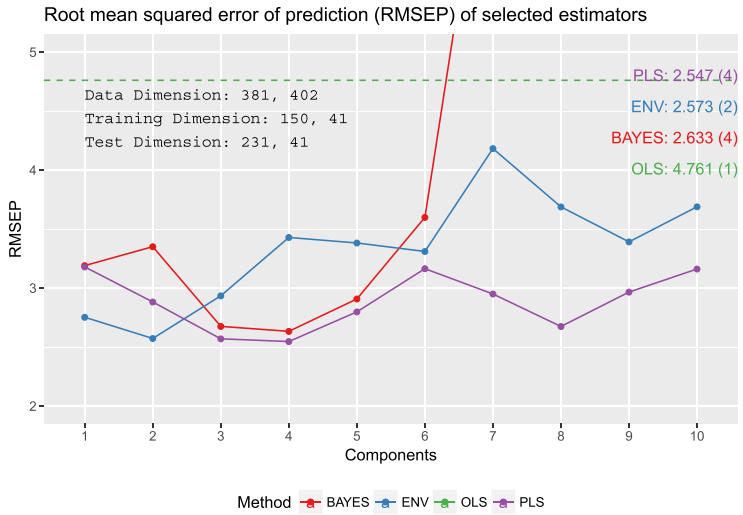


FIGURE 7 Root mean square error of prediction (RMSEP) from different estimators. Missing values were omitted in training and test datasets. ENV, envelope; OLS, ordinary least squares; PLS, partial least squares

Using the calibration set, a model with 1 to 10 components were fitted using PLS, envelope, and Bayes PLS methods. An OLS method was also fitted for reference. With each of these fitted models, the validation (test) set was used for prediction, and the root MSEP was measured. Based on the prediction error, Figure 7 compares the estimators we have considered.

The results from Figure 7 show quite different results from the systematic simulation study, mainly for Bayes PLS estimation. By using 3 and 4 components, the prediction from PLS and Bayes PLS is similar and can be considered their best. Envelope model is able to attain similar prediction error just in 2 components. It is important to notice that Bayes PLS

and envelope methods here are rather sensitive to the extra number of components, which also suggest that over-fitting must be examined before using the model for predicting new observations. In the example, all the methods have significant better performance than OLS.

8 | DISCUSSION

The purpose of the present article has been to discuss the approach to PLS regression via model reduction in the random $\boldsymbol{x}$ multiple regression model, and to compare estimators in this reduced model.

From simulations, the Bayes estimator under the PLS model seems to have very good properties. In virtually all of the 32 designs, the MSEP curve for Bayes PLS lies below that for ordinary PLS and also that for the maximum likelihood envelope model. A particularly desirable feature of Bayes PLS is that the MSEP curve seems to be almost flat for varying number of components. Thus, the error made by choosing a wrong number of components m by cross-validation must be expected to be small.

Envelope and Bayes PLS estimation methods, when compared with ordinary PLS methods, display better prediction performance (only when p is small for the envelope method). However, both of them have their disadvantages. The envelope method, as based on maximum likelihood, breaks down when p approaches n, while Bayes PLS has time-consuming computation, and in our simulations, it failed to converge for some cases.

However, in the results in the example using real data, the performance of Bayes PLS estimator is in contrast to its result from the simulated data. Since the predictors are highly correlated, only a few number of components are sufficient for the prediction, but when an extra number of components were used, the estimators seem to be influenced by the noise which increases with each additional component. In this respect, a more thorough study on Bayes PLS should be done for its contrast results on simulated and real dataset. A convergence issue in Bayes PLS can be suspected for the reason as seen in the example using simulated data.

For practical purposes, the ordinary PLS algorithm still seems to be a good option for prediction purposes, but from a statistical point of view, a closer study of its properties as $p \to \infty$ seems to be called for. We feel that the model approach of the present paper may give a good framework for such a study, both in terms of asymptotic expansions and in terms of further simulations. Such simulations may also include the cross-validated LASSO and other methods such as ridge regression, but note that these estimators are derived from other considerations than that of predicting the effect of relevant components.

This paper has been concentrated on the case of univariate response. We hope to discuss the multivariate case later.

ORCID

Inge Svein Helland http://orcid.org/0000-0002-7136-873X

REFERENCES

1. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*, 2nd ed. Springer: New York; 2009.
2. Wold S, Martens H, Wold H. The multivariate calibration problem in chemistry solved by the PLS method. In: Ruhe A, Kågstrøm B, eds. *Proc. Conf. Matrix Pencils. March 1982. Lecture Notes in Mathematics*. Heidelberg: Springer Verlag; 1983:286-293.
3. Martens H, Næs T. *Multivariate Calibration*. Chichester and New York: John Wiley & Sons; 1989.
4. Boulesteix AL, Strimmer K. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings Bioinf*. 2007;8(1):32-44.
5. Mehmood T, Ahmed B. The diversity in the applications of partial least squares: an overview. *J Chemom*. 2016;30(1):4-17.
6. Martens H, Martens M. *Multivariate Analysis of Quality. An Introduction*. Bristol, UK: IOP Publishing; 2001.
7. Frank LE, Friedman JH. A statistical view of some chemometrics regression tools. *Technometrics*. 1993;35(2):109-135.
8. Helland IS. On the structure of partial least squares regression. *Commun Stat-Simul Comput*. 1988;17(2):581-607.
9. Helland IS. Partial least squares regression and statistical models. *Scand J Stat*. 1990;17:97-114.
10. Sundberg R. Multivariate calibration—direct and indirect regression methodology. *Scand J Stat*. 1999;26(2):161-207.
11. Munck L, Jespersen BM, Rinnan Å, et al. A physiochemical theory on the applicability of soft mathematical models—experimentally interpreted. *J Chemom*. 2010;24(7-8):481-495.
12. Martens H. Quantitative big data: where chemometrics can cotribute. *J Chemom*. 2015;29:563-581.
13. Chung D, Keleş S. Sparse partial least squares classification for high dimensional data. *Stat Appl Genet Mol Biol*. 2010;9(1):17.

14. Cook RD, Forzani L. Big data and partial least-squares prediction. *Can J Stat*. 2017;46:62-78.
15. Krämer N. An overview on the shrinkage properties of partial least squares regression. *Comput Stat*. 2007;22(2):249-273.
16. Foschi P. *The geometry of PLS shrinkages*, University of Bologna; 2015.
17. Garthwaite PH. An interpretation of partial least squares. *J Am Stat Assoc*. 1994;89(425):122-127.
18. Stone M, Brooks RJ. Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression. *J R Stat Soc Ser B (Methodological)*. 1990;52(2):237-269.
19. Naik P, Tsai CL. Partial least squares estimator for single-index models. *J R Stat Soc: Ser B (Statistical Methodology)*. 2000;62(4):763-771.
20. Stoica P, Söderström T. Partial least squares: a first-order analysis. *Scand J Stat*. 1998;25(1):17-24.
21. Chun H, Keleş S. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *J R Stat Soc: Ser B (Statistical Methodology)*. 2010;72(1):3-25.
22. Krämer N, Sugiyama M. The degrees of freedom of partial least squares regression. *J Am Stat Assoc*. 2012;106(494):697-705.
23. Helland IS, Almøy T. Comparison of prediction methods when only a few components are relevant. *J Am Stat Assoc*. 1994;89(426):583-591.
24. Cook R, Helland I, Su Z. Envelopes and partial least squares regression. *J R Stat Soc: Ser B (Statistical Methodology)*. 2013;75(5):851-877.
25. Cook D, Su Z, Yang Y, et al. envlp: A MATLAB toolbox for computing envelope estimators in multivariate analysis. *J Stat Software*. 2015;62(1):1-20.
26. Cook RD, Forzani L, Su Z. A note on fast envelope estimation. *J Multivariate Anal*. 2016;150:42-54.
27. Cook RD, Zhang X. Algorithms for envelope estimation. *J Comput Graphical Stat*. 2016;25(1):284-300.
28. Helland IS, Saebø S, Tjelmeland H, et al. Near optimal prediction from relevant components. *Scand J Stat*. 2012;39(4):695-713.
29. Sæbø S, Almøy T, Helland IS. simrel–a versatile tool for linear model data simulation based on the concept of a relevant subspace and relevant predictors. *Chemom Intell Lab Syst*. 2015;146:128-135.
30. Cook RD, Zhang X. Simultaneous envelopes for multivariate linear regression. *Technometrics*. 2015;57(1):11-25.
31. Cook RD, Zhang X. Foundations for envelope models and methods. *J Am Stat Assoc*. 2015;110(510):599-611.
32. Helland IS. Reduction of regression models under symmetry. *Contemp Math*. 2001;287:139-154.
33. Cook RD, Li B, Chiaromonte F. Envelope models for parsimonious and efficient multivariate linear regression. *Stat Sin*. 2010;20(3):927-960.
34. Næs T, Helland IS. Relevant components in regression. *Scand J Stat*. 1993;20:239-250.
35. Cook RD, Zhang X. Fast envelope algorithms. *Stat Sin*. 2018;28(3):28.

How to cite this article: Helland IS, Sæbø S, Almøy T, Rimal R. Model and estimators for partial least squares regression. *Journal of Chemometrics*. 2018;32:e3044. https://doi.org/10.1002/cem.3044

COMPARISON OF MULTI-RESPONSE PREDICTION METHODS

Chemometrics and Intelligent Laboratory Systems 190 (2019) 10–21

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Comparison of multi-response prediction methods

Raju Rimal [a,*], Trygve Almøy [a], Solve Sæbø [b]

[a] Faculty of Chemistry and Bioinformatics, Norwegian University of Life Sciences, Ås, Norway
[b] Norwegian University of Life Sciences, Ås, Norway

ARTICLE INFO

Keywords:
Model-comparison
Multi-response
Simrel

ABSTRACT

While data science is battling to extract information from the enormous explosion of data, many estimators and algorithms are being developed for better prediction. Researchers and data scientists often introduce new methods and evaluate them based on various aspects of data. However, studies on the impact of/on a model with multiple response variables are limited. This study compares some newly-developed (envelope) and well-established (PLS, PCR) prediction methods based on real data and simulated data specifically designed by varying properties such as multicollinearity, the correlation between multiple responses and position of relevant principal components of predictors. This study aims to give some insight into these methods and help the researcher to understand and use them in further studies.

1. Introduction

The prediction has been an essential component of modern data science, whether in the discipline of statistical analysis or machine learning. Modern technology has facilitated a massive explosion of data however, such data often contain irrelevant information that consequently makes prediction difficult. Researchers are devising new methods and algorithms in order to extract information to create robust predictive models. Such models mostly contain predictor variables that are directly or indirectly correlated with other predictor variables. In addition, studies often consist of many response variables correlated with each other. These interlinked relationships influence any study, whether it is predictive modelling or inference.

Modern inter-disciplinary research fields such as chemometrics, econometrics and bioinformatics handle multi-response models extensively. This paper attempts to compare some multivariate prediction methods based on their prediction performance on linear model data with specific properties. The properties include the correlation between response variables, the correlation between predictor variables, number of predictor variables and the position of relevant predictor components. These properties are discussed more in the Experimental Design section. Among others, Sæbø et al. [26] and Almøy [2] have conducted a similar comparison in the single response setting. In addition, Rimal et al. [25] have also conducted a basic comparison of some prediction methods and their interaction with the data properties of a multi-response model. The main aim of this paper is to present a comprehensive comparison of contemporary prediction methods such as simultaneous envelope estimation (Senv) [8] and envelope estimation in predictor space (Xenv) [7] with customary prediction methods such as Principal Component Regression (PCR), Partial Least Squares Regression (PLS) using simulated dataset with controlled properties. In the case of PLS, we have used PLS1 which fits individual response separately and PLS2 which fits all the responses together. Experimental design and the methods under comparison are discussed further, followed by a brief discussion of the strategy behind the data simulation.

2. Simulation model

Consider a model where the response vector ($\mathbf{y}$) with m elements and predictor vector ($\mathbf{x}$) with p elements follow a multivariate normal distribution as follows,

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

where, $\boldsymbol{\Sigma}_{xx}$ and $\boldsymbol{\Sigma}_{yy}$ are the variance-covariance matrices of $\mathbf{x}$ and $\mathbf{y}$, respectively, $\boldsymbol{\Sigma}_{xy}$ is the covariance between $\mathbf{x}$ and $\mathbf{y}$ and $\boldsymbol{\mu}_x$ and $\boldsymbol{\mu}_y$ are mean vectors of $\mathbf{x}$ and $\mathbf{y}$, respectively. A linear model based on (1) is,

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

where, $\underset{m\times p}{\boldsymbol{\beta}^t}$ is a matrix of regression coefficients and ε is an error term

* 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.2019.05.004
Received 19 March 2019; Received in revised form 30 April 2019; Accepted 9 May 2019
Available online 15 May 2019

such that $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{y|x})$. Here, $\boldsymbol{\beta}^t = \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}$ and $\boldsymbol{\Sigma}_{y|x} = \boldsymbol{\Sigma}_{yy} - \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}$.

In a model like (2), we assume that the variation in response **y** is partly explained by the predictor **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 relevant space in predictor can explain (Fig. 1). Cook et al. [7] and Cook and Zhang [8] have referred to the relevant space as material space and the irrelevant space as immaterial space.

With an orthogonal transformation of **y** and **x** to latent variables **w** and **z**, respectively, by $\mathbf{w} = \mathbf{Qy}$ and $\mathbf{z} = \mathbf{Rx}$, where **Q** and **R** are orthogonal rotation matrices, an equivalent model to (1) in terms of the latent variables can be written as,

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

where, $\boldsymbol{\Sigma}_{ww}$ and $\boldsymbol{\Sigma}_{zz}$ are the variance-covariance matrices of **w** and **z**, respectively. $\boldsymbol{\Sigma}_{zw}$ is the covariance between **z** and **w**. $\boldsymbol{\mu}_w$ and $\boldsymbol{\mu}_z$ are the mean vector of **z** and **w** respectively.

Here, the elements of **w** and **z** are the principal components of responses and predictors, which will respectively be referred to respectively as "response components" and "predictor components". The column vectors of respective rotation matrices **Q** and **R** are the eigenvectors corresponding to these principal components. We can write a linear model based on (3) as,

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

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

Following the concept of relevant space, a subset of predictor components can be imagined to span the predictor space. These components can be regarded as relevant predictor components. Naes and Martens [22] introduced the concept of relevant components which was explored further by Helland [11], Næs and Helland [21], Helland and Almøy [13] and Helland [12]. The corresponding eigenvectors were referred to as relevant eigenvectors. A similar logic is introduced by Cook et al. [7] and later by Cook et al. [5] as an envelope which is the space spanned by the relevant eigenvectors [4, pp. 101].

In addition, various simulation studies have been performed with the model based on the concept of relevant subspace. A simulation study by Almøy [2] has used a single response simulation model based on reduced regression and has compared some contemporary multivariate estimators. In recent years Helland et al. [15], Sæbø et al. [26], Helland et al. [14] and Rimal et al. [25] implemented similar simulation examples similar to those we are discussing in this study. This paper, however, presents an elaborate comparison of the prediction using multi-response simulated linear model data. The properties of the simulated data are varied through different levels of simulation-parameters based on an experimental design. Rimal et al. [25] provide a detailed discussion of the simulation model that we have adopted here. The following section presents the estimators being compared in more detail.

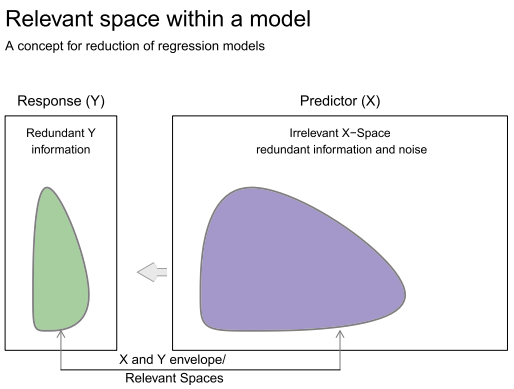


Fig. 1. Relevant space in a regression model.

3. Prediction methods

Partial least squares regression (PLS) and Principal component regression (PCR) have been used in many disciplines such as chemometrics, econometrics, bioinformatics and machine learning, where wide predictor matrices, i.e. p (number or predictors) $> n$ (number of observation) are common. These methods are popular in multivariate analysis, especially for exploratory studies and predictions. In recent years, a concept of envelope introduced by Cook et al. [6] based on the reduction in the regression model was implemented for the development of different estimators. This study compares these prediction methods based on their prediction performance on data simulated with different controlled properties.

Principal Components Regression (PCR): Principal components are the linear combinations of predictor variables such that the transformation makes the new variables uncorrelated. In addition, the variation of the original dataset captured by the new variables is sorted in descending order. In other words, each successive component captures maximum variation left by the preceding components in predictor variables [18]. Principal components regression uses these principal components as a new predictor to explain the variation in the response.

Partial Least Squares (PLS): Two variants of PLS: PLS1 and PLS2 are used for comparison. The first one considers individual response variables separately, i.e. each response is predicted with a single response model, while the latter considers all response variables together. In PLS regression, the components are determined so as to maximize a covariance between response and predictors [10]. Among other, there are three main PLS algorithms NIPALS, SIMPLS and Kernel Algorithm all of which removes the extracted information through deflation and makes the resulting new variables orthogonal. The algorithms differ in the deflation strategy and computation of various weight vectors [1] and here we have used the kernel version of PLS. R-package pls [20] is used for both PCR and PLS methods.

Envelopes: The envelope, introduced by Cook et al. [6], was first used to define response envelope [7] as the smallest subspace in the response space and must be a reducing subspace of $\Sigma_{y|x}$ such that the span of regression coefficients lies in that space. Since a multivariate linear regression model contains relevant (material) and irrelevant (immaterial) variation in both response and predictor, the relevant part provides information, while the irrelevant part increases the estimative variation. The concept of the envelope uses the relevant part for estimation while excluding the irrelevant part consequently increasing the efficiency of the model [9].

The concept was later extended to the predictor space, where the predictor envelope was defined [5]. Further Cook and Zhang [8] used envelopes for joint reduction of the responses and predictors and argued that this produced efficiency gains that were greater than those derived by using individual envelopes for either the responses or the predictors separately. All the variants of envelope estimations are based on maximum likelihood estimation. Here we have used predictor envelope (Xenv) and simultaneous envelope (Senv) for the comparison. R-package Renvlp [19] is used for both Xenv and Senv methods.

3.1. Modification in envelope estimation

Since envelope estimators (Xenv and Senv) are based on maximum likelihood estimation (MLE), it fails to estimate in the case of wide

matrices, i.e. $p > n$. To incorporate these methods in our comparison, we have used the principal components ($\mathbf{z}$) of the predictor variables ($\mathbf{x}$) as predictors, using the required number of components for capturing 97.5% of the variation in $\mathbf{x}$ for the designs where $p > n$. The new set of variables $\mathbf{z}$ were used for envelope estimation. The regression coefficients ($\widehat{\boldsymbol{\alpha}}$) corresponding to these new variables $\mathbf{z}$ were transformed back to obtain coefficients for each predictor variabl

$$\widehat{\boldsymbol{\beta}} = \mathbf{e}_k \widehat{\boldsymbol{\alpha}_k}$$

where $\mathbf{e}_k$ is a matrix of eigenvectors with the first k number of components. Only simultaneous envelope allows to specify the dimension of response envelope and all the simulation is based on a single latent dimension of response, so it is fixed at two in the simulation study. In the case of Senv, when the envelope dimension for response is the same as the number of responses, it degenerates to the Xenv method and if the envelope dimension for the predictor is the same as the number of predictors, it degenerates to the standard multivariate linear regression [19].

4. Experimental design

This study compares prediction methods based on their prediction ability. Data with specific properties are simulated, some of which are easier to predict than others. These data are simulated using the R-package simrel, which is discussed in Sæbø et al. [26] and Rimal et al. [25]. Here we have used four different factors to vary the property of the data: a) Number of predictors (p), b) Multicollinearity in predictor variables (gamma), c) Correlation in response variables (eta) and d) position of predictor components relevant for the response (relpos). Using two levels of p, gamma and relpos and four levels of eta, 32 sets of distinct properties are designed for the simulation.

Number of predictors: To observe the performance of the methods on tall and wide predictor matrices, 20 and 250 predictor variables are simulated with the number of observations fixed at 100. Parameter p controls these properties in the simrel function.

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

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

Here, $\lambda_i, i = 1, 2, \ldots p$ are eigenvalues of the predictor variables. We have used 0.2 and 0.9 as different levels of gamma. The higher the value of gamma, the higher the multicollinearity will be, and vice versa. In our simulations, the higher and lower gamma values corresponded to the maximum correlation between the predictors equal to 0.990 and 0.709, respectively, in the case of $p = 20$ variables. In the case of $p = 250$ the corresponding values for the maximum correlation were 0.998 to 0.923.

Correlation in response variables: Correlation among response variables has been explored to a lesser extent. Here we have tried to explore that part with four levels of correlation in the response variables. We have used the eta (η) parameter of simrel for controlling the decline in eigenvalues corresponding to the response variables as (6).

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

Here, $\kappa_j, i = 1, 2, \ldots m$ are the eigenvalues of the response variables and m is the number of response variables. We have used 0, 0.4, 0.8 and 1.2 as different levels of eta. The larger the value of eta, the larger will be the correlation will be between response variables and vice versa. In our simulation, the different levels of eta from small to large correspond to the maximum correlation of 0, 0.442, 0.729 and 0.878 between the response variables respectively.

Position of predictor components relevant to the response: The principal components of the predictors are ordered. The first principal component captures most of the variation in the predictors. The second captures most of the remainder left by the first principal component and so on. In highly collinear predictors, the variation captured by the first few components is relatively high. However, if those components are not relevant for the response, prediction becomes difficult [13]. Here, two levels of the positions of these relevant components are used as 1, 2, 3, 4 and 5, 6, 7, 8.

Moreover, a complete factorial design from the levels of the above parameters gave us 32 designs. Each design is associated with a dataset having unique properties. Fig. 2, shows all the designs. For each design and prediction method, 50 datasets were simulated as replicates. In total, there were $5 \times 32 \times 50$, i.e. 8000 simulated datasets.

Common parameters: Each dataset was simulated with $n = 100$ number of observation and $m = 4$ response variables. Furthermore, the coefficient of determination corresponding to each response components in all the designs is set to 0.8. The informative and uninformative latent components are generated according to (3). Since $\boldsymbol{\Sigma}_{ww}$ and $\boldsymbol{\Sigma}_{zz}$ are diagonal matrices, the components are independent within $\boldsymbol{w}$ and $\boldsymbol{z}$, but dependence between the latent spaces of $\boldsymbol{x}$ and $\boldsymbol{y}$ are secured through the non-zero elements of $\boldsymbol{\Sigma}_{wz}$ with positions defined by the relpos and ypos parameters. The latent components are subsequently rotated to obtain the population covariance structure of response and predictor variables. In addition, we have assumed that there is only one informative response component. Hence, the informative response component after the orthogonal rotation together with three uninformative response components generates four response variables. This spreads out the information in all simulated response variables. For further details on the simulation tool, see Ref. [25].

An example of simulation parameters for the first design is as follows:

```
simrel(
    n       = 100,                ## Training samples
    p       = 20,                 ## Predictors
    m       = 4,                  ## Responses
    q       = 20,                 ## Relevant predictors
    relpos  = list(c(1, 2, 3, 4)), ## Relevant predictor components index
    eta     = 0,                  ## Decay factor of response eigenvalues
    gamma   = 0.2,                ## Decay factor of predictor eigenvalues
    R2      = 0.8,                ## Coefficient of determination
    ypos    = list(c(1, 2, 3, 4)),
    type    = "multivariate"
)
```

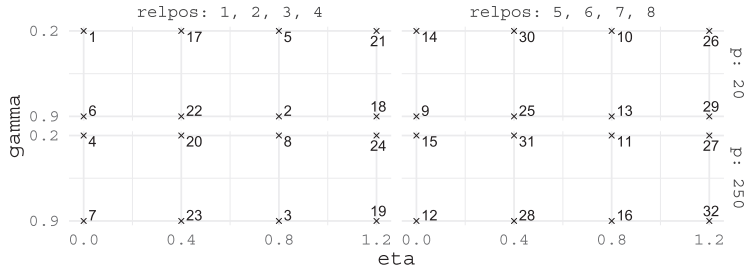


Fig. 2. Experimental Design of simulation parameters. Each point represents a unique data property.

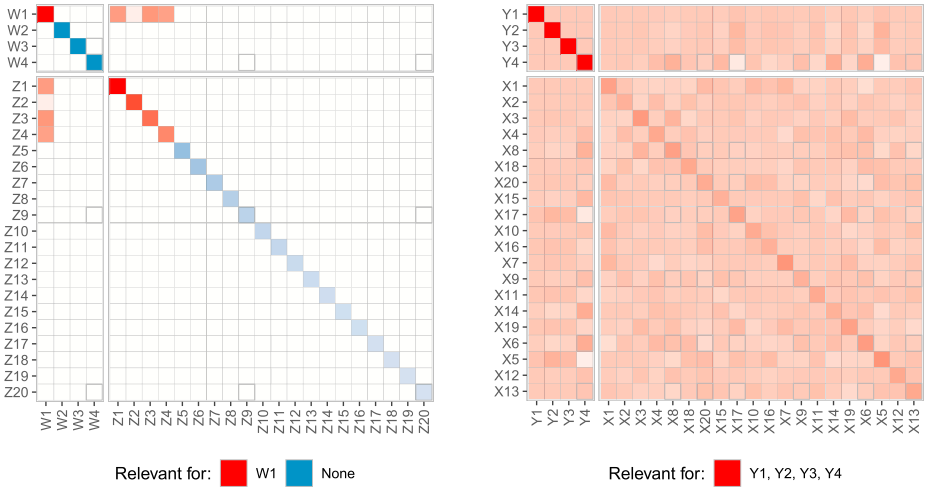


Fig. 3. (left) Covariance structure of latent components (right) Covariance structure of predictor and response.

The covariance structure of the data simulated with this design in Fig. 3 shows that the predictor components at positions 1, 2, 3 and 4 are relevant for the first response component. After the rotation with an orthogonal rotation matrix, all predictor variables are somewhat relevant for all response variables, satisfying other desired properties such as multicollinearity and coefficient of determination. For the same design, Fig. 4 (top left) shows that the predictor components 1, 2, 3 and 4 are relevant for the first response component. All other predictor components are irrelevant and all other response components are uninformative. However, due to the orthogonal rotation of the informative response component together with uninformative response components, all response variables in the population have similar covariance with the relevant predictor components (Fig. 4 (top right)). The sample covariances between the predictor components and predictor variables with response variables are shown in Fig. 4 (bottom left) and (bottom right) respectively.

A similar description can be made for all 32 designs, where each of the designs holds the properties of the data they simulate. These data are used by the prediction methods discussed in the previous section. Each prediction method is given independently simulated datasets in order to give them an equal opportunity to capture the dynamics in the data.

5. Basis of comparison

This study focuses mainly on the prediction performance of the methods with an emphasis specifically on the interaction between the properties of the data controlled by the simulation parameters and the prediction methods. The prediction performance is measured based on the following:

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

Let us define,

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

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

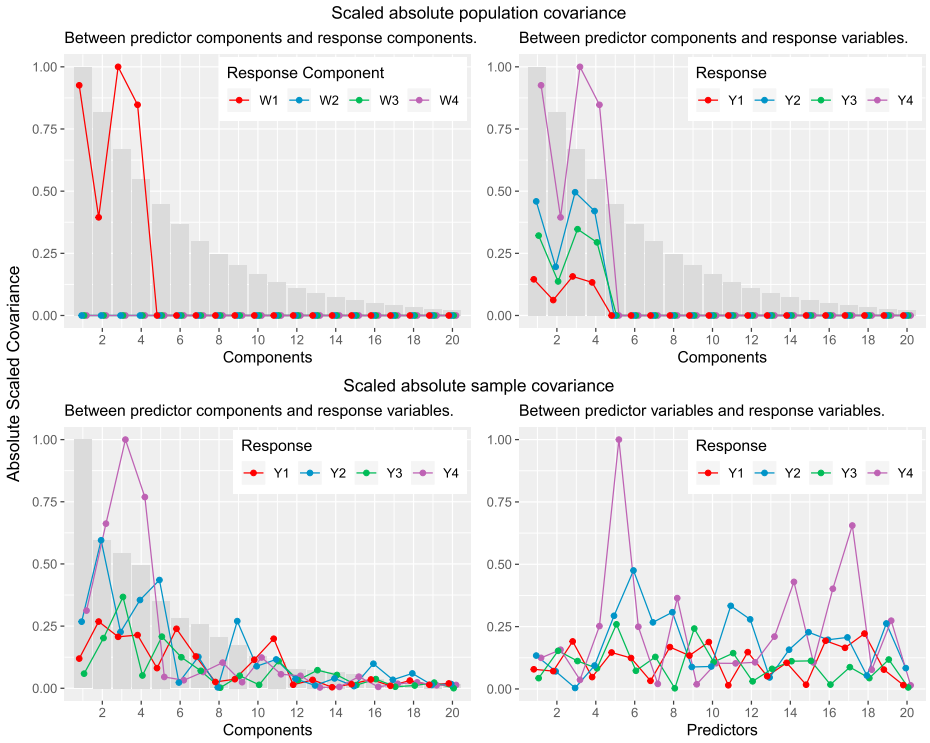


Fig. 4. Expected Scaled absolute covariance between predictor components and response components (top left). Expected Scaled absolute covariance between predictor components and response variables (top right). Sample scaled absolute covariance between predictor components and response variables (bottom left). Sample scaled absolute covariance between predictor variables and response variables (bottom right). The bar graph in the background represents eigenvalues corresponding to each component in the population (top plots) and in the sample (bottom plots). One can compare the top-right plot (true covariance of the population) with bottom-left (covariance in the simulated data) which shows a similar pattern for different components.

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

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

The following section focuses on the data for the estimation of these prediction errors that are used for the two models discussed above in a) and b) of this section.

6. Data preparation

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

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

Thus there are 32 (designs) × 5 (methods) × 11 (number of components) × 50 (replications), i.e. 88000 observations corresponding to the response variables from Y1 to Y4.

Since our discussions focus on the average minimum prediction error that a method can obtain and the average number of components they use to get the minimum prediction error in each replicates, the dataset discussed above is summarized as constructing the following two smaller datasets. Let us call them *Error Dataset* and *Component Dataset*.

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

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

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

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

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

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

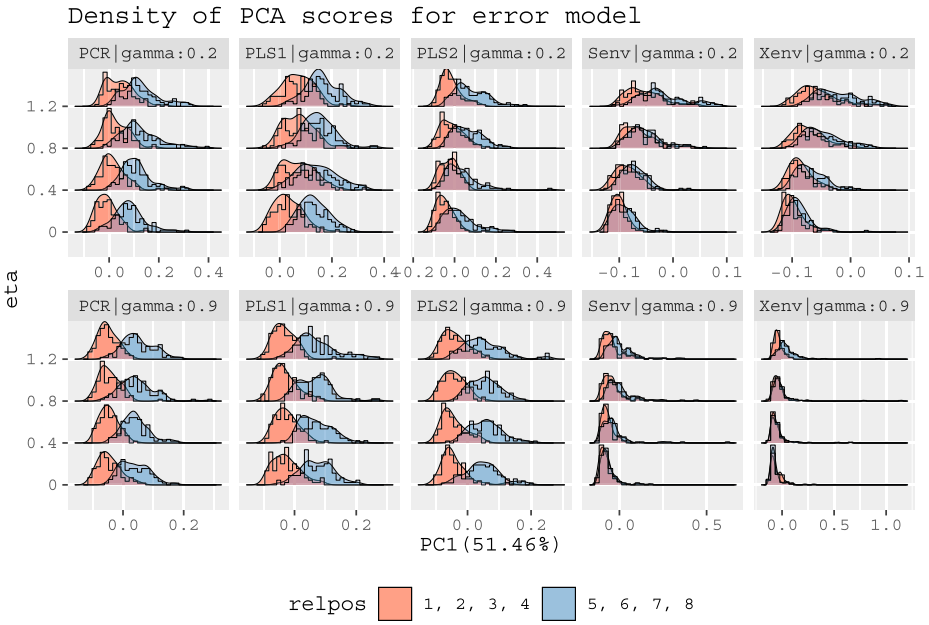


Fig. 5. Scores density corresponding to first principal component of *error dataset* (**u**) subdivided by methods, gamma and eta and grouped by relpos.

7. Exploration

This section explores the variation in the *error dataset* and the *component dataset* for which we have used Principal Component Analysis (PCA). Let $\mathbf{t}_u$ and $\mathbf{t}_v$ be the principal component score sets corresponding to PCA run on the **u** and **v** matrices respectively. The scores density in Fig. 5 corresponds to the first principal component of **u**, i.e. the first column of $\mathbf{t}_u$.

Since higher prediction errors correspond to high scores, the plot shows that the PCR, PLS1 and PLS2 methods are influenced by the two levels of the position of relevant predictor components. When the relevant predictors are at positions 5, 6, 7, 8, the eigenvalues corresponding to them are relatively smaller. This also suggests that PCR, PLS1 and PLS2 depend greatly on the position of the relevant components, and the variation of these components affects their prediction performance. However, the envelope methods appeared to be less influenced by relpos in this regard.

In addition, the plot also shows that the effect of gamma, i.e., the level of multicollinearity, has a lesser effect when the relevant predictors are at positions 1, 2, 3, 4. This indicates that the methods are somewhat robust for handling collinear predictors. Nevertheless, when the relevant predictors are at positions 5, 6, 7, 8, high multicollinearity results in a small variance of these relevant components and consequently yields poor prediction. This is in accordance with the findings of Helland and Almøy [13].

Furthermore, the density curves for PCR, PLS1 and PLS2 are similar for different levels of eta, i.e., the factor controlling the correlation between responses. However, the envelope models have been shown to have distinct interactions between the positions of relevant components (relpos) and eta. Here higher levels of eta have yielded higher scores and clear separation between two levels of relpos. In the case of high multicollinearity, envelope methods have resulted in some large outliers indicating that in some cases that the methods can result in giving an unexpected prediction.

In Fig. 6, the higher scores suggest that methods have used a larger number of components to give minimum prediction error. The plot also shows that the relevant predictor components at 5, 6, 7, 8 give larger prediction errors than those in positions 1, 2, 3, 4. The pattern is more distinct in large multicollinearity cases and PCR and PLS methods. Both the envelope methods have shown equally enhanced performance at both levels of relpos and gamma. However, for data with low multicollinearity ($\gamma = 0.2$), the envelope methods have used a lesser number of components on average than in the high multicollinearity cases to achieve minimum prediction error.

8. Statistical analysis

This section has modelled the *error data* and the *component data* as a function of the simulation parameters to better understand the connection between data properties and prediction methods using multivariate analysis of variation (MANOVA).

Let us consider a model with third order interaction of the simulation parameters (p, gamma, eta and relpos) and Methods as in (11) and (12) using datasets **u** and **v**, respectively. Let us refer to them as the *error model* and the *component model*.

Error Model:

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

Component Model:

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

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

Although there are several test-statistics for MANOVA, all are essentially equivalent for large samples [17]. Here we will use Pillai's trace statistic which is defined as,

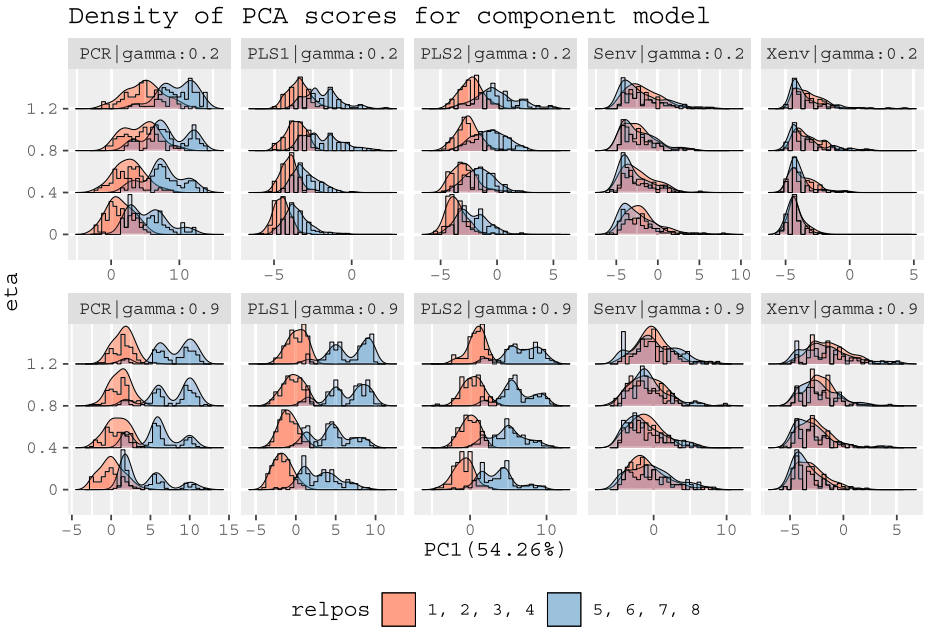


Fig. 6. Score density corresponding to the first principal component of the *component dataset* (v) subdivided by methods, gamma and eta and grouped by relpos.

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

Here the matrix **H** holds between-sum-of-squares and sum-of-products for each of the predictors. The matrix **E** has a within the sum of squares and sum of products for each of the predictors. ν_i represents the eigenvalues corresponding to $\mathbf{E}^{-1}\mathbf{H}$ [24].

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

Error Model: Fig. 7 (left) shows the Pillai's trace statistic for factors of the *error model*. The main effect of Method followed by relpos, eta and gamma have the largest influence on the model. A highly significant two-factor interaction of Method with gamma followed by the relpos and eta clearly shows that methods perform differently for different levels of these data properties. The significant third order interaction between Method, eta and gamma suggest that the performance of a method differs for a given level of multicollinearity and the correlation between the responses. Since only some methods consider modelling predictor and response together, the prediction is affected by the level of correlation between the responses (eta) for a given method.

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

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

8.1. Effect analysis of error model

The large difference in the prediction error for the envelope models in Fig. 8 (left) is intensified when the position of the relevant predictor is at 5, 6, 7, 8. The results also show that the envelope methods are more sensitive to the levels of eta than the rest of the methods. In the case of PCR and PLS, the difference in the effect of levels of eta is small.

In Fig. 8 (right), we can see that the multicollinearity (controlled by gamma) has affected all the methods. However, envelope methods have better performance on low multicollinearity, as opposed to high multicollinearity, and PCR, PLS1 and PLS2 are robust for high multicollinearity. Despite handling high multicollinearity, these methods have higher prediction error in both cases of multicollinearity than the envelope methods.

8.2. Effect analysis of the component model

Unlike for prediction errors, Fig. 9 (left) shows that the number of components used by the methods to obtain minimum prediction error is less affected by the levels of eta. All methods appear to use on average more components when eta increases. Envelope methods are able to obtain minimum prediction error by using components ranging from 1 to 3 in both the cases of relpos. This value is much higher in the case of PCR as its prediction is based only on the principal components of the predictor matrix. The number of components used by this method ranges from 3 to 5 when relevant components are at positions 1, 2, 3, 4 and 5 to 8 when relevant components are at positions 5, 6, 7, 8.

When relevant components are at position 5, 6, 7, 8, the eigenvalues of relevant predictors become smaller and responses are relatively difficult to predict. This becomes more critical for high multicollinearity cases. Fig. 9 (right) shows that the envelope methods are less influenced by the level of relpos and are particularly better in achieving minimum prediction error using a fewer number of components than other methods.

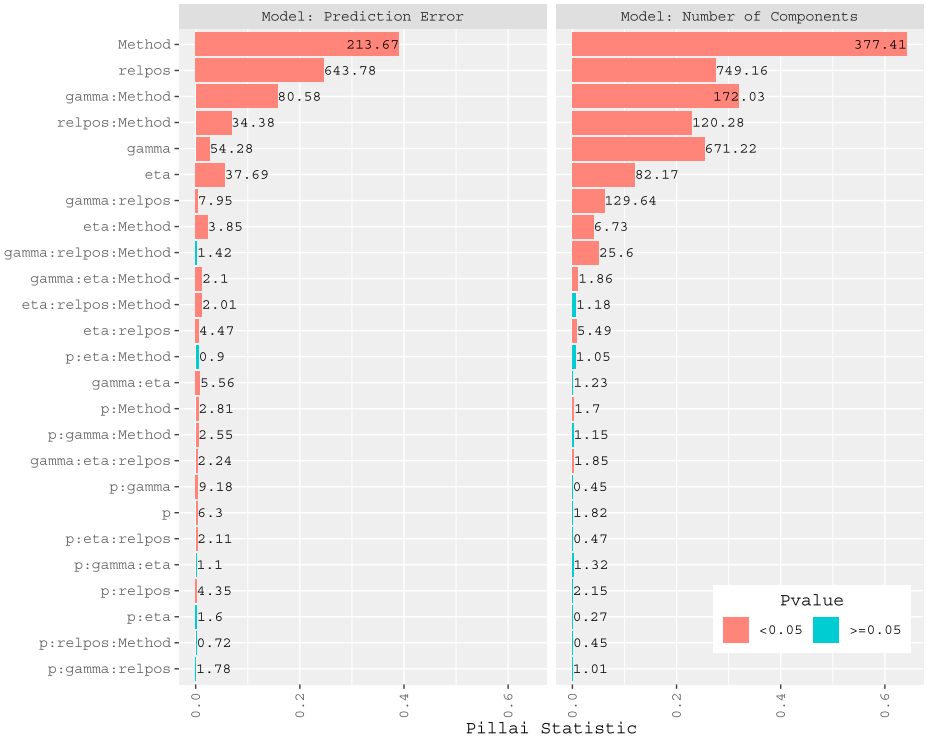


Fig. 7. Pillai Statistic and F-value for the MANOVA model. The bar represents the Pillai Statistic and the text labels are F-value for the corresponding factor.

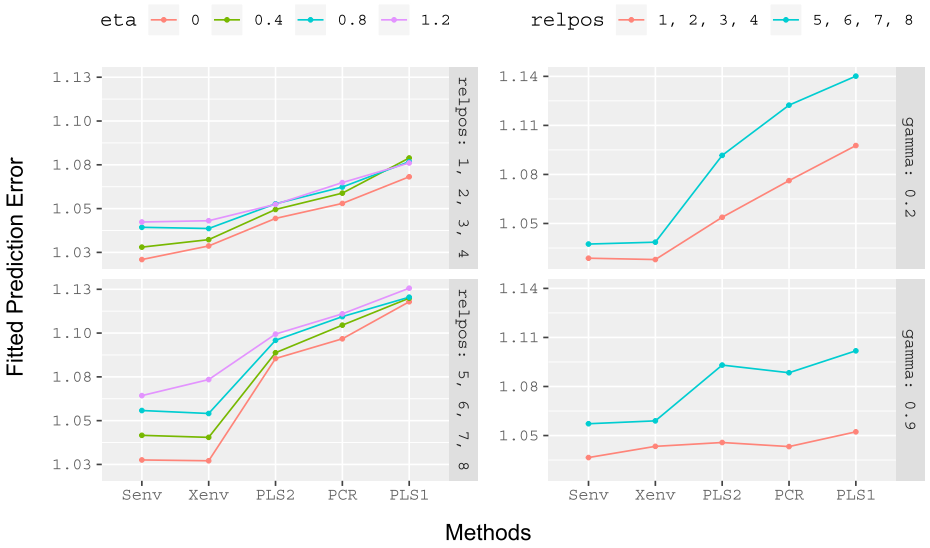


Fig. 8. Effect plot of some interactions of the multivariate linear model of prediction error.

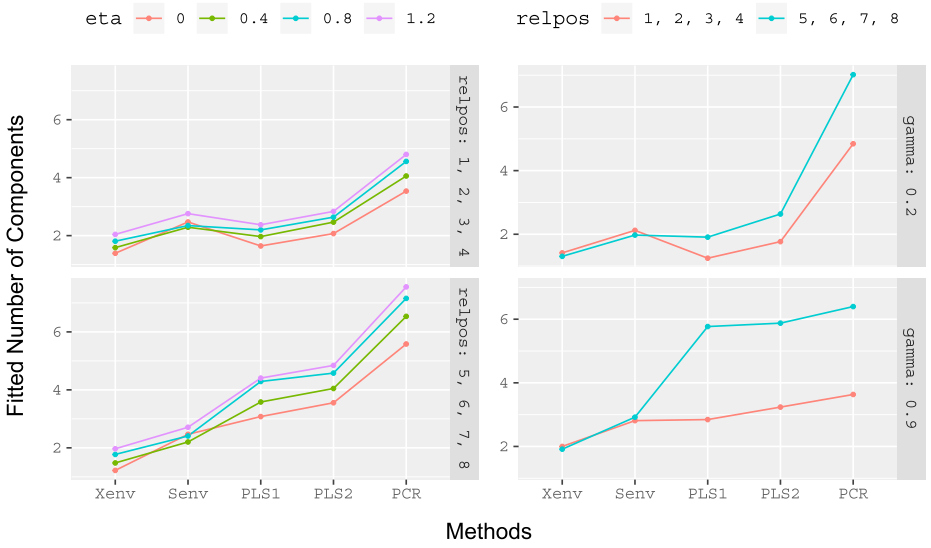


Fig. 9. Effect plot of some interactions of the multivariate linear model of the number of components to get minimum prediction error.

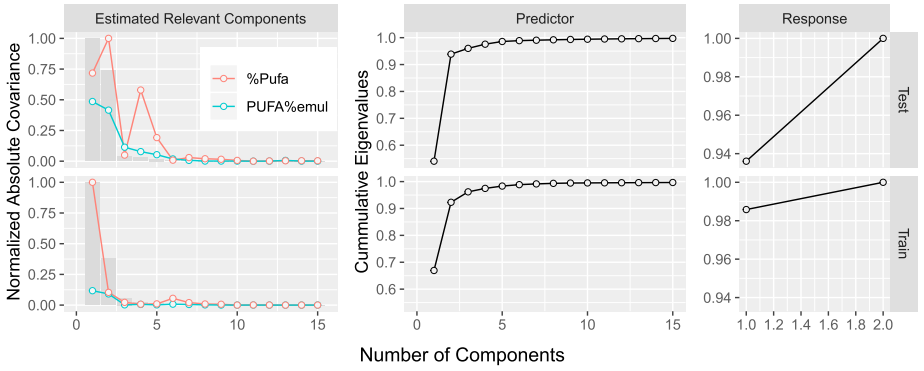


Fig. 10. (Left) Bar represents the eigenvalues corresponding to Raman Spectra. The points and line are the covariances between response and the principal components of Raman Spectra. All the values are normalized to scale from 0 to 1. (Middle) Cumulative sum of eigenvalues corresponding to predictors. (Right) The cumulative sum of eigenvalues corresponding to responses. The top and bottom row corresponds to test and training datasets respectively.

9. Examples

In addition to the analysis with the simulated data, the following two examples explore the prediction performance of the methods using real datasets. Since both examples have wide predictor matrices, principal components explaining 97.5% of the variation in them are used for envelope methods. The coefficients were transformed back after the estimation.

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

This dataset contains 44 training samples and 25 test samples of fatty acid information expressed as a) percentage of total sample weight and b) the percentage of total fat content. The dataset is borrowed from Næs et al. [23] where more information can be found. The samples were analysed using Raman spectroscopy from which 1096 wavelength variables were obtained as predictors. Raman spectroscopy provides detailed chemical information from minor components in food. The aim of this example is to compare how well the prediction methods that we have considered are able to predict the contents of PUFA using these Raman spectra.

Fig. 10 (left) shows that the first few predictor components are somewhat correlated with response variables. In addition, the most variation in predictors is explained by less than five components (middle). Further, the response variables are highly correlated, suggesting that a single latent dimension explains most of the variation (right). We may therefore also believe that the relevant latent space in the response matrix is of dimension one. This resembles Design 19 (Fig. 2) from our simulation.

Using a range of components from 1 to 15, regression models were fitted using each of the methods. The fitted models were used to predict the test observation, and the root mean squared error of prediction (RMSEP) was calculated. Fig. 11 shows that PLS2 obtained a minimum

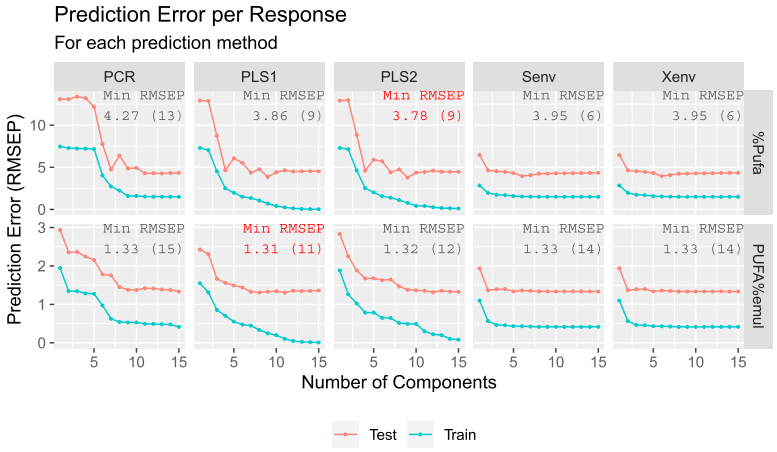


Fig. 11. Prediction Error of different prediction methods using different number of components.

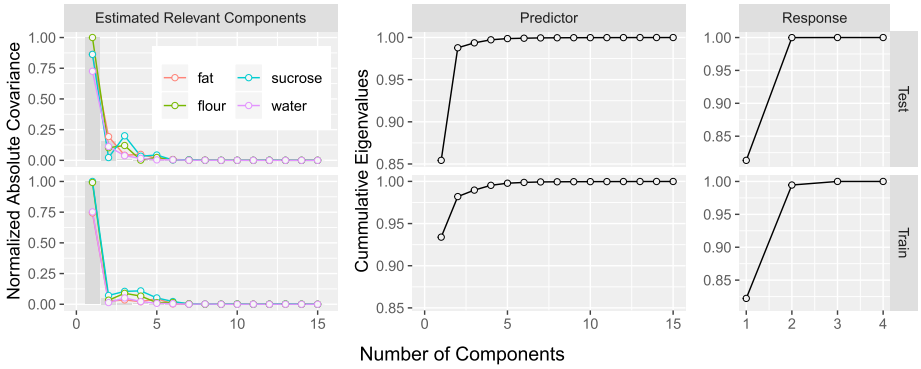


Fig. 12. (Left) Bar represents the eigenvalues corresponding to NIR Spectra. The points and line are the covariances between response and the principal components of NIR Spectra. All the values are normalized to scale from 0 to 1. (Middle) Cumulative sum of eigenvalues corresponding to predictors. (Right)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

prediction error of 3.783 using 9 components in the case of response % Pufa, while PLS1 obtained a minimum prediction error of 1.308 using 11 components in the case of response PUFA%emul. However, the figure also shows that both envelope methods have reached to almost minimum prediction error in fewer number of components. This pattern is also visible in the simulation results (Fig. 9).

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

The dataset consists of 700 wavelengths of NIR spectra (1100–2498 nm in steps of 2 nm) that were used as predictor variables. There are four response variables corresponding to the yield percentages of (a) fat, (b) sucrose, (c) flour and (d) water. The measurements were taken from 40 training observation of biscuit dough. A separate set of 32 samples created and measured on different occasions were used as test observations. The dataset is borrowed from Indahl [16] where further information can be obtained.

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

Fig. 13 (corresponding to Fig. 11) shows the root mean squared error for both test and train prediction of the biscuit dough data. Here four different methods have minimum test prediction error for the four responses. As the structure of the data is similar to that of the first example, the pattern in the prediction is also similar for all methods.

The prediction performance on the test data of the envelope methods appears to be more stable compared to the PCR and PLS methods. Furthermore, the envelope methods achieve good performance generally using fewer components, which is in accordance with Fig. 6.

10. Discussions and conclusion

Analysis using both simulated data and real data has shown that the envelope methods are more stable, less influenced by relpos and gamma and in general, performed better than PCR and PLS methods. These methods are also found to be less dependent on the number of

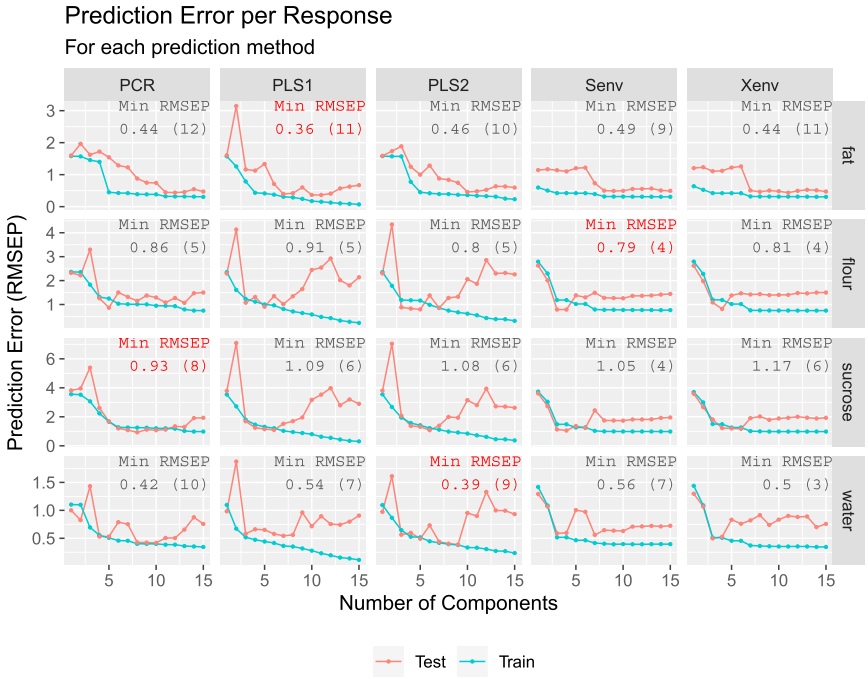


Fig. 13. Prediction Error of different prediction methods using different number of components.

components.

Since the facet in Figs. 5 and 6 have their own scales, despite having some large prediction errors seen at the right tail, envelope methods still have a smaller prediction error and have used a fewer number of components than the other methods.

The envelope methods may have this problem of being caught in a local optimum of the objective function. If these cases of sub-optimal convergence were identified and rerun to obtain better convergence, the envelope results may have become even better. Particularly in the case of the simultaneous envelope, since users can specify the number of dimension for the response envelope, the method can leverage the relevant space of response while PCR, PLS and Xenv are constrained to play only on predictor space.

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

Two clear comments can be made about the effect of correlation of response on the prediction methods. The highly correlated response has shown the highest prediction error in general and the effect is most distinct in envelope methods. Since the envelope methods identify the relevant space as the span of relevant eigenvectors, the methods are able to obtain the minimum average prediction error by using a lesser number of components for all levels of eta.

To our knowledge, the effect of correlation in the response on PCR and PLS methods has been explored only to a limited extent. In this regards, it is interesting to see that these methods have applied a large number of components and returned a larger prediction error than envelope methods in the case of highly correlated responses. To fully understand the effect of eta, it is necessary to study the estimation performance of these methods with different numbers of components.

In addition, since using principal components or actual variables as predictors in envelope methods has shown similar results, we have used principal components that have explained 97.5% of the variation, as mentioned previously, in the cases of envelope methods for the designs where $p > n$. Using 97.5% is slightly arbitrary here, but for the chosen simulation designs this proportion captured a fair amount of variations in predictor variables and also reduce the dimension significantly while enabling us to use envelope methods in all settings. The analyst should choose this number to balance the explained amount of variation to the number of components which is practical for model fitting using the envelope model. The methodology used to adapt envelopes to settings in which $p > n$ is, in fact, the same as that used by PLS: reduce by principal components, run the method, and then back transform to the original scale. The minor relative impact of p shown in Fig. 7 suggests that this adaptation method is useful.

The results from this study will help researchers to understand these methods for their performance in various linear model data and encourage them to use newly developed methods such as the envelopes. Since this study has focused entirely on prediction performance, further analysis of the estimative properties of these methods is required. A study of estimation error and the performance of methods on the non-optimal number of components can give a deeper understanding of these methods.

A shiny application [3] is available at http://therimalaya.shinyapps.io/Comparison where all the results related to this study can be visualized. In addition, a GitHub repository at https://github.com/therimalaya/03-prediction-comparison can be used to reproduce this study.

Acknowledgment

We are grateful to Inge Helland for his inputs on this paper throughout the period. His guidance on the envelope models and his review of the paper helped us greatly. Our gratitude also goes to thank Kristian Lillan, Ulf Indahl, Tormod Næs, Ingrid Måge and the team for

providing the data for analysis. We are also thankful to the reviewers for their comments which helped us to improve this paper.

References

[1] A. Alin, Comparison of pls algorithms when number of objects is much larger than number of variables, Stat. Pap. 50 (4) (2009) 711–720. https://doi.org/10.1007/s00362-009-0251-7.

[2] T. Almøy, A simulation study on comparison of prediction methods when only a few components are relevant, Comput. Stat. Data Anal. 21 (1) (jan 1996) 87–107.

[3] W. Chang, J. Cheng, J. Allaire, Y. Xie, J. McPherson, Shiny: Web Application Framework for R. R Package Version 1.2.0, 2018. https://CRAN.R-project.org/package=shiny.

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

[5] R.D. Cook, I.S. Helland, Z. Su, Envelopes and partial least squares regression, J. R. Stat. Ser. Soc. B Stat. Methodol. 75 (5) (2013) 851–877.

[6] R.D. Cook, B. Li, F. Chiaromonte, Dimension reduction in regression without matrix inversion, Biometrika 94 (3) (aug 2007) 569–584.

[7] R.D. Cook, B. Li, F. Chiaromonte, Envelope models for parsimonious and efficient multivariate linear regression, Stat. Sin. 20 (3) (2010) 927–1010.

[8] R.D. Cook, X. Zhang, Simultaneous envelopes for multivariate linear regression, Technometrics 57 (1) (2015) 11–25.

[9] R.D. Cook, X. Zhang, Algorithms for envelope estimation, J. Comput. Graph. Stat. 25 (1) (2016) 284–300.

[10] S. de Jong, SIMPLS: an alternative approach to partial least squares regression, Chemometr. Intell. Lab. Syst. 18 (3) (mar 1993) 251–263.

[11] I.S. Helland, Partial least squares regression and statistical models, Scand. J. Stat. 17 (2) (1990) 97–114.

[12] I.S. Helland, Model reduction for prediction in regression models, Scand. J. Stat. 27 (1) (mar 2000) 1–20.

[13] I.S. Helland, T. Almøy, Comparison of prediction methods when only a few components are relevant, J. Am. Stat. Assoc. 89 (426) (1994) 583–591.

[14] I.S. Helland, S. Saebø, T. Almøy, R. Rimal, S. Sæbø, T. Almøy, R. Rimal, Model and estimators for partial least squares regression, J. Chemom. 32 (9) (sep 2018), e3044.

[15] I.S. Helland, S. Saebø, H.K. Tjelmeland, Near optimal prediction from relevant components, Scand. J. Stat. 39 (4) (mar 2012) 695–713.

[16] U. Indahl, A twist to partial least squares regression, J. Chemom. 19 (1) (2005) 32–44.

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

[18] I.T. Jolliffe, Principal Component Analysis, second ed., 2002.

[19] M. Lee, Z. Su, Renvlp: Computing Envelope Estimators. R Package Version 2.5, 2018. https://CRAN.R-project.org/package=Renvlp.

[20] B.-H. Mevik, R. Wehrens, K.H. Liland, Pls: Partial Least Squares and Principal Component Regression. R Package Version 2.7-0, 2018. https://CRAN.R-project.org/package=pls.

[21] T. Næs, I.S. Helland, Relevant components in regression, Scand. J. Stat. 20 (3) (1993) 239–250.

[22] T. Naes, H. Martens, Comparison of prediction methods for multicollinear data, Commun. Stat. Simulat. Comput. 14 (3) (jan 1985) 545–576.

[23] T. Næs, O. Tomic, N.K. Afseth, V. Segtnan, I. Måge, Multi-block regression based on combinations of orthogonalisation, pls-regression and canonical correlation analysis, Chemometr. Intell. Lab. Syst. 124 (2013) 32–42.

[24] A.C. Rencher, Methods of Multivariate Analysis, vol. 492, John Wiley & Sons, 2003.

[25] R. Rimal, T. Almøy, S. Sæbø, A tool for simulating multi-response linear model data, Chemometr. Intell. Lab. Syst. 176 (may 2018) 1–10.

[26] 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.

COMPARISON OF MULTI-RESPONSE ESTIMATION METHODS

Comparison of Multi-response Estimation Methods

Raju Rimal[a,*], Trygve Almøy[a], Solve Sæbø[a]

[a]*Faculty of Chemistry and Bioinformatics, Norwegian University of Life Sciences, Ås, Norway*

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.

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

*Corresponding Author

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

1. Introduction

Estimation of parameters in linear regression models is an integral part of many research studies. Research fields such as social science, econometrics, chemometrics, psychology and 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 (High-Level Expert Group on Artificial Intelligence, 2019). 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 (Rimal et al., 2019) 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) (Cook et al., 2010) and simultaneous estimation of the envelope (Senv) (Cook and Zhang, 2015). 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.

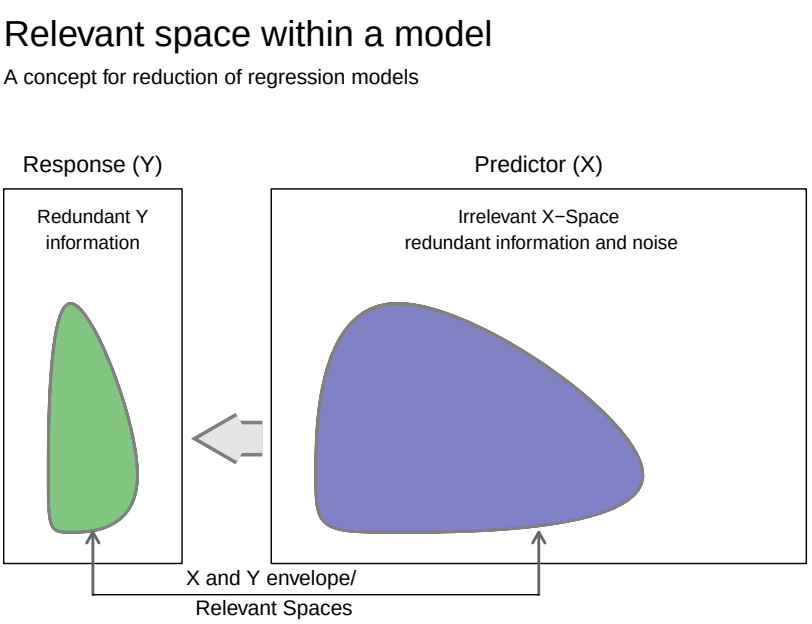


Figure 1: Relevant space in a regression model

2. Simulation Model

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

Following the concept of relevant space, a subset of predictor components can be imagined to span the predictor space. These components can be regarded as relevant predictor components. Naes and Martens (1985) introduced the concept of relevant components, which was explored further by Helland (1990), Næs and Helland (1993), Helland and

Almøy (1994) and Helland (2000). The corresponding eigenvectors were referred to as relevant eigenvectors. A similar logic is introduced by Cook et al. (2010) and later by Cook et al. (2013) as an envelope, as space spanned by the relevant eigenvectors (Cook, 2018, p.101). See Rimal et al. (2018), Sæbø et al. (2015) and Rimal et al. (2019) for in-depth background on the model.

3. Estimation Methods

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

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

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

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

where $\boldsymbol{\beta} = \boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}$ is the regression coefficients that define the relationship between $\mathbf{x}$ and $\mathbf{y}$. With n samples, the least-squares estimate of $\boldsymbol{\beta}$ can be written as $\hat{\boldsymbol{\beta}} = \mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}$. Here, 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 $\mathbf{y}$ or both. Some of the methods we have used for comparison are discussed below.

Principal Components Regression (PCR) uses k eigenvectors of $\mathbf{S}_{xx}$ as the number of components 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 (Jolliffe, 2002). 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 (Almøy, 1996).

Partial Least Squares (PLS) regression aims to maximize the covariance between the predictor and response components (scores) (de Jong, 1993). 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 (Wold, 1975), SIMPLS (de Jong, 1993) and KernelPLS (Lindgren et al., 1993), we will be using KernelPLS for this study, which gives equivalent results to the classical NIPALS algorithm and is default in R-package `pls` (Mevik and Wehrens, 2007).

Envelopes was first introduced by (Cook et al., 2007) 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 $\mathbf{\Sigma}_{xx}$ into relevant (material) and irrelevant (immaterial) parts, such that the response $\mathbf{y}$ is uncorrelated with the irrelevant part given the relevant one. In addition, relevant and irrelevant parts are also uncorrelated. Such separation of the covariance matrix is made using the data through the optimization of an objective function. Further, the regression coefficients are estimated using only 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 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 result will be equivalent to ordinary least squares. Cook and Zhang (2015) and Cook (2018) have discussed the method in detail. Further, Helland et al. (2018) have discussed how the population models of PCR, PLS and Xenv are equivalent.

4. Experimental Design

An R (R Core Team, 2018) package `simrel` (Rimal et al., 2018; Sæbø et al., 2015) is used to simulate the data for comparison. In the simulation, the number of observations is fixed at $n = 100$, 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 $\mathbf{\Sigma}_{xx}(\lambda_i, i = 1, \dots p)$ as,

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

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 $\mathbf{x}$ 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 we will use two different levels of a position index of predictor components (`relpos`): a) 1, 2, 3, 4 and b) 5, 6, 7, 8. Predictor components irrelevant for a response make prediction difficult (Helland and Almøy, 1994). When combined with multicollinearity, this factor can create both easy and difficult cases for both estimation and prediction.

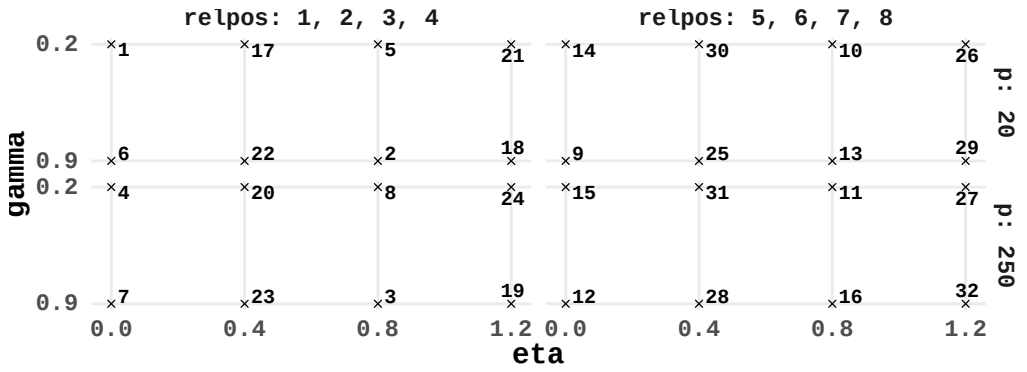


Figure 2: Experimental Design of simulation parameters. Each point represents an unique data property.

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,

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

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

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

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

datasets are simulated.

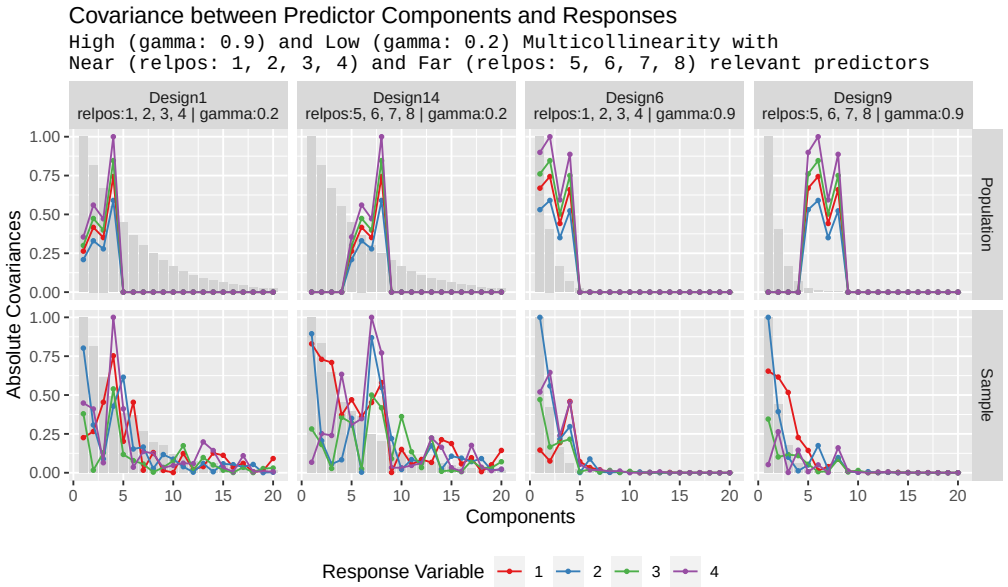


Figure 3: Covariance between predictor components and each response variable in the population (top), and in the simulated data (bottom) for four different designs. The bars in the background represent the variance of the corresponding components (eigenvalues).

The simulation properties are directly reflected in the simulated data. For example, in Figure 3, design pairs 1 and 14 as well as 6 and 9 differ in their properties only in terms of 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 Figure **??**). The combination of these factor levels creates datasets that are easy or difficult with regard to estimation and prediction. We observe from Figure 3 that it may be difficult to infer the structure of the latent relevant space of $\mathbf{x}$ 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. (2019) to compare the estimation performance of PCR, PLS1, PLS2, Xenv and Senv methods. The performance is measured on the basis of,

a) average estimation error computed as in (6)
b) the average number of components used by the methods to give minimum estimation error

Let us define the expected estimation error as

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

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

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

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

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

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. (2019). For a given estimation method, design, and response, the component that gives the minimum

estimation error averaged over all replicates is selected as,

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

Here we have skipped further indices on $\widehat{\boldsymbol{\beta}}$ for brevity. The estimation error $\widehat{\mathrm{MSE}_\circ \left(\widehat{\boldsymbol{\beta}} \right)}$ for every method, design and response corresponding to component $l_\circ$, computed as (7), 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 $j = 1, \ldots 4$ in the context of this dataset. Further, let the number of components that result in minimum estimation error in each replication and computed as (8), 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 $j = 1, \ldots 4$.

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

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 $\mathbf{u}$ and $\mathbf{v}$ matrices, respectively. The density of the scores in Figure 4 and Figure 5 correspond to the first principal component of $\mathbf{u}$ and $\mathbf{v}$, i.e. the first column of $\mathbf{t}_u$ and $\mathbf{t}_v$ respectively. Here higher scores correspond to larger estimation error and vice versa.

Figure 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. On the other hand, the

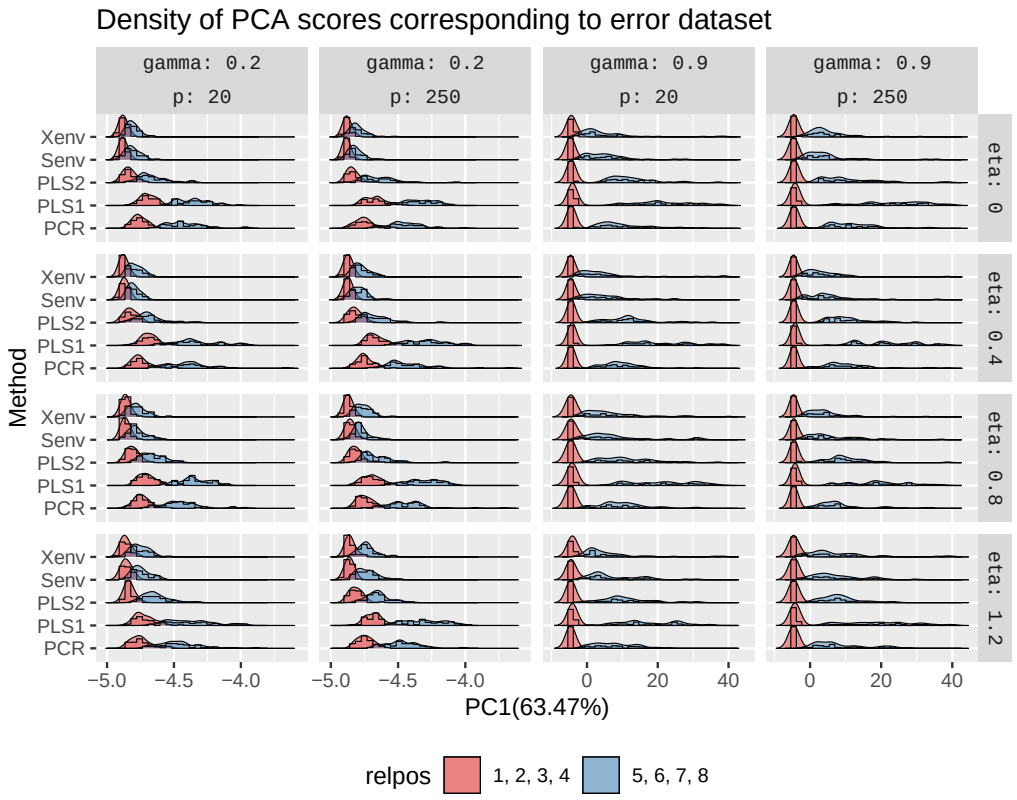


Figure 4: Scores density corresponding to first principal component of *error dataset* (**u**) subdivided by `methods`, `gamma` and `eta` and grouped by `relpos`.

envelope methods tend to have increased estimation error in cases of highly correlated responses (`eta: 1.2`), whereas there is no effect of this correlation in other methods. Furthermore, position of the relevant predictor components has a noticeable effect on estimation error 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 (Rimal et al., 2019).

In the case of the *component dataset* (Figure 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.

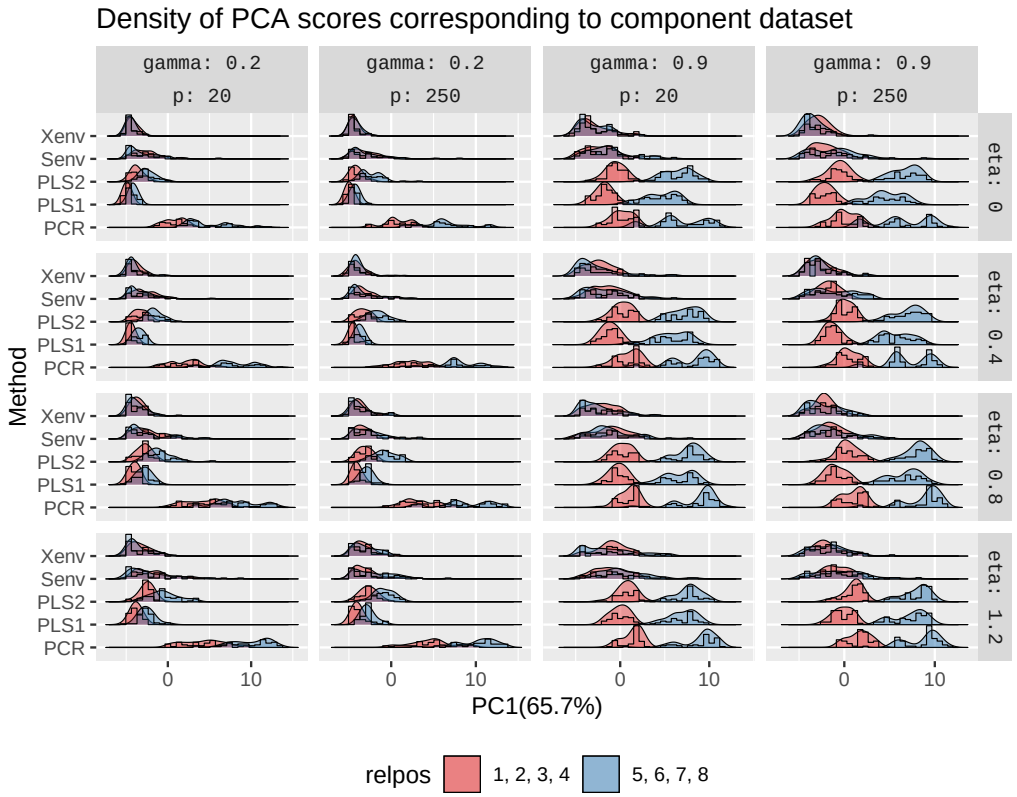


Figure 5: Score density corresponding to the first principal component of *component dataset* (**v**) subdivided by `methods`, `gamma` and `eta` and grouped by `relpos`.

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. (2019)), 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.

Figure 7 shows a clear distinction between the modelling approach of PLS2 and Senv methods for the same model based on Design 9 (top) and Design 29 (bottom). In both of the designs, PLS2 has both minimum prediction error and minimum estimation error obtained using seven to eight components and the estimated regression coefficients approximate the true coefficients. In contrast, the Senv method has approached the minimum prediction and minimum estimation error using one to two components and the corresponding estimated regression coefficients approximate the true coefficients (Figure 6). Despite having contrasted modelling results for a dataset with similar properties, the minimum errors produced by them are comparable in the case of Design 9 (See Table 1). However, in the case of Design 29, estimation error corresponding to PLS1 and envelope methods are much higher than PCR and PLS2. It is interesting to see that despite having large estimation error, the prediction error corresponding to the envelope methods are much

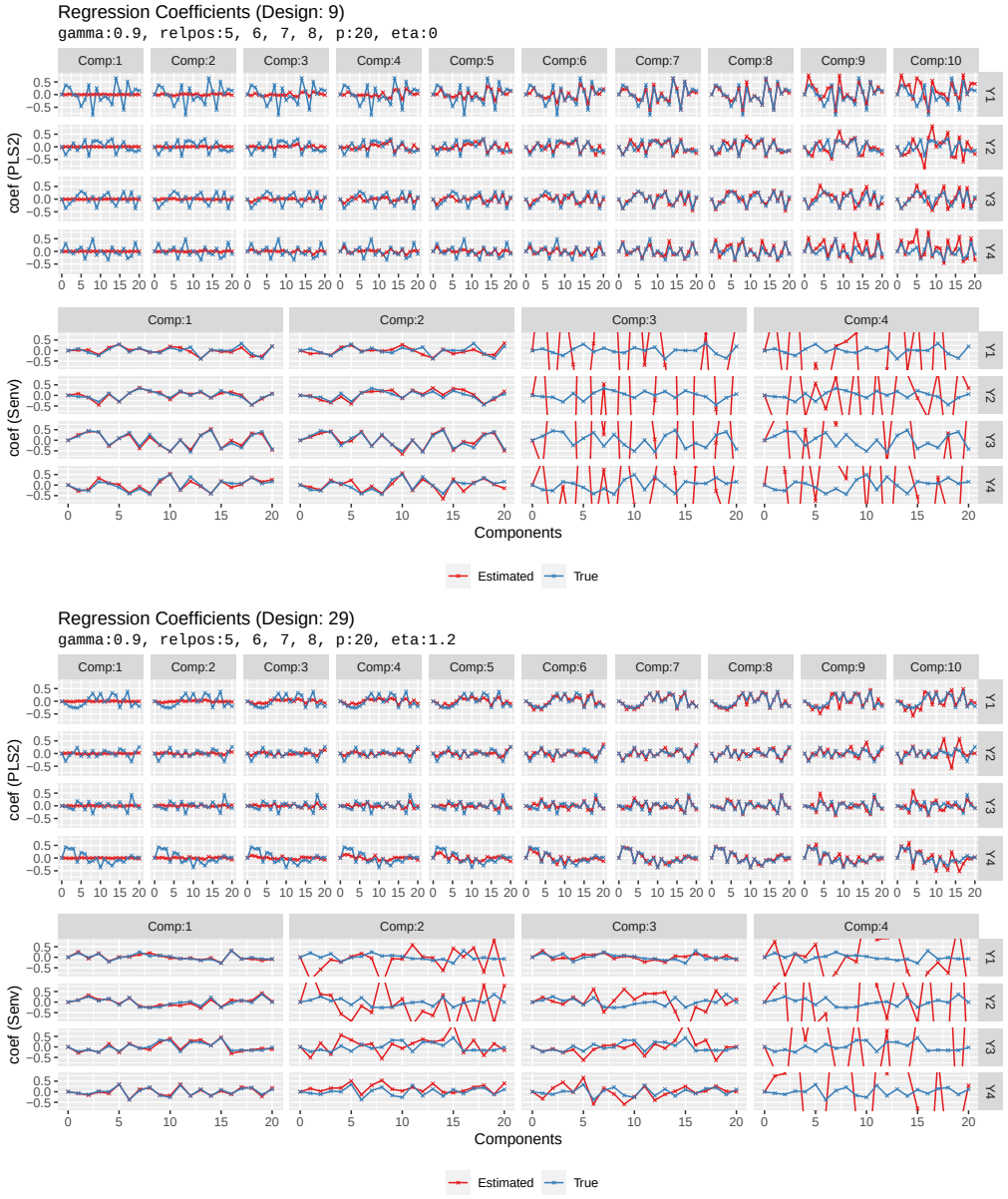


Figure 6: Regression Coefficients (coef) estimated by PLS2 and Simultaneous Envelope methods on the data based on Design 9 and 29.

smaller in this design.

Here the response dimension for the simultaneous envelope has been fixed at two components, which might have affected its performance, however, both envelope methods had performed much better with the same restriction in the case of prediction.

Figure 7 also shows 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).

In addition to the prediction and estimation error, Figure 6 gives a closer view of how the average coefficients corresponding to these methods approximate to the true values. Here 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 components in these methods due to a highly unstable and large error in non-optimal cases.

Despite having a large variation in prediction and estimation error, the envelope based methods have produced a better result even 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* (9) and the *component model* (10) in the

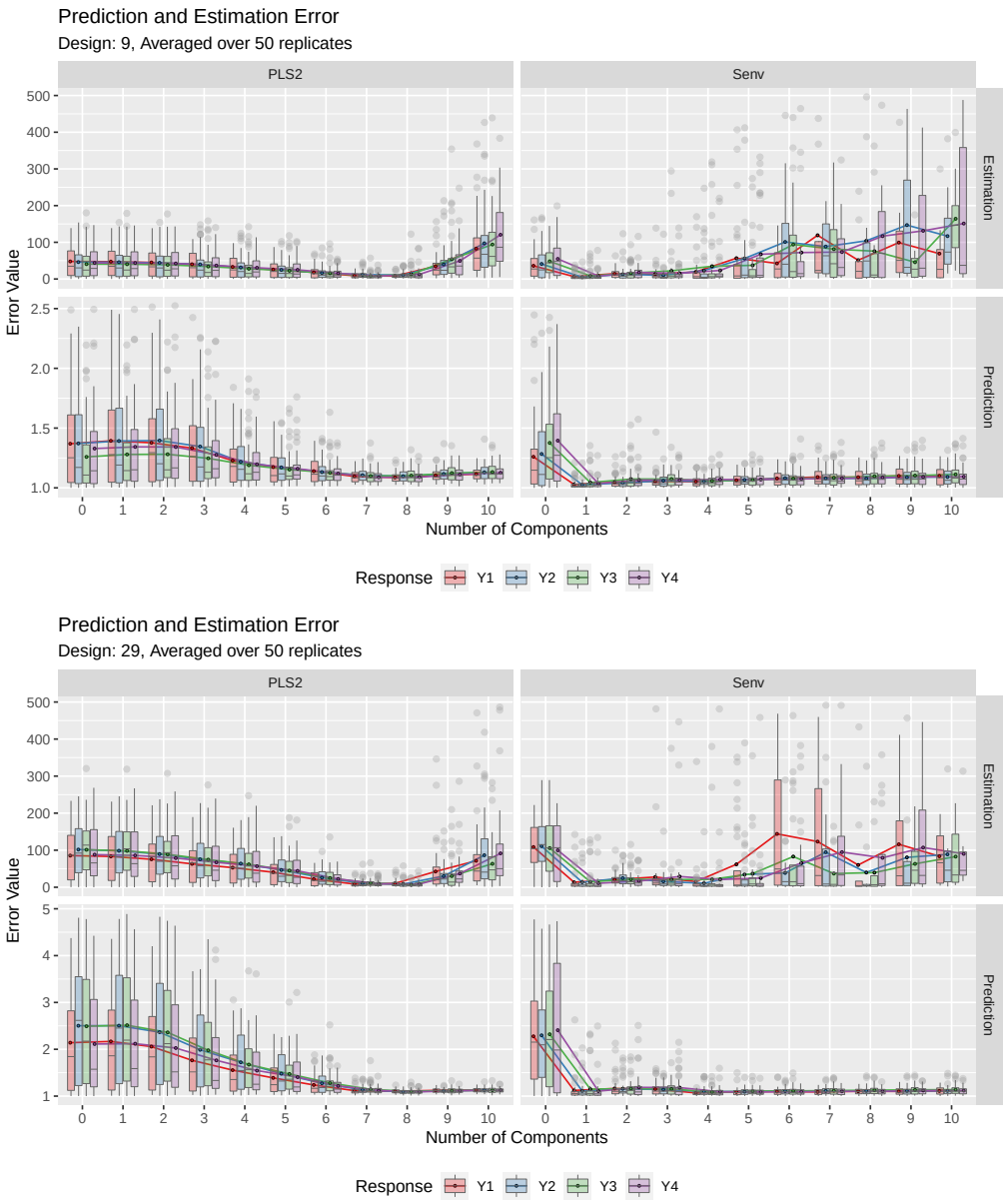


Figure 7: Minimum prediction and estimation error for PLS2 and Simultaneous Envelope methods. The point and lines are averaged over 50 replications.

Table 1: Minimum Prediction and Estimation Error for Design 9

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)

following. In the MANOVA model, we will consider the interaction of simulation parameters (`p`, `gamma`, `eta`, and `relpos`) and `Method` The models are fitted using correspondingly the *error dataset* (**u**) and the *component dataset* (**v**).

Error Model:

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

Component Model:

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

where, **u** corresponds to the estimation errors in *error dataset* and **v** corresponds to the

number of components used by a method to obtain minimum estimation error in the *component dataset*.

To make the analysis equivalent to Rimal et al. (2019), we have also used Pillai's trace statistic for accessing the result of MANOVA. Figure 8 plots the Pillai's trace statistics as bars with corresponding F-values as text labels. The 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.

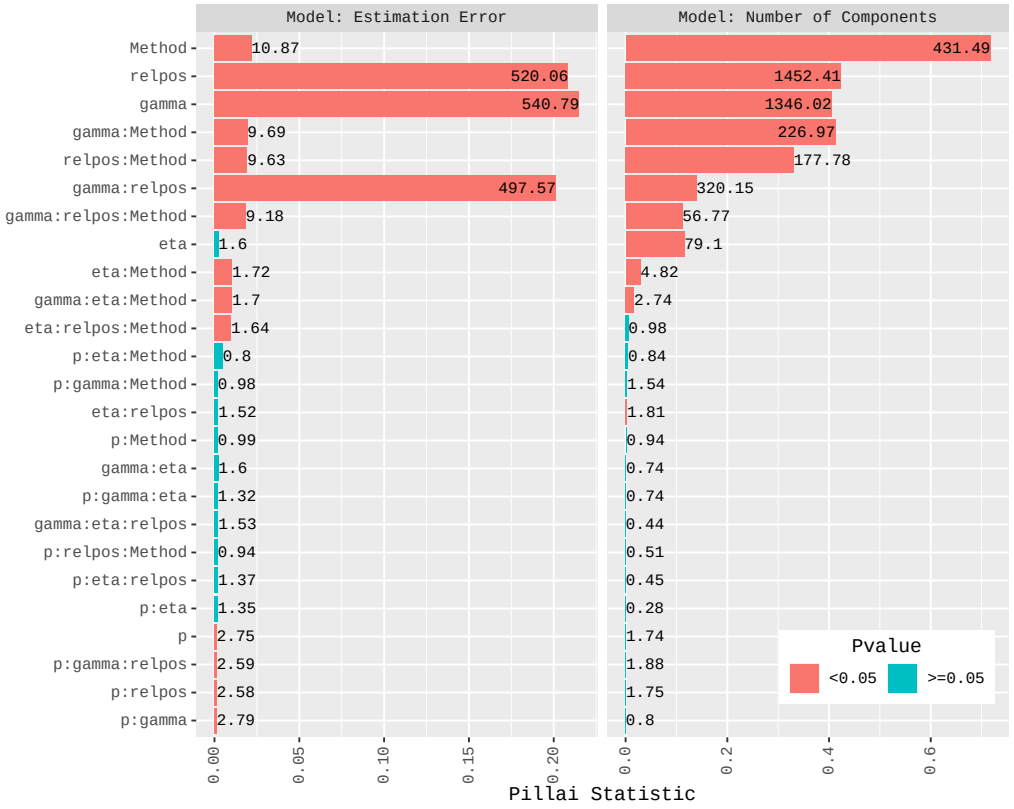


Figure 8: Pillai Statistic and F-value for the MANOVA model. The bar represents the Pillai Statistic and the text labels are F-value for the corresponding factor.

Error Model: Unlike for the prediction error in Rimal et al. (2019), `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 (Figure 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 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 Figure 8 (right) shows that the methods are significantly different and has a huge effect on the number of components they use to obtain the minimum estimation error. The result also corresponds to the case of prediction error in Rimal et al. (2019). However, the F-value corresponding the `relpos` and `gamma` shows that the importance of these factors is much stronger compared to the case of prediction error.

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

7.1. Effect Analysis of the Error Model

In figure 9 (left), the effect of correlation between the responses controlled by the `eta` parameter has a clear influence on the estimation error for the envelope methods. 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.

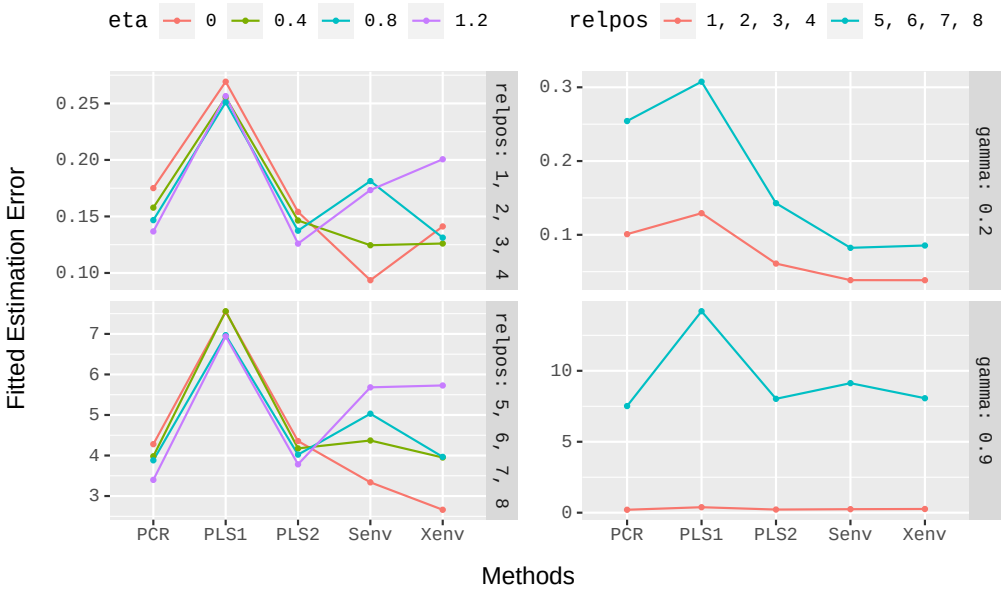


Figure 9: Effect plot of some interactions of the MANOVA corresponding to fitted *error model*

Figure 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 Figure 10. As seen previously, PCR uses, in general, a larger number of components than the other methods.

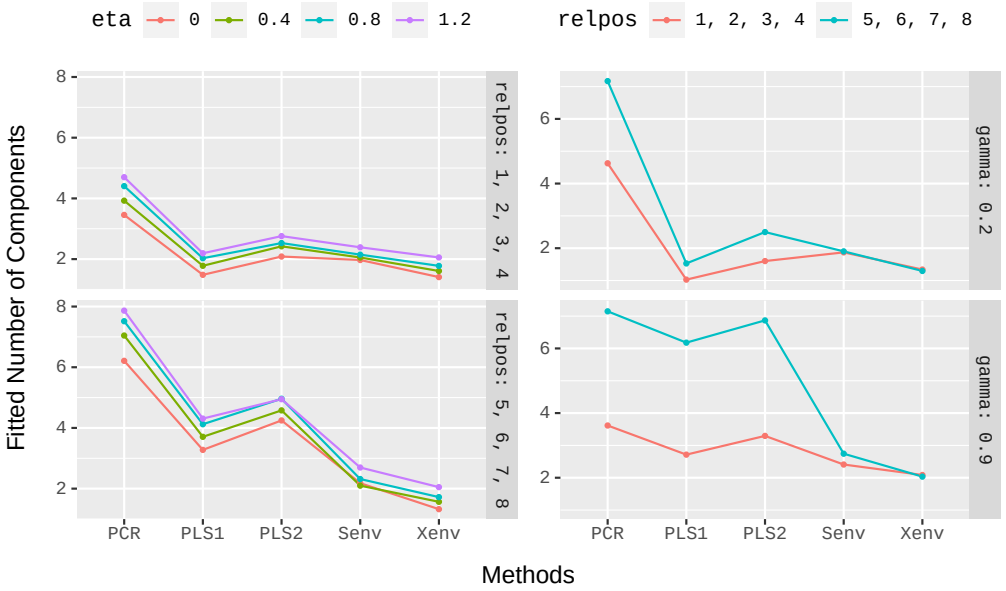


Figure 10: Effect plots of some interactions of the multivariate linear model corresponding to the *component model*.

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. Despite the interaction of the `eta` parameter with the method is significant, the extent of its effect is rather small compared to both main and interaction effect of `gamma` and `relpos`.

The effect of the number of variables is negligible in all cases for all designs. Here the use of principal components for reducing the dimension of $n < p$ designs, as in Rimal et al. (2019), 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. However, non-optimal number of components can lead to large estimation error, so one should be careful in this respect while using the envelope methods.

Both PLS1 and PLS2 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 (Rimal et al., 2019) on prediction comparison should be considered to obtain a comprehensive view of this comparison. A shiny (Chang et al., 2018) web application at http://therimalaya.shinyapps.io/Comparison allows readers to explore all the visualizations for both prediction and estimation comparisons. In addition, a GitHub repository at https://github.com/therimalaya/04-estimation-comparison can be used to reproduce this study.

References

Almøy, T., jan 1996. A simulation study on comparison of prediction methods when only a few components are relevant. Computational Statistics & Data Analysis 21 (1), 87–107.

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

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

Cook, R. D., Helland, I. S., Su, Z., 2013. Envelopes and partial least squares regression. Journal of the Royal Statistical Society. Series B: Statistical Methodology 75 (5), 851–877.

Cook, R. D., Li, B., Chiaromonte, F., aug 2007. Dimension reduction in regression without matrix inversion. Biometrika 94 (3), 569–584.

Cook, R. D., Li, B., Chiaromonte, F., 2010. Envelope Models for Parsimonious and Efficient Multivariate Linear Regression. Statistica Sinica 20 (3), 927–1010.

Cook, R. D., Zhang, X., 2015. Simultaneous envelopes for multivariate linear regression. Technometrics 57 (1), 11–25.

de Jong, S., mar 1993. SIMPLS: An alternative approach to partial least squares regression. Chemometrics and Intelligent Laboratory Systems 18 (3), 251–263.

Helland, I. S., 1990. Partial least squares regression and statistical models. Scandinavian Journal of Statistics 17 (2), 97–114.

Helland, I. S., mar 2000. Model Reduction for Prediction in Regression Models. Scandinavian Journal of Statistics 27 (1), 1–20.

Helland, I. S., Almøy, T., 1994. Comparison of prediction methods when only a few components are relevant. Journal of the American Statistical Association 89 (426), 583–591.

Helland, I. S., Saebø, S., Almøy, T., Rimal, R., Sæbø, S., Almøy, T., Rimal, R., sep 2018. Model and estimators for partial least squares regression. Journal of Chemometrics 32 (9), e3044.

High-Level Expert Group on Artificial Intelligence, 2019. Ethics guidelines for trustworthy ai. Tech. rep., The European Commission.

Jolliffe, I. T., 2002. Principal Component Analysis, Second Edition.

Lindgren, F., Geladi, P., Wold, S., Jan 1993. The kernel algorithm for pls. Journal of Chemometrics 7 (1), 45–59.
URL http://dx.doi.org/10.1002/cem.1180070104

Mevik, B.-H., Wehrens, R., 2007. Theplspackage: Principal component and partial least squares regression inr. Journal of Statistical Software 18 (2), nil.
URL https://doi.org/10.18637/jss.v018.i02

Næs, T., Helland, I. S., 1993. Relevant components in regression. Scandinavian Journal of Statistics 20 (3), 239–250.

Naes, T., Martens, H., jan 1985. Comparison of prediction methods for multicollinear data. Communications in Statistics - Simulation and Computation 14 (3), 545–576.

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., may 2018. A tool for simulating multi-response linear model data. Chemometrics and Intelligent Laboratory Systems 176, 1–10.

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

Wold, H., 1975. Soft modelling by latent variables: the non-linear iterative partial least squares (nipals) approach. Journal of Applied Probability 12 (S1), 117–142.
URL https://doi.org/10.1017/s0021900200047604