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

Thesis Number: 1234:56
ISSN: 1234-5678
ISBN: 123-45-678-1234-5

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, JUL © RAJU RIMAL

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 variations are relevant for regression, i.e. only a certain amount of variations in predictors are relevant for a part of variations in 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 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 on 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.

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 alle variasjoner relevante for regresjon, dvs. bare en viss mengde variasjoner i prediktorer er relevante for en del av variasjoner i respons. Dette fører til en reduksjon av den lineære regresjonsmodellen der man kan forestille seg et underområde av plassen som spennes av prediktorvariablene som inneholder all relevant informasjon for et underområde spandert av responsvariablene.

I denne avhandlingen prøver vi å sammenligne noen nye metoder som er basert på konvoluttmodellen og noen etablerte metoder som hovedkomponenter 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 designet egenskaper. For simuleringen har vi også laget en R-pakke kalt `simrel` med et webgrensesnitt.

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

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 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 insight, suggestion and comment 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 of 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 to simulated data from multi-response linear model. The package can be install from R-package repository CRAN or from GitHub (https://github.com/simulatr/simrel). 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 softwares and packages required are 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 list 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 heap of data has become another problem. Many statistical and machine learning tools devised for this purpose, Most of which, focus to identify the relationships between different variables. A linear relationship is the most common one. This thesis confined itself in the exploration of linear relationship 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 predictor and response are termed as 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 predictor and the response. 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 defind 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.

A latent components corresponding to predictor variables, which we will refer to as “predictor components”, are the linear combination of the predictor variables. Naes and Martens [1985] and later by Helland [1990] and Næs and Helland [1993] have defined a set of predictor components as irrelevant components such that they have no correlation with the response variables and the relevant part. Using only a subset of the latent component for modeling or other analysis is often termed as “dimension reduction”. Methods like PCR, PLS and many other variants of PLS has leveraged this concept and are serving as a prime tool in many disciplines, most notably in chemometrics.

Relatively newer 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 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 estimates the model parameters differently. Model parameters are the unknowns which helps to define a complex relationships between the variables. Regression coefficients (β) in (2) is an example of model parameter. All methods uses data to estimate these parameters. So, the pros and cons of a dataset affect the estimation of a method and consequently their prediction performance. 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 add 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 μ_y and another random variable-vector $\mathbf{x}$ of p predictor variables with mean μ_x as,

$$\begin{bmatrix}\mathbf{y}\\ \mathbf{x}\end{bmatrix} \sim \mathsf{N}\left(\begin{bmatrix}\mu_y\\ \mu_x\end{bmatrix}, \begin{bmatrix}\Sigma_{yy} & \Sigma_{yx}\\ \Sigma_{xy} & \Sigma_{xx}\end{bmatrix}\right) \tag{1}$$

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

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

$$\mathbf{y} = \mu_y + \beta^t(\mathbf{x} - \mu_x) + \varepsilon \tag{2}$$

where $\varepsilon \sim \mathsf{N}\left(\mathbf{0}, \Sigma_{y|x}\right)$

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

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

$$\begin{aligned}\begin{bmatrix}\mathbf{w}\\ \mathbf{z}\end{bmatrix} \sim \mathsf{N}(\mu, \Sigma) &= \mathsf{N}\left(\begin{bmatrix}\mu_w\\ \mu_z\end{bmatrix}, \begin{bmatrix}\Sigma_{ww} & \Sigma_{wz}\\ \Sigma_{zw} & \Sigma_{zz}\end{bmatrix}\right)\\ &= \mathsf{N}\left(\begin{bmatrix}\mathbf{Q}\mu_y\\ \mathbf{R}\mu_x\end{bmatrix}, \begin{bmatrix}\mathbf{Q}\Sigma_{yy}\mathbf{Q}^t & \mathbf{Q}\Sigma_{yx}\mathbf{R}^t\\ \mathbf{R}\Sigma_{xy}\mathbf{Q}^t & \mathbf{R}\Sigma_{xx}\mathbf{R}^t\end{bmatrix}\right)\end{aligned} \tag{3}$$

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,

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

Here, $\mathbf{\Sigma}_{zz}$ and $\mathbf{\Sigma}_{ww}$ are diagonal matrices of eigenvalues corresponding to predictors and responses respectively. Following the concept of relevant components $\mathbf{\Sigma}_{wz} = \mathbf{\Sigma}_{zw}^t$ has non-zero columns (rows) with relevant components. With some random orthogonal rotation matrix $\mathbf{R}$ and $\mathbf{Q}$, which can be easily generated, the unknowns required for simulation drastically decrease. Following the idea from Sæbø et al. [2015], Paper I uses exponential decay of eigenvalues as in (5) that fills of the diagonal of $\mathbf{\Sigma}_{zz}$ and $\mathbf{\Sigma}_{ww}$. 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 following the concept of "relevant components" can be found in Paper I. 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 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 response space as well. The relevant and irrelevant space 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.

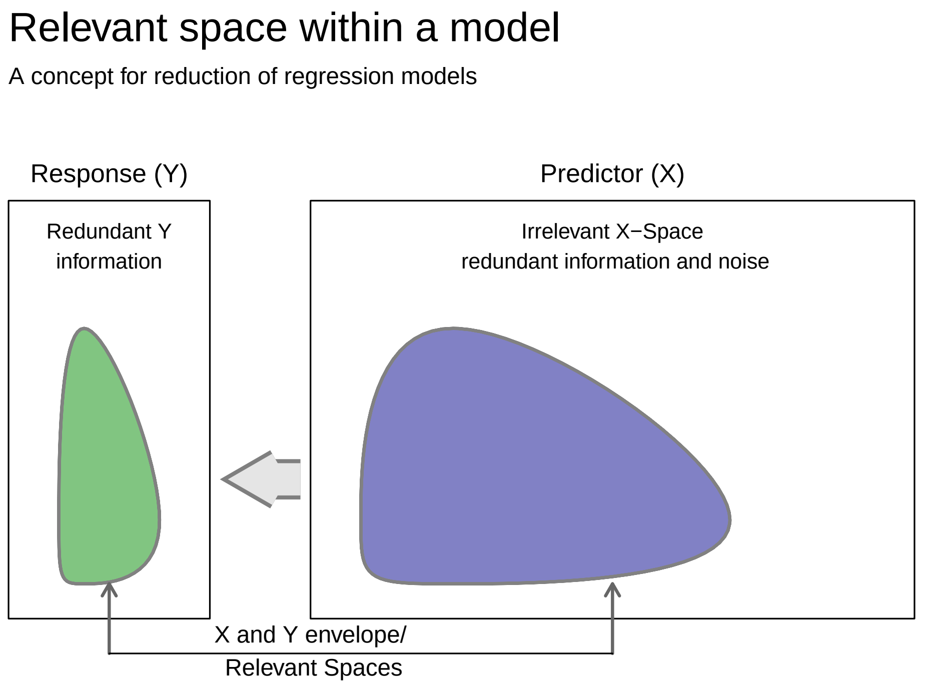
Relevant space within a model
A concept for reduction of regression models
Response (Y)
Predictor (X)
Redundant Y information
Irrelevant X−Space redundant information and noise
X and Y envelope/ Relevant Spaces

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. Let 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 (**x**). These four components are the relevant components. Consider two cases:

CASE 1 (FIGURE 2, LEFT): The position of these relevant components are at 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.

CASE 2 (FIGURE 2, RIGHT): The position of these 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.

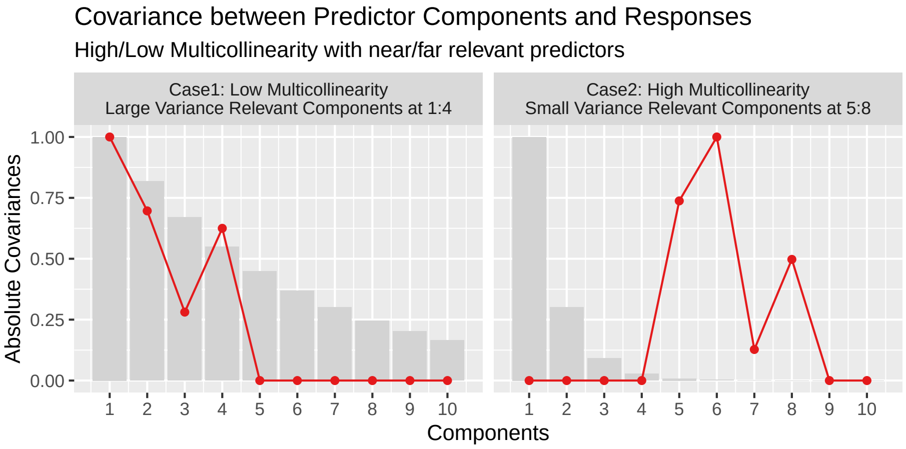

Figure 2: Relevant components at two different set of positions and two different levels of multicollinearity. The points represents the correlation of predictor components and the response variable.

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 **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 reducting from the first component and stabilize after the fourth component while in the second case, prediction error only starts reducting 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 a fewer number of components to get the minimum prediction error than the 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.

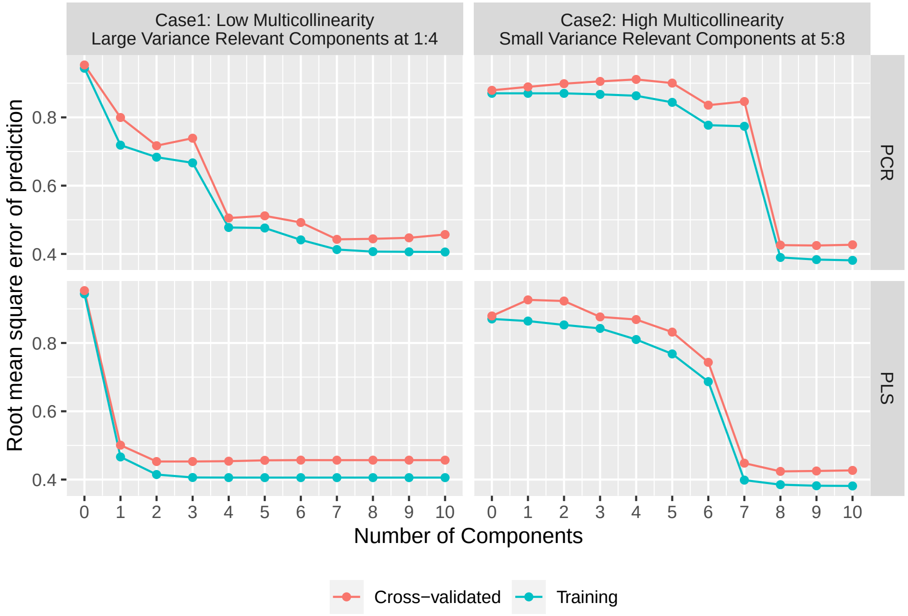

Figure 3: Root mean square error of cross-validation from PCR and PLSR

The concept of relevant components also can 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 by generating 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. Apart from many other methods, one can use the central limit theorem to simulate standard normal variates from the uniform random variates. With standard normal variate $Z \sim N(0, 1)$, 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 also can be widely 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 introduced in 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 parameters 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. The estimated value is considered better if the expected squared difference between the estimated and true value is small and has small variance. The goodness of the estimates depends on the nature of the data and the method that is used to estimate them. Estimation error with true and estimated regression coefficient $\boldsymbol{\beta}$ and $\widehat{\boldsymbol{\beta}}$ respectively, can be defined as in (6).

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

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).

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

From (6) and (7), we can see that the prediction errors are influenced by the covariance of the predictor 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 comparison 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] perform dimension reduction only in the response space while Predictor envelope (Xenv) [Cook et al., 2013] perform dimension reduction only in the predictor space. The simultaneous envelope (Senv) [Cook and Zhang, 2015b] perform 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 has shown encouraging results for envelope methods in both easy and difficult models.

PLS and its derivatives

Since the PLS method has been both popular and productive in the field 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 gives 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 Bayes PLS method. The method only works with a single response model and have 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 less number of 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 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 comparison, simulation parameters are considered as factors and the prediction and estimation error are considered as outcome variables. 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 through and rigorous comparison possible.

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 predictors are at 1, 2, 3, 4 while Design 9 have high multicollinearity and the position index of relevant predictors are at 5, 6, 7, 8. All the other factors or properties of the data being the same for both, the difference is these two

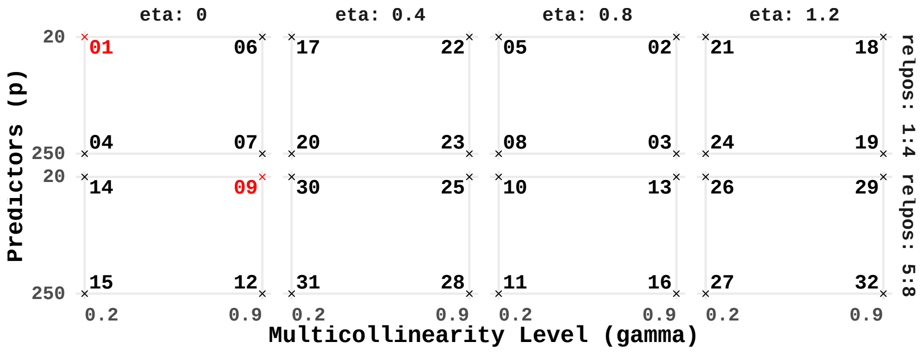

Figure 4: An example of a factorial design used in the third and fourth paper.

design helps us to analyse the interaction between the multicollinearity in the data and the position of relevant components.

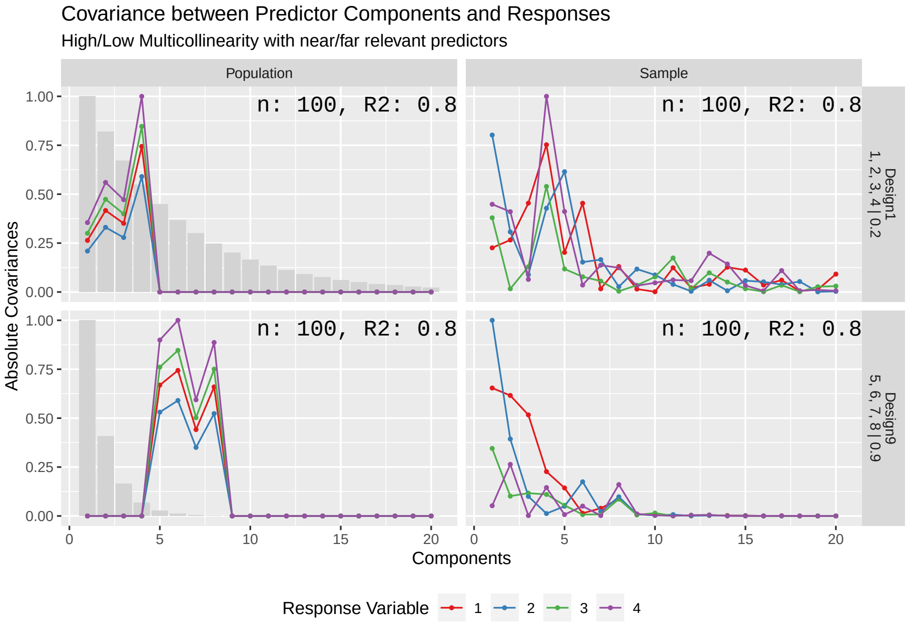

Figure 5: *Design 1*: Relevant components have large variation, *Design 9*: irrelevant components have large variation and relevant components have small variation.

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 has small variation and irrelevant components at position 1, 2, 3, 4 have large variation. This design simulates data that are difficult to model my most methods. The population covariances in the figure give clear and distinct relationship while the sample covariances have a somewhat rough approximation of the population.

Analysis of Variance

Various exploratory plots of prediction error, estimation error and the number of components used by different methods are used for analysis. 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 analyse the effect of the interaction of these properties with the methods. The third and fourth paper uses 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 is 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 matrix. 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 the simulation and compare some multivariate methods using the simulated data. Additionally, it also includes an R-package called `simrel` that has been 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 also 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 multicollinearity problem. The second example compared PLS1 and PLS2 where, on most occasions, the latter dominates the earlier with regards 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 show 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 BayesPLS 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 do the dimension reduction 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 using principal components. This paper also illustrates the use of principal components for using 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 them is higher than for 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. Low multicollinearity with independent response variables is in favour of envelope methods as they have both smaller prediction and estimation error using fewer components. 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 as well. Both prediction and estimation error from PCR is more stable than other methods, while as PLS1 method models each response separately, the performance is in general poorer than for 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 simulate data easily 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. Although the Bayes PLS method has shown the best performance in these simulation results, its performance in the real data was satisfactory. This pointed us to explore the methods comprehensively. However, due to the time-consuming computation and since the 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.

Since prediction and estimation each have many aspects to be discussed, we 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 the estimation and the prediction. These factors were used as simulation parameters in addition to a factor that controls the correlation between the response variables. The response correlation and its interaction with these methods and other simulation parameters have limited studies. The study of response correlation, its interaction with other factors and different methods 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 when a non-optimal number of components is used. 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 not only to evaluate the methods but also 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 as well. 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. However we did some basic comparison by including them, but they require a separate and a more comprehensive study.

The study is 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 needs 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

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 Journal of 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.

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

COMPARISON OF MULTI-RESPONSE ESTIMATION METHODS

Comparison of Multivariate 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, 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)

Preprint submitted to Elsevier *July 9, 2019*