Universidade do Minho
Escola de Ciências

Shirin Shahriari

Variable Selection in Linear Regression Models with Large Number of Predictors

Outubro de 2014

Universidade do Minho
Escola de Ciências

Shirin Shahriari

Variable Selection in Linear Regression Models with Large Number of Predictors

Tese de Doutoramento
Programa Doutoral de Matemática e Aplicações

Trabalho efectuado sob a orientação de
Professora Susana Faria
Professora A. Manuela Gonçalves

Outubro de 2014

STATMENT OF INTEGRITY

I hereby declare having conducted my thesis with integrity, I confirm that I have not used plagiarism or any form of falsification of results in the process of thesis elaboration.
I further declare that I have fully acknowledge the Code of Ethical Conduct of the University of Minho.

University of Minho, ______________________________

Full name: ____

Signature: ____

Acknowledgments

This work was funded by the Portuguese Foundation for Science and Technology (FCT) under the grant SFRH/BD/51164/2010, gratefully acknowledged.

I would like to express my sincere gratitude to my supervisor professor Susana Faria for her inspiration and excellent guidance. Also, I offer my special thanks to my supervisor professor A. Manuela Gonçalves for her guidance and generous support throughout this work.

I would like to thank professor Stefan Van Aelst (KU Leuven, Belgium) for his time and generous suggestions throughout this study, and for his help during my stay in Belgium.

My respectful thanks are to my loving parents Ebrahim Shahriari and Shahin Ahmarirad for their constant encouragement and support in all stages of my studies.

I would like to warmly thank my brother Shahriar and my sister Shahla whose emotional support and assistance have been beyond what one could expect.

I am grateful to my husband, Nima, for his care, support, and constant love during the preparation of this dissertation. His encouragement has facilitated for me writing this thesis.

POPH
PROGRAMA OPERACIONAL **POTENCIAL HUMANO**

QREN
QUADRO
DE REFERÊNCIA
ESTRATÉGICO
NACIONAL
PORTUGAL 2007.2013

UNIÃO EUROPEIA
Fundo Social Europeu

FCT
Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR

To my loving parents, Shahin and Ebrahim

Abstract

In this thesis, we study the problem of variable selection in linear regression models in the presence of a large number of predictors. Usually, some of these predictors are correlated, so including all of them in a regression model will not essentially improve the model's predictive ability. Also, models with reasonable and tractable amount of predictors are easier to interpret than models with a large number of predictors. Therefore, variable selection is an important problem to study. Given that there are some popular regression methods capable of handling collinearity in data but still requiring the removal of irrelevant predictors, so we present an algorithm that enable these methods to perform variable selection. We review the well-known variable selection methods, and investigate the performance of these methods as well as the proposed approach on both simulated and real data sets. The results show that the new algorithm performs well in selecting the relevant variables.

Also, when the data contains outliers, outlier detection and variable selection are not two separable problems. Therefore, we propose a method capable of outlier detection and variable selection. We review the well-known robust variable selection methods and evaluate the performance of these methods with the proposed approach on contaminated simulation data sets as well as on real data. The results show that the proposed method performs well concerning both outlier detection and robust variable selection.

Keywords: Bootstrap, least angle regression (LARS), linear regression, partial least squares regression (PLSR), principal components regression (PCR), outlier detection, variable selection

Resumo

Nesta dissertação foi estudado o problema da seleção de variáveis em modelos de regressão linear, na presença de um grande número de variáveis explicativas ou preditoras, em que usualmente, algumas das variáveis explicativas estão correlacionadas. Um princípio a ser levado em consideração é o "princípio da parcimónia": modelos mais simples devem ser escolhidos aos mais complexos, desde que a qualidade do ajustamento/previsão seja similar. Estes modelos são mais fáceis de interpretar do que os modelos com um grande número de preditores. Portanto, o estudo de métodos de seleção de variáveis é um problema muito importante em modelos de regressão. Dado que existem alguns métodos de regressão, já bem conhecidos, capazes de lidar com a multicolinearidade entre os dados, mas ainda não removendo os preditores irrelevantes, apresentamos um algoritmo que permite realizar a seleção de variáveis.
São estudados métodos de seleção de variáveis e investigados os desempenhos desses métodos, bem como o desempenho do algoritmo proposto, com dados simulados e com dados reais. Os resultados mostram que o novo algoritmo tem um bom desempenho na seleção das variáveis relevantes para o modelo. Além disso, quando os dados contêm valores atípicos, a detecção de *outliers* e a seleção de variáveis não podem ser estudados como dois problemas separáveis. Assim, nesta dissertação foi proposto um método capaz de deteção de *outliers* e de seleção de variáveis, em simultâneo.
Foram estudados os métodos de seleção de variáveis robustos mais conhecidos, de forma a avaliar e comparar o desempenho desses métodos com a abordagem proposta neste trabalho com estudos de simulação em situações de contaminação, bem como com dados reais. Os resultados mostram que o método desenvolvido tem um bom desempenho tanto em termos de detecção de *outliers*, assim como na seleção robusta de variáveis.

Palavras-chave: *Bootstrap*, deteção de *outliers*, *least angle regression* (LARS), *partial least squares regression* (PLSR), *principal components regression* (PCR), regressão linear, seleção de variáveis

Contents

1 Introduction 1

1.1 Variable Selection 2

1.1.1 Need for Variable Selection 3

1.2 Outlier Detection and Robust Variable Selection 4

1.3 Thesis Overview 4

1.3.1 Objectives and Achievements 4

1.3.2 Variable Selection Ability 5

1.3.3 Outlier Detection and Robust Variable Selection Development 6

1.3.4 Thesis Structure 7

2 Review: Variable Selection in Linear Regression 9

2.1 Introduction 9

2.2 Potential Subset Selection 13

2.2.1 All Possible Subset Selection 13

2.2.2 Forward Stepwise and Backward Elimination Selection 14

2.2.3 Forward Stagewise Regression 14

2.3 Shrinkage Methods 15

2.3.1 Ridge Regression 15

2.3.2 Least Absolute Shrinkage and Selection Operator 16

2.4 Modified Versions of Stepwise Regression 18

2.4.1 Least Angle Regression 18
2.4.2 Variance Inflation Factor (VIF) 19
2.5 Discussion and Conclusions 21

3 Sparsity for Principal Component and Partial Least Squares Regression 25

3.1 Introduction 25
3.2 Principal Components Regression and Partial Least Squares Regression 28
3.2.1 Principal Components Regression Algorithm 31
3.2.2 Partial Least Squares Regression Algorithm 36
3.3 Resampling Methods 40
3.3.1 Jack-knife 40
3.3.2 Bootstrap 41
3.4 Variable Selection for PCR and PLSR 43
3.4.1 Sparse Formulation Based Methods 44
3.4.2 Jack-knife PLSR/PCR Algorithm 46
3.4.3 Bootstrap PLSR/PCR Algorithm 47
3.5 Simulation Study 49
3.5.1 Performance Measurements 53
3.6 Simulation Study Results 54
3.7 Application to Real Data Set 59
3.8 Discussion and Conclusions 60

4 Outlier Detection and Robust Variable Selection 63

4.1 Introduction 63
4.2 Robust Variable Selection Methods 67
4.2.1 Robust Least Angle Regression 68
4.2.2 Robust Variance Inflation Factor (RobVIF) Regression 71
4.3 Jack-knife Robust LARS (JKRLARS) Algorithm 72

4.4 Simulation Study 75

4.4.1 Performance measurements 77

4.4.2 Simulation Study Results 78

4.5 Application to Real Data 81

4.6 Discussion and Conclusions 82

5 Conclusions and Future Work 85

5.1 Conclusions 85

5.2 Future Developments 87

Bibliography 89

List of Tables

3.1 TPR, FNR and RP values for each variable selection method for models. 58

3.2 $RMSE$, $RPred$ values and the number of selected predictors ($Pred$) for each method on yarn data. 59

4.1 The true positive rate (TPR) and the false negative rate (FNR) averaged over 200 simulation runs, are reported for JKRLARS. 81

List of Figures

2.1 Estimation picture for LASSO (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions $|\beta_1| + |\beta_2| \leq t$ and $\beta_1^2 + \beta_2^2 \leq t^2$, respectively, while the red ellipses are the contours of the least squares error function [Tibshirani (1996)]. 17

3.1 PCR and PLSR schematic representation. The left hand plot shows the first stage of the bilinear model. The red arrows represent the direction of the first 2 components. The general process of PCR and PLSR are illustrated in the right plot. 30

3.2 Flowchart summarizes BootPLS algorithm. 50

3.3 Comparison between the predictive ability of PLSR, PCR, JKPLS, JKPC, BootPLS, BootPC, SPLS, and SPC for Model 1. 55

3.4 Comparison between the predictive ability of PLSR, PCR, JKPLS, JKPC, BootPLS, BootPC, SPLS, and SPC for Model 2. 56

3.5 Comparison between the predictive ability of PLSR, PCR, JKPLS, JKPC, BootPLS, BootPC, SPLS, and SPC for Model 3. 56

4.1 Limitation of separate univariate winsorizations ($c = 2$). The correlation outliers (red points) are only shrunken to the boundary of the square(Khan *et al.* (2007)). 69

4.2 Adjusted winsorization (for initial estimate $\boldsymbol{R}_0$) with $c_1 = 2,\ c_2 = \sqrt{h}c_1$ (Khan *et al.* (2007)). 70

4.3 Average number of target variables t_m versus m for LARS, RLARS, RobVIF, and JKRLARS for scenarios (a)-(e). The lines shown in all plots follow the legend of figure (a). 80

4.4 Learning curves for LARS, RLARS, BRLARS, RobVIF, and JKRLARS on Census data. Each learning curve suggests the optimal number of predictors for each method. 83

4.5 Box plots of MAPE values of the optimal number of selected variables over 10 test subsets with 10 times replication for each method. . . . 84

Chapter 1

Introduction

Regression models have been widely used in many different areas such as health, biology, environment, management, and etc. Motivated by various applications, there has been a dramatic growth in the automated means of data collection, yielding data sets with a huge number of observations and larger numbers of potentially relevant predictor variables than before.

Usually, there are some correlated predictor variables, and including all of them in a statistical model will not necessarily improve the model's prediction performance. On the other hand, the interpretation of models with fewer predictors is easier and more tractable than for models with a large number of predictors. Therefore, finding the best variables among all candidate predictor variables is an important and challenging problem to study.

When the data contains atypical observations such as outliers, we need a robust variable selection method that is resistant to outliers in order to select variables reliably. In this situation, when the data contains atypical observations, outlier detection and variable selection are inseparable problems. Therefore, a robust method that can simultaneously detect outliers and select variables is needed.

In this thesis, we are interested in the following problems

- selecting the relevant predictor variables for regression models with large number of predictors;
- detecting outliers and robustly selecting the relevant predictor variables in the presence of outliers in data sets with large number of predictors.

Given that there are some regression methods which are well-known and popular to reduce the dimension of the data but still requiring the elimination of irrelevant predictor variables, we plan to present a variable selection procedure for these regression methods in the presence of large number of predictor variables in data.

We will also present an outlier detection and robust variable selection method using computationally efficient regression methods in dealing with contaminated data sets with large number of predictors.

1.1 Variable Selection

Consider the design matrix $\mathbf{X} = [\mathbf{1}\ \boldsymbol{x}_1 \ldots \boldsymbol{x}_p]$ with p predictor variables and a response variable $\mathbf{y}$, with n observations, the classical linear regression model supposes

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{1.1}$$

with parameters $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_p]^T$. The errors $\boldsymbol{e} = [e_1,\ e_2, \ldots, e_n]^T$ are assumed to have $E(\boldsymbol{e}) = \mathbf{0}$ and $var(\boldsymbol{e}) = \sigma^2\mathbf{I}$, where $\mathbf{I}$ is the identity matrix. We aim to select the relevant predictor variables to entering the regression model.

In the variable selection problem, there is uncertainty as to which subset of predictor variables to use. This problem receives more attention especially when the number of variables p is large and there are irrelevant and/or redundant variables in the set of predictors. For fitting the regression model (1.1), which is as follows

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}, \tag{1.2}$$

where

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \tag{1.3}$$

a variable selection technique should identify which subset of p predictors truly have nonzero coefficients and which have zero coefficients. In order to build an accurate linear predictor of type (1.2), we should force the coefficients of this set of variables to nonzero and the rest to zero. In data analysis, this is a fundamental problem and is not limited only to linear regression or to the field of statistics.

1.1.1 Need for Variable Selection

There are several advantages in using an efficient variable selection algorithm. One reason is simplicity and interpretability. In order to understand the functional relation between the response variable (output) and the predictor variables, finding the most relevant and potential predictor variables leads to the dimension reduction of the number of variables and to a simpler interpretation and understanding of the model.

The next reason is saving time. The best subset selection methods involve considering all possible subsets of candidate predictors for calculating a suitable model. These methods are also difficult to apply to data sets with large number of predictors. For instance, with p number of predictor variables 2^p different models should be tried for an extensive search of all possible subsets, and therefore considering all of these models is very time-consuming. (Kadane and Lazar (2004)). Thus, an efficient algorithm can save a lot of time and also the costs of finding the true model out of a large number of models.

The other reason is prediction performance, which can be improved by selecting the true predictors (i.e., those predictors whose regression coefficients are nonzero). In fitting the models using least squares regression, adding each new predictors to the model adds to the variance of the predicted values. Hence, the fewer the number of estimated coefficients the lower the variance. By using efficient variable selection methods, we can be confident to select the correct number of variables small enough to have a small variance, but large enough to yield a good prediction performance (see details in Hastie *et al.* (2009), Miller (2012)).

1.2 Outlier Detection and Robust Variable Selection

If the errors in the classical linear regression model (1.1) follow a mixture distribution, for instance a mixture of Normal distribution and some general distribution F, that is $\boldsymbol{e} \sim (1-a)N(0,\sigma^2) + aF$, where $0 < a \leq 1$, then they do not follow the assumptions of the classical linear regression model. Variable selection problem has not been studied well in these situations, and most of the variable selection methods use least squares in some way. Many estimation procedures like least squares can be highly effected by a small proportion of outliers in data. Hence, robust variable selection methods are needed to remedy this problem and do show a reliable behavior in the presence of outliers in data. In contaminated data sets, outlier detection and variable selection are not two separable things since selection may have effects on what is considered as an outlier and vice versa.

1.3 Thesis Overview

1.3.1 Objectives and Achievements

In this thesis we develop principal components regression (PCR) and partial least squares regression (PLSR) as two popular regression methods to enable them for variable selection (Massy (1965), Helland (1988), Helland (1990), Zou *et al.* (2006), Chung *et al.* (2013)).

We evaluate and compare the performance of the proposed methods with their counterparts on both different simulation studies and real data sets. The proposed variable selection algorithm for both PCR and PLSR is competitive with its counterparts presented in the literature review or even outperforms them according to the considered simulation studies. The application of the methods on real data also confirms the good performance of the proposed variable selection algorithms (Shahriari *et al.* (2014)).

We also aim to develop and enable the powerful modified version of the forward stage-wise procedure to perform outlier detection and robust variable selection simultaneously (Efron *et al.* (2004), Khan *et al.* (2007)). We investigate and compare the performance of the proposed outlier detection and robust variable selection method on both simulated and real data sets. This method presents a robust version of the classical least angle regression (LARS) (Efron *et al.* (2004)) and is competitive with its counterparts presented in literature review or even outperforms them (Shahriari *et al.* (2014)).

1.3.2 Variable Selection Ability

We contribute with variable selection algorithms for principal components regression (PCR) (Massy (1965)) and partial least squares regression (PLSR) (Helland (1988), Helland (1990)) as two popular applicable high-dimensional regression methods (i.e., regression methods that can be used even when $p \gg n$). The basic idea of these algorithms is based on the significance tests of the regression coefficients in the PCR/PLSR model. This technique automatically generates a standard error estimates of the coefficients parameters and then, by dividing the estimated regression coefficients by their estimated standard errors, the t-test values will be obtained, giving the significance level for each parameter. The estimated regression coefficients are calculated using bootstrap (Efron (1982)). Based on such bootstrap estimates of the regression coefficients, useless variables may be eliminated automatically in order to simplify the model, which makes it more reliable. This approach enables PCR and PLSR to perform variable selection.

By conducting different simulation studies, the proposed method is evaluated and compared with its counterparts presented in the literature review. The results of different simulation scenarios showed that the performance of the proposed method is competitive or outperforms in the considered simulation studies in comparison with its counterparts. We also applied the methods to real data and the results yielded the same achievements in simulation study.

1.3.3 Outlier Detection and Robust Variable Selection Development

We present an algorithm for the least angle regression (LARS) (Efron *et al.* (2004)) which is a modified version of the forward stage-wise procedure. The LARS algorithm is a powerful and computationally efficient method to sequence the predictor variables for least squares regression. Since LARS is based on the pairwise correlation between the predictor variables and the response variable, it is not robust to the presence of a small amount of anomalies in data. We propose a method capable of outlier detection and robust variable selection simultaneously which presents a robust sequence of predictor variables for LARS.

This method is obtained by combining robust least angle regression with least trimmed squares regression (Rousseeuw (1984), Rousseeuw and Van Driessen (2006)) on jackknife subsets (Efron (1982)) to detect outliers. The detected outliers are then removed to obtain the clean data, then standard least angle regression is applied on this clean data to robustly sequence the predictor variables in order of their importance.

1.3.4 Thesis Structure

The first section of this chapter consists of the description of the variable selection problem and its advantages for the regression models with a large number of predictors. Then we point to the problem of variable selection when the data contains outliers, thus revealing the necessity for a robust variable selection method.

The following chapters of this thesis are organized as follows. In Chapter 2, we will describe the literature related to variable selection and some popular regression methods for data sets with large number of predictors.

In Chapter 3, first we will provide the methodology of principal components regression (PCR) and partial least squares regression (PLSR). We will also address why these two popular methods need to be developed to perform variable selection. Second, we will describe the existing sparse based methods for principal components regression and partial least squares regression, and then we will explain resampling methods which will be used inside our variable selection approach to produce automatic estimates for the regression parameters. Finally, we will propose variable selection algorithms to enable principal components regression and partial least squares regression in order to find the most relevant predictor variables. Different simulation scenarios will be conducted along by using real data to evaluate and to compare the performance of the mentioned methods, and the results will be discussed.

In Chapter 4, we will address the sensitivity of linear regression followed by least squares to the presence of outliers in data and will describe some popular robust variable selection methods which are resistant in these situations. We will point out the least angle regression and its sensitivity to the presence of outliers in data and will review the existing robust least angle regression method, as our method's counterpart. We will also describe another robust variable selection method based on the variance inflation factor (VIF), which also inherits the spirit of a variation of forward stage-wise regression. Then, we will propose our approach for identifying outliers and robustly selecting variables for least angle regression, simultaneously. Finally, by conducting different simulation studies we will investigate and compare the performance of the proposed method with its counterparts. We will also apply the mentioned methods to real data. The results of both simulation study and real data will be discussed.

Finally, in Chapter 5, the conclusions of our study will be drawn by summarizing the main ideas and the results of this thesis, and future developments will be addressed.

Chapter 2

Review: Variable Selection in Linear Regression

2.1 Introduction

The method of least squares (LS) is a standard approach to estimate the parameters in regression analysis. LS minimizes the sum of squared errors, in a regression model. There are two main reasons why we are often not convinced with least squares estimates:

- *Predictive ability*: least squares estimates has low bias (zero bias) but suffers from high variance. Therefore, it may be worth sacrificing some bias to reduce the variance of the predicted values, and, as a consequent, to increase the model's predictive ability. This can be achieved by shrinking or setting some coefficients to zero (Hastie *et al.* (2009), Miller (2012));

- *Interpretation*: it can be helpful to determine a smaller subset of variables through a large number of predictor variables in order to simplify the model for interpretation.

As we represented in Section 1.1 in Chapter 1, our aim is to select the most relevant predictor variables in linear regression models. The selection criteria are numerous and can be based on prediction, fit, and etc. We define some notation in order to explain these selection criteria.

Suppose we are fitting a model, which we call submodel, with a subset of the possible predictor variables. We define $q_{submodel}$ as the number of variables in the submodel plus one for the intercept, which is the number of nonzero coefficients we would fit in this model. When we use least squares to predict the response variable, given the submodel, we name this prediction, $\hat{\boldsymbol{y}}_{submodel}$.

We define the total sum of squares, $SST_{submodel} = \sum_{i=1}^{n}(y_i - \bar{y})^2$, with $\bar{y}$ the mean of the response variable, the regression sum of squares, $SSR_{submodel} = \sum_{i=1}^{n}(\hat{y}_{submodel,i} - \bar{y})^2$, the residual sum of squares, $RSS_{submodel} = \sum_{i=1}^{n}(y_i - \hat{y}_{submodel,i})^2$, and the hat matrix $\boldsymbol{H} = \mathbf{X}(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}^T$ for the full model which contains p predictors plus one for the intercept.

One simple idea in full model is to use adjusted R^2 which adjusts the multiple correlation coefficient for the number of parameters fitted in the model with n number of observations

$$R^2_{adj} = 1 - \frac{n-1}{n - q_{submodel} - 1}(1 - R^2), \tag{2.1}$$

where

$$\begin{aligned} R^2 &= \frac{SSR_{submodel}}{SST_{submodel}} \\ &= 1 - \frac{RSS_{submodel}}{SST_{submodel}}. \end{aligned}$$

A related criterion is Mallow's C_p statistic (Mallows (1973)). For a given model where we fit $q_{submodel}$ parameters,

$$C_p = \frac{RSS_{submodel}}{\hat{\sigma}^2} + (2q_{submodel} - n), \tag{2.2}$$

where $\hat{\sigma}^2 = \dfrac{RSS_{full}}{n-p-1}$, where p is the number of predictors, and $RSS_{full} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. The motivation behind C_p was to create a statistic to estimate the mean square prediction error (Mallows (1995)).

Some other selection criteria, such as akaike information criterion (AIC), and the corrected AIC (AIC_c) (Akaike (1970), Akaike (1974), Sugiura (1978)) are as follows

$$AIC = nlog(\frac{RSS_{submodel}}{n}) + 2q_{submodel};$$

$$AIC_c = nlog(\frac{RSS_{submodel}}{n}) + 2(q_{submodel} + 1) \times \frac{n}{(n - q_{submodel} - 2)}.$$

C_p, AIC, and AIC_c tend to select too large of a subset and the data is overfit (Shao (1993)).

The existing selection procedures can be broadly categorized into three classes according to their general strategy and respective computational efficiency.

A first class considers the potential subsets of predictors. *All possible subset selection* considers all the combination subsets of the predictors and evaluates each of these submodels according to a fixed criterion. The model which best satisfies the selection criteria is chosen as the optimal model. *Stepwise subset selection* is built on the basis of sequential selection procedures in which a predictor at a time is added to (or removed from) the model, based on a criterion that can change from one step to the next and that is calculated for all potential predictor variables to add (or to exit) until another criterion is satisfied. The second class shrinks the regression coefficients by imposing a penalized term on their size in order to satisfy bias to reduce the variance of the predicted values, and, as a consequent, to increase the model's predictive ability. The third class of selection procedures is also formed of sequential procedures in which each predictor is evaluated to enter into a model as a potential predictor. Each predictor is evaluated once to enter as a potential predictor.

The selection criteria such as AIC, AIC_c (Akaike (1970), Akaike (1974)), Mallow's C_p (Mallows (1973)), cross-validation (see also Efron (2004)) may be categorized into the first class. When the number p of the predictor variables is small, one may compute these selection criteria for all possible subsets of predictors. But as the number p of the predictor variables increases, the number of subsets grows dramatically; as a result, the computational burden of using this approach is increased and makes it infeasible for computation. This problem leads to the popularity of step-by-step algorithms like forward selection and stepwise (SW) selection procedures (Weisberg (2014)).

The second class may contain shrinkage methods in which the estimator of the regression coefficients $\boldsymbol{\beta}$ minimizes the sum of squared errors according to a norm q as a penalized term, i.e.,

$$\arg\min_{\boldsymbol{\beta}} \left\{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_q \|\boldsymbol{\beta}\|_q^q \right\}, \tag{2.3}$$

where $\mathbf{X} = [\mathbf{1} \;\; \boldsymbol{x}_j]_{j=1,\dots,p}$ with $\mathbf{1}$ a column of ones, $\|\boldsymbol{\beta}\|_q = \left(\sum_{j=1}^{p} |\beta_j|^q \right)^{1/q}$ for $q > 0$, and $\|\boldsymbol{\beta}\|_0 = \sum_{j=1}^{p} \mathbf{I}_{\{\beta_j \neq 0\}}$ with $\mathbf{I}_{\{\beta_j \neq 0\}} = 1$ if $\beta_j \neq 0$ and 0 otherwise (Lin *et al.* (2011)). Shrinkage methods that use $q = 1$ norm penalization are least absolute shrinkage and selection operator (LASSO) (Tibshirani (1996)), and the Dantzig selector (Candes and Tao (2007)). Ridge regression is another shrinkage method which uses $q = 2$ norm penalization.

The third class may include a variation of the stepwise regression that evaluates predictors sequentially. Least angle regression (LARS) (Efron *et al.* (2004)) can be categorized into this class. Streamwise regression (Zhou *et al.* (2006)), which evaluates predictors sequentially, though each of them is considered once to enter to the model, can also be categorized into this class. This approach uses the α-investing rule (Foster and Stine (2008)) and is a fast procedure.

Variance inflation factor (VIF) is an improved streamwise regression approach where the selected model is given by computing a fast test statistic based on the variance

inflation factor of the candidate variable (Lin *et al.* (2011)).
In Section 2.2 of this chapter we will describe all possible subset selection procedures, along with some classical step-by-step algorithms, such as forward stepwise and backward elimination selection. In Section 2.3 some popular shrinkage methods such as ridge regression and least absolute shrinkage and selection operator (LASSO) will be reviewed. Finally, in Section 2.4 the modified versions of stagewise regression, least angle regression (LARS) and variance inflation factor (VIF) regression will be described.

2.2 Potential Subset Selection

In subset selection only a subset of variables is kept and the rest of them is removed from the model. The coefficients of the potential predictors as the inputs of the model are estimated using least squares regression. There are a number of different strategies for subset selection.

2.2.1 All Possible Subset Selection

This approach considers all 2^p possible subsets of the predictor variables and determines the subset of the predictors of a given size that maximizes a measure of fit or minimizes an information criterion based on a monotone function of the residual sum of squares. The efficient algorithm "simple leap and bound" (Furnival and Wilson (1974)) can find the best subsets without examining all possible subsets. With a fixed number of terms in the regression model, there are a number of criteria such as the AIC (Akaike (1970)), Mallow's C_p (Mallows (1973)) that one may use to evaluate a subset of predictor variables; typically we choose the smallest model that minimizes the residual sum of squares. So, for instance, if a subset with a fixed number of terms satisfies a criteria among all subsets of size k, then this subset will also satisfies other selection criteria among all subsets of fixed size k. This process is not feasible for a large number p of predictor variables, and this problem leads to the popularity of step-by-step algorithms.

2.2.2 Forward Stepwise and Backward Elimination Selection

As we mentioned before, finding the best model among all the possible submodels will become infeasible when p is large. Therefore, in order to decrease the computation time we can search through all possible subsets to find a good path through them. In this section, we review some of the most important of this kind of algorithms. Forward stepwise selection starts with all coefficients equal to zero, and then sequentially adds to the model the predictor that most improves the fit (in the other words, the predictor that minimizes the criterion of interest or if an information criterion is used, then this amounts to find the most correlated predictor with the response variable). The procedure is continued until all predictors have been added to the model.

Backward elimination selection starts with the full model and at each step drops a predictor with the least impact on the fit; therefore it is the opposite of forward stepwise selection procedure. The procedure is continued until all predictors have been deleted from the model (Weisberg (2014)).

2.2.3 Forward Stagewise Regression

Forward stagewise (FS) regression is similar to forward stepwise regression but it is more constrained (Hastie *et al.* (2009)). It starts like forward stepwise regression with an intercept equal to $\bar{y}$, with all coefficients equal to zero. At each step, the most correlated variable with the current residual enters into the model. Then, the simple linear regression coefficient of the residual on this selected variable is estimated and then added to the current coefficient for that variable. This process is continued until none of the variables have correlation with the residuals. Since none of the variables is adjusted when a predictor is entered into the model, therefore it can take many steps in the direction of the same predictor in order to reach the least squares fit.

2.3 Shrinkage Methods

In subset selection procedure, a subset of the predictor variables is kept and the rest are removed; as a result, a model that is interpretable with lower prediction error than the full model is obtained. But since this procedure proceeds in a discreet way, i.e. variables are either kept or removed, it often shows high variance, and so does not reduce the prediction error of the full model. Shrinkage methods are more continuous, and do not suffer as much from high variability (Hastie *et al.* (2009)).

2.3.1 Ridge Regression

Ridge regression (Hoerl and Kennard (1970)) shrinks regression coefficients by penalizing their size. The ridge estimates minimize a penalized residual sum of squares,

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{ridge} &= \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \\ &= \arg\min_{\boldsymbol{\beta}} \left\{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2 \right\} \qquad (2.4)\end{aligned}$$

Here $\lambda \geq 0$ is a penalty parameter which controls the amount of shrinkage; i.e. the larger the value of λ, the greater the amount of shrinkage. When $\lambda = 0$, we get the least squares estimates; when $\lambda \simeq \infty$, we get $\hat{\boldsymbol{\beta}}^{ridge} = 0$; when λ is between these two values, we are balancing two ideas: fitting the linear regression model, and shrinking the coefficients.

An equivalent form of writing the ridge problem is

$$\hat{\boldsymbol{\beta}}^{ridge} = \arg\min_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^2 , \tag{2.5}$$

$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 \leq t,$$

where t is a tuning parameter varied over a certain range.

This rewriting form makes explicit how ridge regression puts constraint on the size of the parameters. Parameter λ in (2.4) has a one-to-one correspondence with t in (2.5).

In the presence of many correlated variables in a linear regression model, their coefficients can become poorly determined and show high variance. One variable with a large positive coefficient can be canceled by a similarly negative coefficient as its correlated cousin. Imposing a constraint on the size of the coefficients, as in (2.5), this problem is thus conciliated. It should be noticed that the intercept term has left out the penalty term, the solution for $\hat{\beta}_0 = \bar{y}$, and so we assume that the inputs have been standardized before solving (2.4).

2.3.2 Least Absolute Shrinkage and Selection Operator

Least absolute shrinkage and selection operator (LASSO) (Tibshirani (1996)) shrinks the regression coefficients by imposing norm 1 as the penalty parameter. LASSO is like ridge regression but the penalty term is $\|\boldsymbol{\beta}\|_1$ instead of $\|\boldsymbol{\beta}\|_2$ as in ridge regression. The LASSO estimate is defined as

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{lasso} &= \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \\ &= \arg\min_{\boldsymbol{\beta}} \left\{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \textstyle\sum_{j=1}^{p} |\beta_j| \right\} \\ &= \arg\min_{\boldsymbol{\beta}} \left\{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right\} . \end{aligned} \tag{2.6}$$

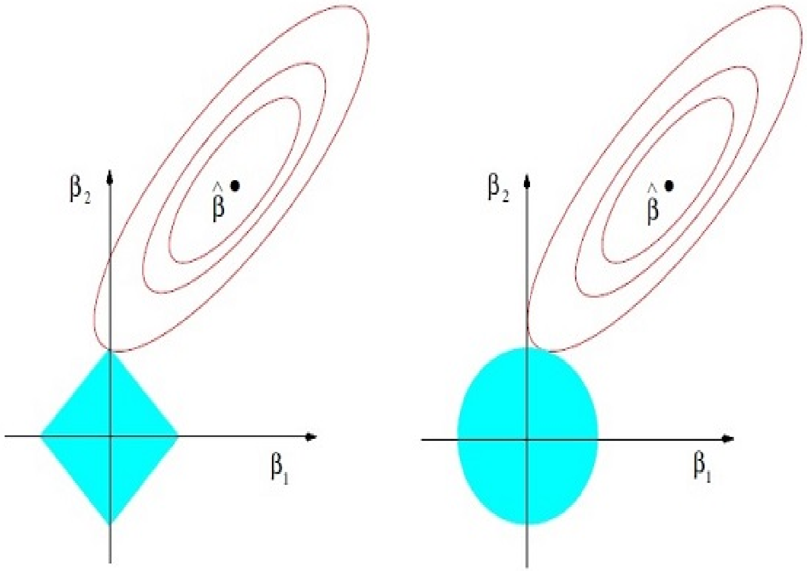

Figure 2.1: Estimation picture for LASSO (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions $|\beta_1| + |\beta_2| \leq t$ and $\beta_1^2 + \beta_2^2 \leq t^2$, respectively, while the red ellipses are the contours of the least squares error function [Tibshirani (1996)].

Just as in ridge regression, we can rewrite (2.6) as

$$\hat{\boldsymbol{\beta}}^{lasso} = \arg\min_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^2 , \tag{2.7}$$

$$\text{subject to} \sum_{j=1}^{p} |\beta_j| \leq t,$$

where t is a tuning parameter over a certain range.

Because of using $q = 1$ norm penalization as you can see in Figure 2.1, LASSO can shrink the regression coefficients exactly to zero, and therefore can perform variable selection. The LASSO estimator can be computed using the efficient least angle regression algorithm (LARS), which is a modified version of forward stagewise procedure. Like ridge regression, the inputs are standardized and the solution for $\hat{\beta}_0 = \bar{y}$, and thereafter we fit a model without an intercept.

LASSO and Ridge Regression As we mentioned in Section 2.3.1, when there are many correlated variables in a linear regression model, the estimates may exhibit high variance. Ridge regression, by imposing a restriction on the size of the regression coefficients, tries to alleviate this problem.

Ridge regression shrinks the regression coefficient toward zero but is not able to set some coefficients exactly equal to zero; thus, unlike LASSO, cannot perform variable selection.

2.4 Modified Versions of Stepwise Regression

In this section we describe two modified and improved versions of stepwise regression which are fast to compute and sequence the predictor variables in their importance.

2.4.1 Least Angle Regression

Least angle regression (LARS) (Efron *et al.* (2004)) enters variables in the model in order of their importance. LARS is related to forward stagewise (FS) procedure, which sequentially adds variables to the model in small steps. FS uses in each step the most correlated variable with the residuals, and updates the fit by adding only a small fraction of the least squares contribution of this variable to the model. Usually, FS takes a large number of steps using the same variable before another variable yields a higher correlation with the remaining residual. Therefore, FS is not computationally efficient. By deriving a simple mathematical formula for the optimal step size of a selected variable, LARS largely reduces the number of steps and speeds up the computation time.

The steps of the LARS algorithm can be summarized as follows:

1. Start with all coefficients equal to zero and set the initial residual vector equal

to response vector $\mathbf{y}$;

2. Find the predictor that is most correlated with the response variable, say $\boldsymbol{x}_1$;

3. Take the largest step possible in the direction of this predictor until some other predictor, say $\boldsymbol{x}_2$, has the same maximal correlation (in an absolute value) with the current residual;

4. Proceed in the equiangular direction of the two predictors until a third variable $\boldsymbol{x}_3$ earns its way into the "most correlated" set. The equiangular direction is the direction in which the correlation of the predictors decreases at the same pace so that these correlations remain equal at all times;

5. Repeat until all predictors have been entered.

LARS is only based on the means, variances and correlations of the data, and so it is sensitive to the presence of contamination in the data. In Chapter 4 we will discuss about its sensitivity and to remedy this problem we will present a robust version of this algorithm and evaluate the performance of our method against its counterparts.

LARS and LASSO LARS and LASSO are closely connected. LARS can provide an efficient algorithm to compute the LASSO path. By modifying the LARS algorithm, the LASSO path can be obtained as follows: if a non-zero coefficient crosses zero, remove its variable from the active set of variables and recompute the best direction, that is, current joint least squares direction.

2.4.2 Variance Inflation Factor (VIF)

Variance inflation factor (VIF) regression (Lin *et al.* (2011)) also inherits the spirit of a variation of stepwise regression. VIF regression searches over the predictor variables to test each potential predictor variable for addition to the model. VIF regression

selects those predictors among other available predictors that can reduce a statistically sufficient part of the variance in the predictive model. VIF regression approximates the partial correlation of each candidate variable with response variable by correcting (using VIF) the marginal correlations.
The characteristics of the VIF algorithm are the following two components:

- *Evaluation step*: the partial correlation of each candidate variable $\boldsymbol{x}_j, j = 1, \ldots, p$, with the response variable $\mathbf{y}$ is approximated by correcting (using the variance inflation factor) the marginal correlation using a small presampled set of data;

- *Search step*: each variable is tested sequentially using an α-investing rule (Foster and Stine (2008)). The α-investing rule purveys high accurate models and guarantees no model overfitting.

Let $\mathbf{X}_S$ be the design matrix that contains the selected variables at stage S, and $\tilde{\mathbf{X}}_S = [\mathbf{X}_S \;\; \boldsymbol{z}_j]$ with $\boldsymbol{z}_j$ the new candidate to be evaluated for inclusion. Consider the following models:

$$\mathbf{y} = \mathbf{X}_S\boldsymbol{\beta}_S + \boldsymbol{z}_j\beta_j + \boldsymbol{e}_{step}, \quad \boldsymbol{e}_{step} \sim N(\mathbf{0}, \sigma^2_{step}\mathbf{I}), \tag{2.8}$$

$$\mathbf{r}_S = \boldsymbol{z}_j\gamma_j + \boldsymbol{e}_{stage}, \quad \boldsymbol{e}_{stage} \sim N(\mathbf{0}, \sigma^2_{stage}\mathbf{I}), \tag{2.9}$$

where $\mathbf{r}_S = \left(\mathbf{I} - \mathbf{X}_S(\mathbf{X}_S^T\mathbf{X}_S)^{-1}\mathbf{X}_S^T\right)\mathbf{y}$ are the residuals of the fit of $\mathbf{y}$ on $\mathbf{X}_S$, $\mathbf{I}$ is the identity matrix, $\boldsymbol{\beta}_S$ are slope parameters, and γ_j is the slope parameter of the fit of $\boldsymbol{z}_j$ on the residuals $\mathbf{r}_S$.

When there are some collinear predictors in the data, all known estimators of parameters β_j, σ^2_{step}, and γ_j, σ^2_{stage} will provide different estimates and, as a consequent, significance tests based on estimates of β_j or γ_j do not necessarily lead to the same conclusions. While in stepwise regression the significance of β_j in model (2.8) is at the core of the selection procedure, in VIF regression it is more convenient to estimate γ_j.

When least squares (LS) are used to estimate,

$$\hat{\gamma}_j = \rho^2 \hat{\beta}_j, \tag{2.10}$$

where $\rho^2 = \boldsymbol{z}_j^T \left(\mathbf{I} - \mathbf{X}_S \left(\mathbf{X}_S^T \mathbf{X}_S \right)^{-1} \mathbf{X}_S^T \right) \mathbf{z}_j$ Lin *et al.* (2011). Then, a t-statistic $T_{\gamma_j} = \frac{\hat{\gamma}_j}{\hat{\rho}\hat{\sigma}}$ with proper estimates for ρ and σ is computed and compared to the standard normal distribution to decide whether or not $\mathbf{z}_j$ should be entered to the current model. This procedure is named VIF regression because $1/\rho$ is called the VIF for $\mathbf{z}_j$ (see Marquaridt (1970)).

Test statistic T_{γ_j} is very sensitive to outliers since it is based on the following:

- the LS estimator $\hat{\gamma}_j$;
- ρ, in turn based on the design matrix $\mathbf{X}_S$ and $\boldsymbol{z}_j$;
- the classical estimator of σ, a form of the model deviation.

In Chapter 4 we will describe the robust version of this method and discuss its performance in different simulation scenarios as well as on real data.

2.5 Discussion and Conclusions

All possible subset selection procedure which seeks through all possible 2^p subsets of the predictor variables in order to find the best subset of a given size, which maximizes a measure of fit or minimizes an information criterion based on a monotone function of the residual sum of squares, is not computationally efficient, and will be infeasible when the number p of the predictor variables is large.

Forward stepwise selection is a greedy algorithm which produces a nested sequence of models, therefore in this sense it might seem suboptimal compared to all possible subset selection procedure. But there are two main reasons why forward stepwise selection procedure might be preferred rather than all possible subset selection procedure:

- *Computationally*: when the number p of the predictor variables is large it is not feasible to compute the best subset sequence; however, forward stepwise selection can always compute the best subset (even when $p \gg n$, n is the number of observations);
- *Statistically*: a cost is paid in variance for determining the best subset of each size; forward stepwise is a more constrained search, with lower variance but perhaps more bias.

The drawback of backward stepwise selection in comparison with forward stepwise is that it can only be used when $n > p$, while the later can be used at all times. In forward stagewise, unlike forward stepwise regression, none of the variables are adjusted when a term is entered to the model, and therefore forward stagewise takes many more than p steps to achieve the least squares fit.

By keeping a subset of the predictor variables and dropping the rest, subset selection produces an interpretable model which has lower prediction error compared to the full model. However, since it proceeds in a discrete path, that is, variables are either kept or dropped, therefore it does not reduce the prediction error of the full model. Shrinkage methods are more continuous, and are not affected as much by high variability.

Ridge regression does a proportional shrinkage in order to shrink the regression coefficients toward zero but does not set them exactly equal to zero. LASSO performs variable selection by setting the predictor variables exactly equal to zero.

LARS and VIF regression inherit the spirit of a variation of stepwise regression. These methods are sensitive when data contains outliers, and in Chapter 4 we will discuss their sensitivity to outliers.

Chapter 3

Sparsity for Principal Component and Partial Least Squares Regression

3.1 Introduction

High-dimensional regression (regression analysis when the number of predictors is larger than the number of observations) has captured the attention of many statisticians worldwide, because of its interesting applications, as well as the unique challenges faced. Motivated by interesting applications, there has been a dramatic growth in the development of statistical methodology in the analysis of high-dimensional data ($p \gg n$), particularly related to regression (model selection, estimation and prediction).
The problem of high-dimensional variable selection has received tremendous attention during the last decades because variable selection plays an important role in high-dimensional statistical modeling (Fan and Li (2001) and Lin *et al.* (2011) among many others). The usual variable selection procedure using least squares and a penalty

which involves the number of parameters in the candidate submodel (e.g. AIC) is infeasible to compute exhaustively (Bühlmann and Van de Geer (2011)). Forward selection strategies are computationally fast but they can be very instable, yielding poor results (Breiman (1996)). Best subset selection methods involve considering all possible subsets of candidate predictors for calculating a suitable model. These methods are also difficult to apply to high dimensional data. For instance, if the number of predictors is p, then 2^p different models should be tried for an extensive search of all possible subsets. This value, even for moderate values of p, grows exponentially and makes the search impractical (Trevor *et al.* (2001)). Although several search-based strategies such as artificial neural network (Burden *et al.* (1997)) and genetic algorithms (Jouan-Rimbaud *et al.* (1995)) have been successfully applied to variable selection problems, these methods still have some drawbacks. Several variable selection (VS) methods in high dimensional data in the unified framework of penalized likelihood estimator are ridge regression (Frank and Friedman (1993)), LASSO (Tibshirani (1996)), elastic net (Zou and Hastie (2005)), the new smoothed LASSO (Meier and Bühlmann (2007)) and boosting algorithms (Fan and Lv (2010) and Bühlmann and Van de Geer (2011)).

Principal components regression (PCR) and partial least squares regression (PLSR) are well-known techniques in dealing with high-dimensional data whose number of predictors p is larger than sample size n (Engelen *et al.* (2004), Helland (1988), Helland (1990) and Hubert and Verboven (2003)). PCR (Massy (1965)) and PLSR (Wold (1966)) are two related families of methods that contain selecting a subspace of the column space of the predictors on which to project the response vector. These methods latter appeared in chemometrics, sensory evaluation and food research (Wold (1975b), Wold *et al.* (1984), Geladi and Kowalski (1986), Martens and Martens (2000), Martens and Næs (1989), Stone and Brooks (1990), and De Jong (1993b)). PCR and PLSR as two dimension reduction techniques that have recently obtained much attention in high dimensional genomic data (Boulesteix and Strimmer (2007)).

Although PCR and PLSR are two popular techniques that can handle multicollinearity among predictors even when there are more predictors than observations, it is still necessary to eliminate irrelevant predictors in high-dimensional data. The drawback of PCR and PLSR is that, when using these two methods, the number of variables used is not reduced, since the components are linear combinations of all the predictors.

This drawback makes it often difficult to interpret the derived principal components (PC). In order to help practitioners to interpret principal components, rotation techniques are commonly used (Jolliffe (1995)). There is a technique which restricts the loadings in principal components to take values from a small set of given integers such as 0, 1, -1 (Vines (2000)).

Existing irrelevant variables can influence the PCR and PLSR models. Imposing sparsity in the midst of dimension reduction step may lead to variable selection. Huang et al. (2004) recently proposed a penalized PLS method that thresholds the final PLS estimator (Huang *et al.* (2004)). Chun and Keleş (2010) formulated sparse partial least squares (SPLS) regression by relating it to sparse principal component analysis (SPCA) (Jolliffe *et al.* (2003), Zou *et al.* (2006)). SPCA involves formulating principal component (PC) analysis as a regression-type optimization problem, and then obtaining sparse loadings by imposing LASSO (Tibshirani (1996)) (and a generalization of LASSO, elastic net) constraint on the regression coefficients. Therefore, it is important to reduce the size of the explicitly used variables (i.e., variable selection) besides having dimension reduction for PCR and PLSR. We introduce an approach which is based on bootstrap technique for significance tests of the regression coefficients to perform variable selection inside PCR and PLSR.

The rest of this chapter is organized as follows. First, in Section 3.2 we will give an introduction of the methodology for PCR and PLSR, and will compare PCR and PLSR. Then, in Section 3.3 we will review jack-knife and bootstrap as two resampling methods. We will explain some variable selection methods for PCR and PLSR, along with our proposed approach in Section 3.4. We will conduct a simulation study in

Section 3.5 to compare the performance of the presented variable selection methods for PCR and PLSR. In Section 3.6, we will show the results of the simulation study. We will also apply these methods to a real data in 3.7; and finally in Section 3.8 we will conclude.

3.2 Principal Components Regression and Partial Least Squares Regression

As mentioned before in Section 1.1 in Chapter 1, a linear regression model assumes that there is a linear relationship between the response variable and the set of the predictor variables. There are various methods to estimate the slope parameters $\boldsymbol{\beta}$ in a linear regression model. In the presence of large number of predictors in the model, LS can be used; however, if the number of predictors gets too large (more than the number of observations, $p \gg n$, the matrix $\boldsymbol{X}^T\boldsymbol{X}$ gets singular), the corresponding multiple linear regression (MLR) model obtained by LS is not able to fit the sampled data well and simultaneously predict the new data well; that is the problem of overfitting. To remedy this problem, PCR and PLSR can be used, since they reduce the dimension of the design matrix $\boldsymbol{X}$. Also, they can handle the multicollinearity in the data.

PCR and PLSR are two multivariate regression methods concerned with explaining the variance-covariance structure of the data through a small number of components which are linear combinations of the original variables (Engelen *et al.* (2004)). The basis of both techniques is a bilinear model that explains the existence of a relation between a set of p-dimensional predictors and a response variable through the latent component matrix $\boldsymbol{T}_{n\times k} = (\boldsymbol{t}_1, \ldots, \boldsymbol{t}_k)$ where the $\boldsymbol{t}_i, i = 1, \ldots, k$ are the score vectors, with $k \ll p$. By considering the observations $(\boldsymbol{X}, \mathbf{y})$, we assume that,

$$\boldsymbol{X} = \mathbf{1}\bar{\boldsymbol{x}} + \boldsymbol{T}\boldsymbol{P}^T + \boldsymbol{\nu}, \tag{3.1}$$

$$\mathbf{y} = \mathbf{1}\bar{y} + \boldsymbol{T}\boldsymbol{q} + \boldsymbol{\xi}. \tag{3.2}$$

Here $\bar{\boldsymbol{x}}$ is the p-dimensional mean vector of predictors, $\bar{y}$ the mean of the response variable, $\boldsymbol{P}$ the $p \times k$ matrix of loadings and $\boldsymbol{q}$ represents the k-component slope parameters in the regression of $\mathbf{y}$ on the $\boldsymbol{t}_i$. The error terms are denoted by the $n \times p$ matrix $\boldsymbol{\nu}$ and the vector $\boldsymbol{\xi}$ with n elements. These corresponding errors show the unique variation in $\boldsymbol{X}$ and $\mathbf{y}$ that is not explained by the k-component bilinear structure. The errors may be due to operator mistakes, measurement noise, mis-specified model, etc. In order to make inference on the ordinary least squares estimator of $\boldsymbol{q}$, an assumption on the errors of $\boldsymbol{\xi} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ is needed. This condition is equivalent to assuming $\mathbf{y}|\boldsymbol{q} \sim N(\mathbf{1}\bar{y} + \boldsymbol{T}\mathbf{q}, \sigma^2\mathbf{I})$ which represents that the variance-covariance matrix of $\mathbf{y}$ depends only on one parameter σ^2; such a matrix is also known as a scalar covariance matrix.

Figure 3.1 shows the major two stages of PCR and PLSR. In the first step, due to (3.1), the latent components $\boldsymbol{t}_i$ are built as linear combinations of the original predictor variables or, in the other words, the high-dimensional observations $\boldsymbol{X}$ are summarized in latent components $\boldsymbol{t}_i$s of dimension $k \ll p$. The model's predictive ability depends on the number k of components selected. If too few components are chosen, the valid structure in the data is left and the model *under-fits* the data. Choosing too many components will lead to *over-fitting* in the data, pulling too much noise which will diminish model prediction. Various different criteria (such as Adjusted R^2, root mean squared error ($RMSE$)) can be used in order to select the number of components k (Aucott *et al.* (2000)). Cross-validation (CV) or independent test set validation are usually applied to select the optimal number of components. If the amount of existing

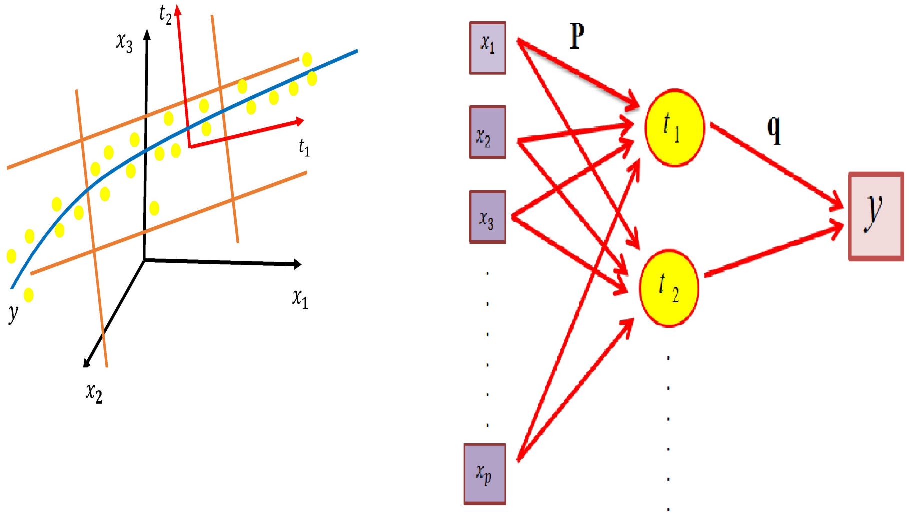

Figure 3.1: PCR and PLSR schematic representation. The left hand plot shows the first stage of the bilinear model. The red arrows represent the direction of the first 2 components. The general process of PCR and PLSR are illustrated in the right plot.

data is limited, especially for high-dimensional data, cross-validation is preferable to independent test set validation (Martens and Næs (1989)).

For the second step of the algorithm, due to (3.2), these k latent components will become the regressors for the $\mathbf{y}$. The objective of the second step is to find a k-dimensional regression hyperplane that best fits the data $(\mathbf{y}, \boldsymbol{T}_{n\times k})$. At last, the estimates for the parameters of the original predictor variables, $\hat{\boldsymbol{\beta}}$, can be obtained from the bilinear model.

A large body of literature has compared PCR and PLSR (Martens and Næs (1989), Stone and Brooks (1990), Wentzell and Vega Montoto (2003), Schumann *et al.* (2013)). Both methods construct new predictors, known as components, as linear combinations of the original predictors, but they construct those components in different ways. The construction of the k-dimensional scores vector $\boldsymbol{t}_i$ by PCR and PLSR is not the same. PCR creates components to explain the observed variability in the predictors, without considering the response variable at all, and thus using a variance criterion. On the other hand, PLSR does take the response variable into account, and therefore often leads to models that are able to fit the response variable with fewer components, i.e., PLSR finds linear combinations of the predictors that best explain the response by maximizing a covariance criterion between the predictors and the response variable (Frank (1987), Næs *et al.* (2002)). With the same number of latent variables, PLSR will cover more of the variation in response variable and PCR will cover more variation of the original predictor variables.

In practice, however, there is hardly any difference between the use of PCR and PLSR; by construction, it is thus expected that PLSR requires less components than PCR (Frank and Friedman (1993), Mevik and Wehrens (2007)). In most situations, both methods achieve similar prediction performances, although by construction it is expected that PLSR gives improved prediction results with fewer components than PCR (De Jong (1993a), Garthwaite (1994)). Some components from a PCR model fit may serve primarily to describe the variation in the predictors, and may include large weights for variables that are not strongly correlated with the response. Thus, PCR can lead to retaining variables that are unnecessary for prediction.

3.2.1 Principal Components Regression Algorithm

Principal component analysis (PCA) is a popular statistical method which can explain the covariance structure of the data through a small number of components which are linear combinations of the original predictor variables (Jolliffe (2005)). These components can be used for an interpretation and a better understanding of the different sources of variation. PCR is based on principal component analysis. PCR preforms a least-squares regression of projections of data onto the basis vectors of a factor space (Kramer (1998)). In PCR, instead of directly regressing the response variable on the predictor variables, the principal components are used as regressors. One typically uses only a subset of all the principal components for regression. Often the principal components with higher variances (those that are based on eigenvectors corresponding to the higher eigenvalues of the sample variance-covariance matrix of the predictors) are chosen as regressors. However, for predicting the response variable, the principal components with low variances may also be important, and in some cases even more important (Jolliffe (1982)). In the classical approach, the first component corresponds to the direction in which the projected observations have the largest variance. The second component is then orthogonal to the first and again maximizes the variance of the data points projected onto it. Proceeding in this way produces all the principal components, which correspond to the eigenvectors of the empirical covariance matrix.

First, PCR centers the data using the column mean $\bar{\boldsymbol{x}}$ of the $\boldsymbol{x}$-variables and the mean $\bar{y}$ of the $\mathbf{y}$-variable. Denote the centered observations by

$$\begin{aligned}\tilde{\boldsymbol{X}} &= \boldsymbol{X} - \mathbf{1}\bar{\boldsymbol{x}},\\ \tilde{\mathbf{y}} &= \mathbf{y} - \mathbf{1}\bar{y}.\end{aligned}$$

Then, for dealing with the multicollinearity in the $\boldsymbol{x}$-variables, the first k ($k \leq min(n,p)$) principal components of $\boldsymbol{X}$ are calculated. These loading vectors $\hat{\boldsymbol{P}} = (\hat{\boldsymbol{p}}_1, \ldots, \hat{\boldsymbol{p}}_k)$ are the k eigenvectors that correspond to the k largest eigenvalues of the empirical covariance matrix $\boldsymbol{S}_{\boldsymbol{x}} = \dfrac{1}{n-1}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}$ (Jolliffe (1982), Jolliffe (2005)). Therefore, these loading vectors are uncorrelated, orthogonal and they construct a new coordinate system in the k-dimensional subspace that they span. Denote the eigenvalues of the $\boldsymbol{S}_{\boldsymbol{x}}$, by $\hat{\tau}_1, \ldots, \hat{\tau}_p$. The sums-of-squares of the component scores $\hat{\boldsymbol{t}}_1, \ldots, \hat{\boldsymbol{t}}_p$ are calculated as $\hat{\tau}_a = \hat{\boldsymbol{t}}_a^T\hat{\boldsymbol{t}}_a, a = 1, \ldots, p$. The orthogonality properties of the loadings and scores can be written in matrix languages as

$$\hat{\boldsymbol{P}}^T \hat{\boldsymbol{P}} = \mathbf{I},$$
$$\hat{\boldsymbol{T}}^T \hat{\boldsymbol{T}} = diag(\hat{\tau}_1, \ldots, \hat{\tau}_k),$$

where $\mathbf{I}$ is the identity matrix and $diag()$ denotes a diagonal matrix with zero elements off the leading diagonal. Then, the k-dimensional scores of each data point $\hat{\boldsymbol{t}}_i$ are calculated as the coordinates of the projections of $\tilde{\boldsymbol{X}}$ onto this subspace, or equivalently

$$\hat{\boldsymbol{T}} = \tilde{\boldsymbol{X}}\hat{\boldsymbol{P}}(\hat{\boldsymbol{P}}^T\hat{\boldsymbol{P}})^{-1} = \tilde{\boldsymbol{X}}\hat{\boldsymbol{P}}. \tag{3.3}$$

It can be shown that for centered $\boldsymbol{x}$-variables the $\hat{\boldsymbol{p}}_a, a = 1, \ldots, k$ dimensional vectors are eigenvectors of $\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}$ with the $\hat{\tau}$'s as eigenvalues. This means that all $\hat{\boldsymbol{p}}$'s satisfy the equation

$$\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}\hat{\boldsymbol{p}}_a = \hat{\boldsymbol{p}}_a\hat{\tau}_a.$$

Likewise, it is possible to show that the scores $\hat{\boldsymbol{t}}_a, a = 1, \ldots, k$ represent the corresponding eigenvalues of $\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}$, scaled to modulus $\sqrt{\hat{\tau}_a}$. In the final step, $\hat{\boldsymbol{t}}_i$ are used as regressors for the centered response variable $\tilde{\mathbf{y}}$. We thus fit the linear regression model using ordinary least squares as

$$\tilde{\mathbf{y}} = \hat{\boldsymbol{T}}\boldsymbol{q} + \boldsymbol{\xi},$$

since the data are centered, the above equation does not contain intercept. The estimated parameters are obtained as

$$\hat{\boldsymbol{q}} = (\hat{\boldsymbol{T}}^T\hat{\boldsymbol{T}})^{-1}\hat{\boldsymbol{T}}^T\tilde{\mathbf{y}},$$

and the fitted values

$$\hat{\mathbf{y}} = \hat{\boldsymbol{T}}\hat{\boldsymbol{q}} + \mathbf{1}\bar{y}$$
$$= (\boldsymbol{X} - \mathbf{1}\bar{\boldsymbol{x}})\hat{\boldsymbol{P}}\hat{\boldsymbol{q}} + \mathbf{1}\bar{y}.$$

The unknown regression parameters in terms of the original variables are estimated as

$$\hat{\boldsymbol{\beta}}_{PCR} = \hat{\boldsymbol{P}}\hat{\boldsymbol{q}},$$

$$\hat{\beta}_{0PCR} = \bar{y} - \bar{\boldsymbol{x}}\hat{\boldsymbol{\beta}}_{PCR}. \tag{3.4}$$

Finally, an estimator for the variance of the errors $\boldsymbol{\xi}$ can be calculated as

$$\begin{aligned} S_{\xi} &= \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)(y_i - \hat{y}_i)^T \\ &= S_y - \hat{\boldsymbol{q}}^T \boldsymbol{S}_t \hat{\boldsymbol{q}}, \end{aligned} \tag{3.5}$$

where S_y is variance of $\mathbf{y}$ and $\boldsymbol{S}_t$ is the empirical covariance matrix of the $\boldsymbol{t}$-variables. Note that equality (3.5) follows from the fact that the fitted values are orthogonal to the residuals (Johnson and Wichern (2002)).

For non high-dimensional data, it is possible to calculate all the positive eigenvalues and their eigenvectors simultaneously, either from the centered data $\tilde{\boldsymbol{X}}$, or through the $\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}$ cross product matrix using principal components decomposition. But the above algorithm is not proper for high-dimensional data. In such situations, PCR can be constructed with a singular value decomposition method or with the NIPALS (Nonlinear Iterative Partial Least Squares) algorithm (Wold (1966), Wold (1975a)) which is not respective to the number of observations to predictors ratio. NIPALS algorithm uses the fact that the components are orthogonal both in scores and loadings to extract one single factor at a time, $a = 1, \ldots, k$, starting from the factor with the largest eigenvalue. We use this algorithm since it can help us to understand the PCR method. Basically, NIPALS is just an iterative algorithm based on simple least squares regressions for computing principal components. For each factor, NIPALS algorithm uses an iterative method to obtain the loading vector $\hat{\boldsymbol{p}}_a$ and the score vector $\hat{\boldsymbol{t}}_a$ from the residual power-matrix (obtained after estimation of the previous $a-1$ factors). That initial residual matrix can be called $\hat{\boldsymbol{\nu}}$, but for facilitating the description of the PCA algorithm we may name it $\tilde{\boldsymbol{X}}_{a-1}$. It is defined by $\tilde{\boldsymbol{X}}_{a-1} = \tilde{\boldsymbol{X}} - \hat{\boldsymbol{t}}_1\hat{\boldsymbol{p}}_1^T - \cdots - \hat{\boldsymbol{t}}_{a-1}\hat{\boldsymbol{p}}_{a-1}^T$, where $\tilde{\boldsymbol{X}}$ is the matrix of mean centered $\boldsymbol{X}$-variables. A simple algorithm for computing the $a = 1, \ldots, k$ largest eigenvalues with corresponding eigenvectors proceeds as follows

1. Choose starting values, for instance $\hat{\boldsymbol{t}}_a$ is the column in $\tilde{\boldsymbol{X}}_{a-1}$ which has the largest remaining sum of squares (There are no restrictions on choosing the initial value, but choosing an appropriate initial value will speed up the convergence process);

2. Project the matrix $\tilde{\boldsymbol{X}}_{a-1}$ on $\hat{\boldsymbol{t}}_a$ to improve estimate of loading vector $\hat{\boldsymbol{p}}_a$

$$\hat{\boldsymbol{p}}_a^T = (\hat{\boldsymbol{t}}_a^T\hat{\boldsymbol{t}}_a)^{-1}\hat{\boldsymbol{t}}_a^T\tilde{\boldsymbol{X}}_{a-1};$$

3. Normalize loading vector $\hat{\boldsymbol{p}}$ to length 1 in order to avoid scaling ambiguity

$$\hat{\boldsymbol{p}}_a = \frac{\hat{\boldsymbol{p}}_a}{\sqrt{\hat{\boldsymbol{p}}_a^T\hat{\boldsymbol{p}}_a}};$$

4. Project the matrix $\tilde{\boldsymbol{X}}_{a-1}$ on $\hat{\boldsymbol{p}}_a$ in order to improve estimate of score $\hat{\boldsymbol{t}}_a$ for this factor

$$\hat{\boldsymbol{t}}_a = \tilde{\boldsymbol{X}}_{a-1}\hat{\boldsymbol{p}}_a(\hat{\boldsymbol{p}}_a^T\hat{\boldsymbol{p}}_a)^{-1};$$

5. Improve estimate of the eigenvalue $\hat{\tau}_a$

$$\hat{\tau}_a = \hat{\boldsymbol{t}}_a^T\hat{\boldsymbol{t}}_a;$$

6. Check for convergence: if the difference between $\hat{\tau}_a$, (i.e., $\hat{\tau}_{a(new)}$) and the $\hat{\tau}_a$, in the previous iteration (i.e., $\hat{\tau}_{a(old)}$) is smaller than $threshold \times \hat{\tau}_{a(new)}$ (for instance, threshold may be 0.0001), the method has converged for this factor; if not, return to step 2.

   Subtract the effect of this factor

$$\tilde{\boldsymbol{X}}_a = \tilde{\boldsymbol{X}}_{a-1} - \hat{\boldsymbol{t}}_a \hat{\boldsymbol{p}}_a^T$$

and go to step 1 for the next factor. Repeat steps 2-6 until convergence.

In the ideal situation, only a few eigenvalues will be large such that almost all the variation in $\boldsymbol{X}$ will be described by those first few components. One of the advantages of PCR is that it can yield good results where $\boldsymbol{X}$-variables are highly correlated, as long as the predictive information is in the first eigenvectors. However, PCR is not effective in the following situations:

1. Since the principal components are linear combinations of the predictor variables, little is achieved if these are not interpretable. In general, the predictor variables should have measurements of comparable quantities for interpretation to be possible;

2. Principal component uses only the $\boldsymbol{X}$ matrix and not the response, therefore it may happen that some low eigenvalue principal component highly contributes to the overall predictive ability of the regression model. Thus, there is no definite guarantee that PCR predicts well.

3.2.2 Partial Least Squares Regression Algorithm

Partial least squares is a family of regression based methods which combines features from principal component analysis and multiple regression. Therefore, it contains regression and classification tasks, as well as dimension reduction techniques and modeling tools. The underlying assumption of all PLS methods is that the observed data is generated by a system or process which is driven by a small number of latent (not directly observed or measured) variables (Rosipal and Krämer (2006)). Projections of the observed data onto its latent structure by means of PLS was developed by Herman Wold for econometrics (Wold (1966)) and then applied to other fields such as chemometrics, social science, food research, etc (Wold *et al.* (1984), Geladi and Kowalski (1986), Martens and Martens (2000), Hulland (1999)).

In its general form, PLS constructs orthogonal score vectors (also called latent vectors or components) by maximizing the covariance between $\mathbf{y}$ and all possible linear functions of $\boldsymbol{X}$ (Frank (1987), Næs *et al.* (2002)). PLSR can be useful when the purpose is to predict the response and not necessarily trying to understand the underlying relationship between the predictor variables.

According to (3.1) and (3.2), the bilinear structure proceeds in two steps in PLSR algorithm. Similar to PCR, denote the mean-centered of the $\boldsymbol{X}$ and $\mathbf{y}$ with $\tilde{\boldsymbol{X}}$ and $\tilde{\mathbf{y}}$, respectively. The normalized PLSR weight vectors $\boldsymbol{w}_a$ $(\|\boldsymbol{w}_a\| = 1)$ are then defined as the vectors that maximize for each $a = 1, \ldots, k$

$$cov(\tilde{\mathbf{y}}, \tilde{\boldsymbol{X}}\boldsymbol{w}_a) = \frac{\tilde{\mathbf{y}}^T\tilde{\boldsymbol{X}}}{n-1}\boldsymbol{w}_a = \boldsymbol{S}_{yx}\boldsymbol{w}_a, \tag{3.6}$$

where $\boldsymbol{S}_{yx}^T = \boldsymbol{S}_{xy} = \dfrac{\tilde{\boldsymbol{X}}^T\tilde{\mathbf{y}}}{n-1}$ is the empirical cross-covariance matrix between the $\boldsymbol{X}$- and $\mathbf{y}$-variables. The maximization problem of (3.6) is equivalent to finding the unit weight vector $\hat{\boldsymbol{w}}_a$ that maximizes $\tilde{\mathbf{y}}^T\tilde{\boldsymbol{X}}\boldsymbol{w}_a$. For each latent variable $a = 1, \ldots, k$, PLSR algorithm is proceeded as follows:

1. Use the variability remaining in $\tilde{\mathbf{y}}$ to find the loading weight $\hat{\boldsymbol{w}}_a$, using least squares and the local model

   $$\tilde{\boldsymbol{X}} = \tilde{\mathbf{y}}\boldsymbol{w}_a^T + \boldsymbol{\nu}_1$$

   and scale the vector to modulus 1. The solution is of the form

   $$\hat{\boldsymbol{w}}_a = c\tilde{\boldsymbol{X}}^T\tilde{\mathbf{y}}, \tag{3.7}$$

   where c is the scaling factor that makes the modulus of the final $\hat{\boldsymbol{w}}_a$ equal to 1, i.e.

$$c = (\tilde{\mathbf{y}}^T \tilde{\boldsymbol{X}} \tilde{\boldsymbol{X}}^T \tilde{\mathbf{y}})^{-1/2};$$

2. The elements of the scores $\hat{\boldsymbol{t}}_a$ are then defined as linear combinations of the mean-centered data. It is estimated as the projection of $\tilde{\boldsymbol{X}}$ on $\hat{\boldsymbol{w}}_a$, i.e.,

$$\tilde{\boldsymbol{X}} = \boldsymbol{t}_a \hat{\boldsymbol{w}}_a^T + \boldsymbol{\nu}_2.$$

The least squares solution is (since $\hat{\boldsymbol{w}}_a^T \boldsymbol{w}_a = 1$)

$$\hat{\boldsymbol{t}}_a = \tilde{\boldsymbol{X}}_{n \times p} \hat{\boldsymbol{w}}_a; \tag{3.8}$$

3. The loading $\hat{\boldsymbol{p}}_a$ is obtained from simple linear regression of the columns of $\tilde{\boldsymbol{X}}$ on $\hat{\boldsymbol{t}}_a$, i.e.

$$\tilde{\boldsymbol{X}} = \boldsymbol{t}_a \boldsymbol{p}_a^T + \boldsymbol{\nu}_3,$$

which gives the least squares solution

$$\hat{\boldsymbol{p}}_a = \frac{\tilde{\boldsymbol{X}}^T \hat{\boldsymbol{t}}_a}{\hat{\boldsymbol{t}}_a^T \hat{\boldsymbol{t}}_a}; \tag{3.9}$$

4. Regress $\tilde{\mathbf{y}}$ on $\hat{\boldsymbol{t}}_a$ to find the loading $\hat{q}_a$, i.e.

$$\tilde{\mathbf{y}} = \hat{\boldsymbol{t}}_a q_a + \boldsymbol{\xi};$$

which gives the solution

$$\hat{q}_a = \frac{\tilde{\mathbf{y}}^T \hat{\boldsymbol{t}}_a}{\hat{\boldsymbol{t}}_a^T \hat{\boldsymbol{t}}_a}; \tag{3.10}$$

5. Create new $\tilde{\boldsymbol{X}}$ and $\tilde{\mathbf{y}}$ matrix by subtracting the estimated effect of this factor

$$\hat{\boldsymbol{\nu}} = \tilde{\boldsymbol{X}} - \hat{\boldsymbol{t}}_a \hat{\boldsymbol{p}}_a^T, \tag{3.11}$$

$$\hat{\boldsymbol{\xi}} = \tilde{\mathbf{y}} - \hat{\boldsymbol{t}}_a \hat{q}_a. \tag{3.12}$$

Note that $\boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \boldsymbol{\nu}_3$ represent different residual in steps 1-3, respectively, although little distinction has been made notationally. Update the former $\tilde{\boldsymbol{X}}$ and $\tilde{\mathbf{y}}$ by the new residuals $\hat{\boldsymbol{\nu}}$ and $\hat{\boldsymbol{\xi}}$, i.e. set

$$\tilde{\boldsymbol{X}} = \hat{\boldsymbol{\nu}},$$
$$\tilde{\mathbf{y}} = \hat{\boldsymbol{\xi}};$$

6. If a is less than k, then increase its value by one and return to step 1. If a is equal to k, then the algorithm stops and the estimated vector $\hat{\mathbf{y}}$ is

$$\hat{\mathbf{y}} = \hat{q}_1 \hat{\boldsymbol{t}}_1 + \hat{q}_2 \hat{\boldsymbol{t}}_2 + \cdots + \hat{q}_k \hat{\boldsymbol{t}}_k + \mathbf{1}\bar{y}.$$

In terms of the original predictor variables, the estimates for $\hat{\boldsymbol{\beta}}_{PLSR}$ and $\hat{\beta}_{0PLSR}$ are achieved through equations (3.7) to (3.10)

$$\hat{\boldsymbol{\beta}}_{PLSR} = \hat{\boldsymbol{W}}(\hat{\boldsymbol{P}}^T \hat{\boldsymbol{W}})^{-1} \hat{\boldsymbol{q}},$$
$$\hat{\beta}_{0PLSR} = \bar{y} - \bar{\boldsymbol{x}} \hat{\boldsymbol{\beta}}_{PLSR}. \tag{3.13}$$

Here, the loading weights $\hat{\boldsymbol{W}} = (\hat{\boldsymbol{w}}_1, \ldots, \hat{\boldsymbol{w}}_k)$ achieved are orthogonal, and so are the scores $\hat{\boldsymbol{T}} = (\hat{\boldsymbol{t}}_1, \ldots, \hat{\boldsymbol{t}}_k)$. The estimated loadings $\hat{\boldsymbol{P}} = (\hat{\boldsymbol{p}}_1, \ldots, \hat{\boldsymbol{p}}_k)$ are however generally nonorthogonal for PLSR, although they usually resemble $\hat{\boldsymbol{W}}$. It should also be mentioned that substraction from $\tilde{\mathbf{y}}$ in step 5 is unnecessary for the model estimation (Manne (1987)), but it simplifies the computation of the validation statistics.

In order to perform the above algorithm it is necessary that k be known, but in practice k is unknown and should be estimated. In a case when $k = p$, the PLSR predictor is equal to the ordinary LS predictor, if it exists. In another special case, when the $\boldsymbol{x}$-variables are orthonormal, i.e. orthogonal and scaled to modulus one, so that $\boldsymbol{X}^T \boldsymbol{X} = \mathbf{I}$, PLSR with one component and LS gives the same predictor. But this situation is rarely encountered in multivariate regression. Usually k is considered equal to the maximum number of PLSR components to be calculated. This number should be higher than the number of phenomena (informative components) expected to be seen in $\boldsymbol{X}$, in order to allow unexpected phenomena to be modeled.

3.3 Resampling Methods

In this section, we describe two resampling computer-based methods of statistical inference that can answer many real statistical questions. These methods give a direct appreciation of variance, bias, confidence intervals, and other probabilistic phenomena. We use these computational techniques to analyze and understand high-dimensional data sets. Here, these methods are used to estimate the standard error of PCR and PLSR regression coefficients for significance tests of regression coefficients which enable PCR and PLSR to perform variable selection. We will describe in detail the jack-knife and bootstrap based methods for variable selection in Sections 3.4.2 and 3.4.3.

3.3.1 Jack-knife

Jack-knife (JK) was the original computer-based method for estimating biases and standard errors (Efron and Tibshirani (1994)). Quenouille in mid-1950's proposed the idea of jack-knife for bias estimation (Quenouille (1949)). Tukey recognized the potential of jack-knife for estimating the standard errors (Tukey (1958)). Further development was presented by Miller (1964), Miller (1974), Gray and Schucany (1972), Hinkley (1977), Reeds (1978)), Hinkley and WET (1984)), Sen (1988), and Wu (1986) in the linear regression setting. Shao and Wu (1989), and Shao (1991) presented general theoretical results on the deleted-d jack-knife.

Suppose we have a random sample $\mathbf{X} = (\mathrm{X}_1, \ldots, \mathrm{X}_n)$ from an unknown probability

distribution F and we wish to estimate a parameter of interest θ on the basis of **X**. For this purpose, we calculate $\hat{\theta}$ from x. In order to see how accurate $\hat{\theta}$ is, we may estimate the standard error of $\hat{\theta}$. By deleted-d jack-knife the data set is randomly partitioned into L non-overlapping equal size subsets ($d = n/L$ is the size of each subset). Delete one subset from the training set, use the remaining $L-1$ subsets to obtain the estimates, and repeat this process in turn for each of the L subsets. The f-th, $f = 1, \dots, L$ jack-knife sample contains the data set with the d-th observation of the original sample removed. When L is set to the number of observations n, one observation is left out at a time instead of leaving out d observations each time. Consider $\hat{\theta}^{(f)}$, be the f-th replication jack-knife of $\hat{\theta}$. In the other words, $\hat{\theta}^{(f)}$ is the estimate of θ with the f-th subset removed.
The jack-knife estimate of the standard error is defined by

$$\hat{se}_{JK}(\hat{\theta}) = \left(\frac{L-1}{L} \sum_{f=1}^{L} (\hat{\theta}^{(f)} - \bar{\theta}^{L})^2 \right)^{1/2}, \tag{3.14}$$

where

$$\bar{\theta}^{L} = \frac{1}{L} \sum_{f=1}^{L} \hat{\theta}^{(f)}. \tag{3.15}$$

3.3.2 Bootstrap

The basic idea of bootstrap is to generate a large number of bootstrap samples by sampling with replacement from the original data set. Drawing B times, we obtain B independent bootstrap data sets, each of the same size as the original data set. From each bootstrap sample, the statistical parameter of interest is estimated and we obtain B bootstrap estimates of parameter that can be used to assess many aspects of the distribution of the estimator of interest.
Like jack-knife, suppose we have a random sample $\mathbf{X} = (\mathrm{X}_1, \dots, \mathrm{X}_n)$ from an unknown probability distribution function F, the bootstrap (Boot) was introduced in 1979 as a

computer-based method for estimating the standard error of $\hat{\theta}$. It enjoys the advantage of being completely automatic. The bootstrap estimate of standard error requires no theoretical calculations, and is available no matter how mathematically complicated the estimator $\hat{\theta}$ may be (Efron and Tibshirani (1994)).

Bootstrap methods depend on the notion of a bootstrap sample. Let $\hat{F}$ be the empirical distribution, putting probability $1/n$ on each of the observation values $x_i, i = 1, \ldots, n$, a bootstrap sample is defined to be a random sample of size n drawn from $\hat{F}$, say $\mathbf{X}^* = (\mathsf{X}_1^*, \ldots, \mathsf{X}_n^*)$,

$$\hat{F} \to \mathbf{X}^* = (\mathsf{X}_1^*, \ldots, \mathsf{X}_n^*). \tag{3.16}$$

The star notation shows that $\mathbf{X}^*$ is not the actual data set $\mathbf{X}$ but rather a randomized, or resampled version of $\mathbf{X}$.

In the other words, (3.16) indicates that the bootstrap data points $\mathsf{X}_1^*, \ldots, \mathsf{X}_n^*$ are a random sample of size n drawn with replacement from the population of n objects $(\mathsf{X}_1, \ldots, \mathsf{X}_n)$. The bootstrap data set $(\mathsf{X}_1^*, \ldots, \mathsf{X}_n^*)$ consists of members of the original data set $(\mathsf{X}_1, \ldots, \mathsf{X}_n)$, some appearing zero times, some appearing once, some appearing twice, etc.

Corresponding to a bootstrap data set $\mathbf{X}^*$ there is a bootstrap replication of $\hat{\theta}$, which is denoted by $\hat{\theta}^*$. For instance, if $\hat{\theta}$ is the sample mean $\bar{\mathsf{X}}$, then $\hat{\theta}^*$ is the mean of the bootstrap data set, $\bar{\mathsf{X}}^* = \frac{1}{n} \sum_{i=1}^{n} \mathsf{X}_i^*$. The bootstrap estimate of $se_F(\hat{\theta})$, the standard error of a statistic $\hat{\theta}$, is an estimate which uses the empirical distribution $\hat{F}$ in place of the unknown distribution F. The bootstrap estimate of $se_F(\hat{\theta})$ is the standard error of $\hat{\theta}$ for data sets of size n randomly sampled from $\hat{F}$.

The bootstrap algorithm is a computational way of obtaining a good approximation to the numerical value of $se_{\hat{F}}(\hat{\theta}^*)$. The bootstrap sampling can be easily implemented on the computer. A random number device generates integers $i_1, i_2, \ldots, i_n$, each of which can get a value between 1 and n with probability $1/n$. The bootstrap sample contains the corresponding members of **X**,

$$\mathsf{X}_1^* = \mathsf{X}_{i_1}, \mathsf{X}_2^* = \mathsf{X}_{i_2}, \ldots,\ \mathsf{X}_n^* = \mathsf{X}_{i_n}. \tag{3.17}$$

The bootstrap algorithm proceeds by sampling many independent bootstrap samples, evaluating the corresponding bootstrap replications, and estimating the standard error of $\hat{\theta}$ by the empirical standard deviation of the replications. The bootstrap estimate of standard error is denoted by $\hat{se}_{Boot}$.

Next, we describe the bootstrap algorithm for estimating the standard error of $\hat{\theta}$ from the observed data **X**.

1. Select B independent bootstrap samples $\mathbf{X}_1^*, \mathbf{X}_2^*, \ldots, \mathbf{X}_B^*$, each consisting of n data values drawing with replacement from **X**;

2. Evaluate the bootstrap replication corresponding to each bootstrap sample

$$\hat{\theta}^{*(b)}, \quad b = 1, 2, \ldots, B; \tag{3.18}$$

3. Estimate the standard error $se_F(\hat{\theta})$ by the sample standard error of the B replications

$$\hat{se}_{Boot}(\hat{\theta}) = \left\{ \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\theta}^{*(b)} - \bar{\theta}^* \right)^2 \right\}^{1/2}, \tag{3.19}$$

where $\bar{\theta}^* = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}^{*(b)}$.

3.4 Variable Selection for PCR and PLSR

Principal components regression (PCR) and partial least squares regression (PLSR) are widely used in data processing and dimension reduction. However, these two methods suffer from the fact that each component inside them is a linear combination of

all the original variables, and thus it is often difficult to interpret the results. Hence, it is important to select the variables that are relevant in order to obtain the best model in terms of prediction accuracy and with the lowest number of variables.
Here, we describe several existing methods for variable selection in PCR and PLSR such as sparse formulation, jack-knife based approaches, and also we introduce bootstrap-based approach for performing variable selection in PCR and PLSR.

3.4.1 Sparse Formulation Based Methods

Sparse formulation based methods for PCR and PLSR (i.e., sparse principal components (SPC) and sparse partial least squares (SPLS) regression) use this fact that PC can be written as a regression type optimization problem; therefore, LASSO (elastic net) (Tibshirani (1996), Zou and Hastie (2005)) can be integrated into the regression criterion to produce sparse linear combinations of the original predictors.

Sparse Principal Components Regression

Sparse principal component (SPC) constructs modified principal components (PCs) with sparse loadings by considering PC as a regression-type optimization problem and imposing the LASSO (elastic net) constraint into the regression criterion to produce sparse loadings. Suppose we have matrix of predictors, $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p]$, and, without loss of generality, assume the column means of $\boldsymbol{X}$ are all zero. Consider the singular value decomposition (SVD) of $\boldsymbol{X}$ as

$$\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{V}^T, \tag{3.20}$$

where $\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{D}$, are the PCs, and the columns of $\boldsymbol{V}$ are the corresponding loadings of the PCs. The sample variance of the i-th PC is $\frac{\boldsymbol{D}_{ii}^2}{n}$ (Zou *et al.* (2006)).
Considering the fact that PC can be written as a regression-type optimization problem,

the following optimization problem which is called the SPCA criterion

$$(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) = \arg\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \sum_{i=1}^{n} |\mathbf{x}_i - \boldsymbol{\alpha}\boldsymbol{\beta}^T\mathbf{x}_i|^2 + \lambda \sum_{j=1}^{k} |\beta_j|^2 + \sum_{j=1}^{k} \lambda_{1,j}|\beta_j|_1 \qquad (3.21)$$

$$\text{subject to } \boldsymbol{\alpha}^T\boldsymbol{\alpha} = \mathbf{I}_k,$$

should be solved to obtain the sparse loadings, with k, the number of PCs.
Then, $\hat{\beta}_j \propto v_j$ for $j = 1, \ldots, k$. Whereas the same λ (L_2-penalty) is used for all components, different $\lambda_{1,j}, j = 1, \ldots, k$, ($L_1$-penalty) are allowed for penalizing the loadings of different principal components (in high dimensional data, $\lambda \to \infty$ and $\lambda_{1,j}$ were used such that the number of nonzero loadings varied in a rather wide range, Zou *et al.* (2006)).
The obtained sparse principal components will become the regressors for the response variable to perform SPC regression.

Sparse Partial Least Squares Regression

SPLS regression was formulated by relating it to SPCA. To specify the $n \times k$ latent component matrix $\boldsymbol{T} = (\boldsymbol{t}_1, \ldots, \boldsymbol{t}_k)$, such that $\boldsymbol{T} = \boldsymbol{X}\boldsymbol{W}$, the following minimization problem should be solved to find the columns of $\boldsymbol{W} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k)$.

$$\min_{\boldsymbol{w},\boldsymbol{c}}\{-\kappa\boldsymbol{w}^T\boldsymbol{M}\boldsymbol{w} + (1-\kappa)(\boldsymbol{c}-\boldsymbol{w})^T\boldsymbol{M}(\boldsymbol{c}-\boldsymbol{w}) + \lambda_1||\boldsymbol{c}||_1 + \lambda_2||\boldsymbol{c}||_2^2\}, \qquad (3.22)$$

$$\text{subject to } \boldsymbol{w}^T\boldsymbol{w} = 1,$$

where $\boldsymbol{M} = \boldsymbol{X}^T\mathbf{y}\mathbf{y}^T\boldsymbol{X}$. This formulation promotes exact zero property by imposing L_1- penalty (λ_1) onto a surrogate of direction vector ($\boldsymbol{c}$) instead of the original direction vector ($\boldsymbol{w}$), while keeping $\boldsymbol{w}$ and $\boldsymbol{c}$ close to each other (Chun and Keleş (2010)).
In this formulation, the L_1-penalty (λ_1) encourages sparsity on $\boldsymbol{c}$, whereas the L_2-

penalty (λ_2) addresses the potential singularity in $\boldsymbol{M}$ when solving for $\boldsymbol{c}$. Although there are four tuning parameters $(\kappa, \lambda_1, \lambda_2, k)$ in the SPLS regression formulation (3.22), only two of these are main tuning parameters, namely the sparsity parameter λ_1 and the number of hidden components k. These two parameters are chosen by L-fold cross-validation (Chun and Keleş (2010), Chung *et al.* (2013)).
The obtained sparse latent components will become the regressors for the response variable to perform SPLS regression.

3.4.2 Jack-knife PLSR/PCR Algorithm

Jack-knife partial least squares (JKPLS) algorithm for variable selection, proposed by Westad and Martens (2000), is based on the significance tests of the regression coefficients in the PLSR model. This technique automatically generates a standard error estimate of the coefficients parameters and then, by dividing the estimated regression coefficients by their estimated standard errors, the t-test values (or equivalently F-test, (Weisberg (2014))) will be obtained, giving the significance level for each parameter. Therefore, useless variables may be eliminated automatically, in order to simplify the model and make it more reliable. By JK method, the data set is divided into L non-overlapping subsets of equal size. In order to obtain the structure model's regression coefficients, one subset from the training set is deleted and the remaining subsets are used. This process repeats in turn for each of the subsets. The jack-knife estimated standard error for the estimated regression coefficients (similar to (3.14)) is given by

$$\hat{se}_{JK}(\hat{\beta}_j) = \left\{ \frac{L-1}{L} \sum_{f=1}^{L} \left(\hat{\beta}_j^{(f)} - \bar{\beta}_j \right)^2 \right\}^{1/2}, \tag{3.23}$$

where L is the number of JK subsets, $\hat{\beta}_j^{(f)}$ is the estimated coefficient with the f-th subset removed, $\bar{\beta}_j$ is the mean value of $\hat{\beta}_j^{(f)}$, $f = 1, \ldots, L$, for the j-th variable, $j = 1, \ldots, p$.

Since the collinearity between predictors in high-dimensional data is high, it is not to be expected that the significance test for a data set with large number of variables will converge to a stable set of variables in the first iteration. This algorithm eliminates the irrelevant variables iteratively whilst improving the model's predictive ability. We also use the jack-knife approach for principal components (JKPC) regression model.

3.4.3 Bootstrap PLSR/PCR Algorithm

As mentioned in Section 3.3.2, the basic idea of bootstrap is to generate a large number of bootstrap samples by sampling with replacement from the original data set. Drawing B times, we obtain B independent bootstrap data sets, each of the same size as the original data set. From each bootstrap sample, the statistical parameter of interest is estimated and we obtain bootstrap estimates of parameter that can be used to assess many aspects of the distribution of the estimator of interest. The merit of using bootstrap here is that it can automatically generate a standard error estimate which is difficult to compute analytically. Then a t-test statistic (or equivalently F-test, (Weisberg (2014))) is obtained by dividing the regression coefficients themselves to their bootstrap estimated standard error. Using this t-test statistic, useless variables may be eliminated automatically in order to perform variable selection. The bootstrap algorithm on partial least squares (BootPLS) regression is based on significance tests of the estimated regression coefficients in the PLSR model. In this approach, the standard errors of the estimated regression coefficients are estimated by resampling. We consider the following bootstrap method to construct bootstrap samples (Wehrens and Van der Linden (1997)): resample with replacement from the set of pairs $(y_1, \mathbf{x}_1)$, $\ldots, (y_n, \mathbf{x}_n), y_i \in \mathbb{R}, \mathbf{x}_i \in \mathbb{R}^p, i = 1, \ldots, n$, and generate the bootstrapped pairs $(y_1^*, \mathbf{x}_1^*), \ldots,$ $(y_n^*, \mathbf{x}_n^*), y_i^* \in \mathbb{R}, \mathbf{x}_i^* \in \mathbb{R}^p, i = 1, \ldots, n$. Because sampling is performed with replacement, in almost all bootstrap samples, the data occur more than once. Usually, there is collinearity in high-dimensional data, and therefore it is not to be expected that the significance test for such a data set will converge to a stable set of predictor variables in the first iteration. We use an optional procedure in order to eliminate the irrelevant predictors iteratively whilst improving the model's predictive ability. For PCR and PLSR, one of the important aspects is to find the optimal number of components. The optimal number of components is obtained by considering its learning curves. For this purpose, we plot the number of components versus $RMSE$ values as measured by cross-validation. The optimal number of components can be chosen as the point where the learning curve does not show a considerable slope anymore. We fixed the optimum number of components based on the lowest $RMSE$ value of cross-validation,

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}, \quad i = 1, \ldots, n, \tag{3.24}$$

where $\hat{y}_i$ is the fitted value of the response variable.
The steps of the algorithm are as follows:

1. The data set is standardized using its mean and variance. B bootstrap samples are generated from the data and PLSR is performed on each of bootstrap samples to calculate the $RMSE$ value as the model's predictive ability using cross-validation. The minimum $RMSE$ value across these bootstrap samples is used as the starting value;

2. The PLSR method for each $b = 1, \ldots, B$ bootstrap samples is performed and the estimated regression coefficient for the b-th bootstrap replication parameter is stored;

3. The bootstrap estimated standard error (similar to (3.19)) for the estimated regression coefficient, $\hat{\beta}_j, j = 1, \ldots, p$, across the bootstrap samples is computed

$$\hat{se}_{Boot}(\hat{\beta}_j) = \left\{ \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\beta}_j^{*(b)} - \bar{\beta}_j^{*} \right)^2 \right\}^{1/2}, \tag{3.25}$$

where $\bar{\beta}_j^{*} = \frac{1}{B} \sum_{b=1}^{B} \hat{\beta}_j^{*(b)}$;

4. A two-sided t-test ($n-1$ degrees of freedom) is performed for each regression coefficient in the model, giving the significance level for each parameter

$$t_j = \frac{\bar{\beta}_j^{*}}{\hat{se}_{Boot}(\hat{\beta}_j)}; \tag{3.26}$$

5. The regression coefficients with p-values smaller than $\alpha = 0.05$ (we fix a significance level of 0.05 for statistical testing) are selected and PLSR is performed with the new variable set. While the increase in the new $RMSE$ value is lower than a punishment factor (1% is chosen as the punishment factor), the predictor variables for which $H_0 : \beta_j = 0$ are accepted at the fixed significance level (i.e., these predictors are removed to form the new model) and the algorithm repeats from step 2 for the next iteration. If the increase in $RMSE$ value in comparison with the previous $RMSE$ value is greater than the punishment factor, the previous set of predictors is kept and the algorithm is terminated.

In this thesis, we also extend the bootstrap approach for principal components (Boot PC) regression model. The steps of BootPLS algorithm are summarized in Figure 3.2.

3.5 Simulation Study

We perform a simulation study (Luo (2008)) in order to investigate the performance of variable selection algorithms in PCR and PLRS. The simulation is designed to evaluate the algorithms' performance in reducing the number of predictors and the

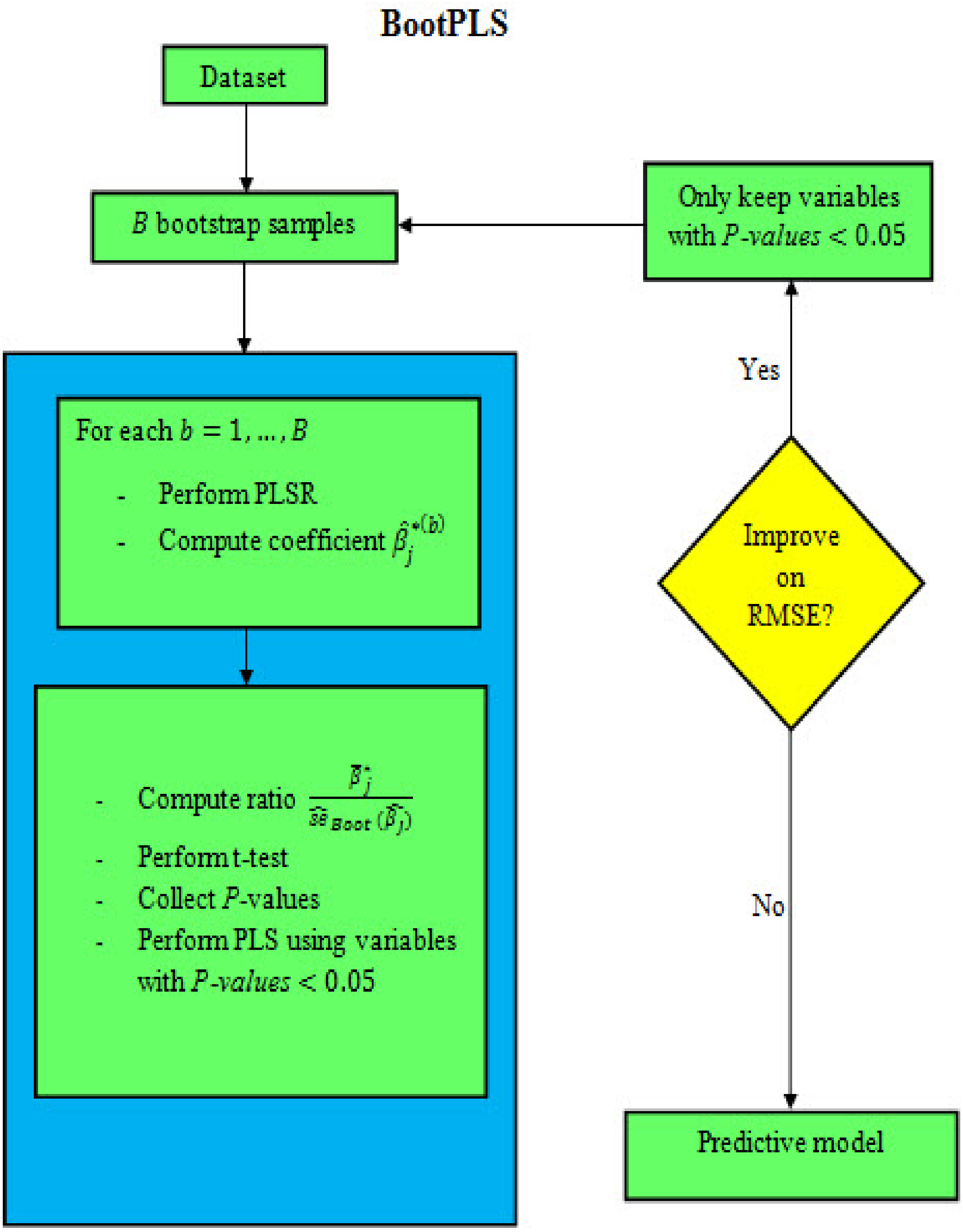
BootPLS
Dataset
B bootstrap samples
Only keep variables with P-values < 0.05
For each $b = 1, \ldots, B$
- Perform PLSR
- Compute coefficient $\hat{\beta}_j^{*(b)}$
Yes
Improve on RMSE?
- Compute ratio $\frac{\bar{\beta}_j^*}{\widehat{se}_{Boot}(\hat{\beta}_j)}$
- Perform t-test
- Collect P-values
- Perform PLS using variables with P-values < 0.05
No
Predictive model

Figure 3.2: Flowchart summarizes BootPLS algorithm.

predictive ability in terms of the $RMSE$ value. We use the freeware R to develop the simulation study (R Core Team (2014)). We use the R-package pls (Mevik *et al.* (2013)) for computing the PCR and PLSR estimates, and the R packages spls (Chung *et al.* (2013)), and elasticnet (Zou and Hastie (2012)) to compute SPLS and SPC estimates, respectively.

In our simulations we generate three models considering the following scenarios with the common model

$$y_i = \beta_0 + \mathbf{x}_i^T \boldsymbol{\beta} + e_i, \quad i = 1, \ldots, n. \tag{3.27}$$

Model 1 - We generate a high-dimensional data set with $p = 80$ predictors and the sample size $n = 20$, with $\beta_0 = 5$ with the following regression coefficients

$$\beta_j = \begin{cases} 3, & if \quad j = 1, 2, 3 \\ -2, & if \quad j = 10, 11 \\ 6, & if \quad j = 21, 22 \\ 5, & if \quad j = 30, 31 \\ 4, & if \quad j = 80 \\ 0, & \textit{otherwise} \end{cases} .$$

Therefore, there are 10 active and 70 inactive predictors in simulation model 1.

Model 2 - We generate a high-dimensional data set with $p = 100$ predictors and the sample size $n = 20$, $\beta_0 = 5$ with the following regression coefficients

$$\beta_j = \begin{cases} 3, & if \quad j = 1, 2, 3 \\ -2, & if \quad j = 11, 12 \\ 6, & if \quad j = 21, 22 \\ 5, & if \quad j = 32, 33 \\ 4, & if \quad j = 100 \\ 0, & \textit{otherwise} \end{cases} .$$

Therefore, there are 10 active and 90 inactive predictors in simulation model 2.

Model 3 - We generate a high-dimensional data set with $p = 100$ predictors and the sample size $n = 20$, $\beta_0 = 18$ with the following regression coefficients

$$\beta_j = \begin{cases} 6, & if \quad j = 10, 11 \\ -8, & if \quad j = 20, 21 \\ 13, & if \quad j = 30, 31 \\ 15, & if \quad j = 40, 41 \\ 13, & if \quad j = 50 \\ 52, & if \quad j = 51 \\ -11, & if \quad j = 71 \\ 6, & if \quad j = 90 \\ 0, & \textit{otherwise} \end{cases} .$$

Therefore, there are 12 active and 88 inactive predictors in simulation model 3.

We generate samples $\mathbf{x}_i$ by drawing from a multivariate normal distribution with mean $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$. The vector $\boldsymbol{\mu}$ and the diagonal entries of the covariance matrix Σ are selected by sampling randomly with replacement from -200, $\ldots$, 200 and 1, $\ldots$, 200, respectively. The block diagonal matrix Σ has 10 equally-sized blocks, and the correlation within all pairs of the first, second,..., and tenth $p/10$ variables is set to 0.9,0.8, $\ldots$, 0.1, 0, respectively.

To compute y_i we sample the error term from a normal distribution with mean 0 and standard deviation σ. The value of σ is chosen such that the signal to noise ratio is equal to 10, i.e.,

$$\sqrt{var(y_i)/var(e_i)} = \sqrt{var(y_i)}/\sigma = 10, \quad i = 1, \ldots, n.$$

We investigate several highly correlated variables and one uncorrelated variable in the first two models, while the third model includes variables with partly large coefficients from each correlation block. We use the partly large coefficients for the third model in comparison with the first two models in order to make the active predictors more correlated to the response variable.

The global minimum $RMSE$ value for the JK based on the $L = 10$ fold cross-validation and for the Boot based on $B = 200$ bootstrap samples from the data was used as a starting value.

3.5.1 Performance Measurements

We generate $G = 100$ data sets of each type of simulation model independently, and each time we perform all the aforementioned variable selection (VS) methods on the same data set.

In order to examine the performance of the variable selection methods in terms of the predictive ability, dimension reduction and accuracy of variable selection, we report:

- The $RMSE$ value;

- The average value of reduced predictors (RP)

$$RP = \frac{\#reduced\ \ predictors}{G}; \tag{3.28}$$

- The average value of true positive rate (TPR)

$$TPR = \frac{1}{G}\sum_{g=1}^{G} TPR^{(g)}, \tag{3.29}$$

where $TPR^{(g)}$ is the true positive rate of the g-th replication

$$TPR^{(g)} = \frac{\#truly\ \ selected\ \ predictors}{\#active\ \ predictors}, \tag{3.30}$$

where the active predictors are those predictors that are in the true model;

- The average value of false negative rate (FNR)

$$FNR = \frac{1}{G}\sum_{g=1}^{G} FNR^{(g)}, \tag{3.31}$$

where $FNR^{(g)}$ is the false negative rate of the g-th replication

$$FNR^{(g)} = \frac{\#falsely\ \ selected\ \ predictors}{\#inactive\ \ predictors}, \tag{3.32}$$

where inactive predictors are those predictors that are not in the true model.

3.6 Simulation Study Results

Figures 3.3, 3.4, and 3.5 show the box plots of the $RMSE$ values across the $G = 100$ replications for PCR, PLSR, and the variable selection methods (JKPLS, JKPC, BootPLS, BootPC, SPLS, and SPC). The $RMSE$ values were obtained by performing

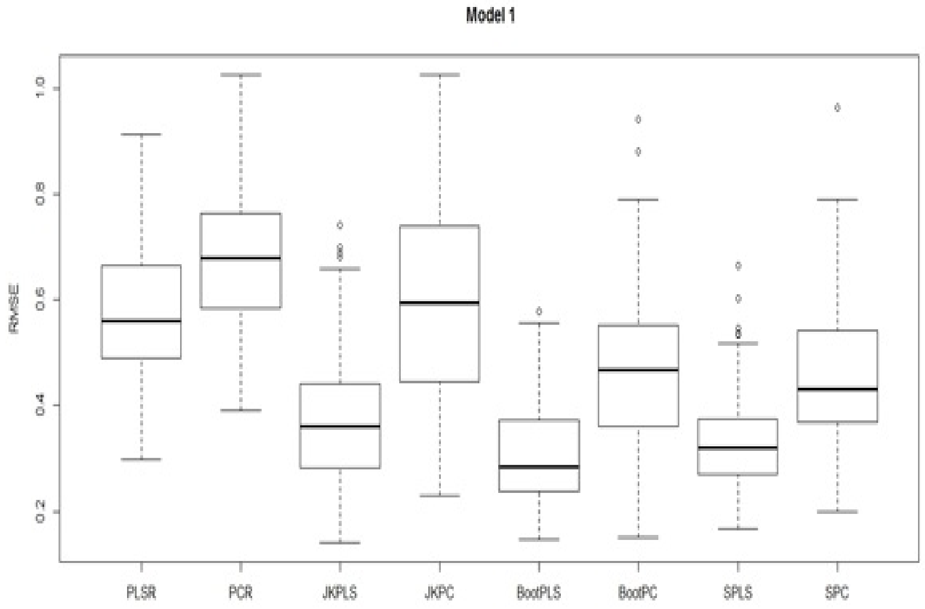

Figure 3.3: Comparison between the predictive ability of PLSR, PCR, JKPLS, JKPC, BootPLS, BootPC, SPLS, and SPC for Model 1.

all methods each time on the same data set through the $G = 100$ independent generated data sets for the first, second and third simulation model, respectively. The optimal number of components for PCR on models 1, 2, and 3 was considered equal to 6, 6, and 8, respectively and the optimal number of components for PLSR on models 1, 2, and 3 was considered equal to 3, 3, and 5, respectively.

Using variable selection methods on PCR and PLSR considerably reduces the $RMSE$ value. PLSR performs better than PCR in prediction, and the prediction performance ability by VS methods on PLSR is higher than the prediction performance ability while applying VS methods on PCR for all three models. BootPLS and SPLS show a similar prediction performance in all three models and both perform better than JKPLS. BootPC and SPC, in all three models, show the similar prediction error. In all three models, the prediction performance ability by JKPC is worse than other

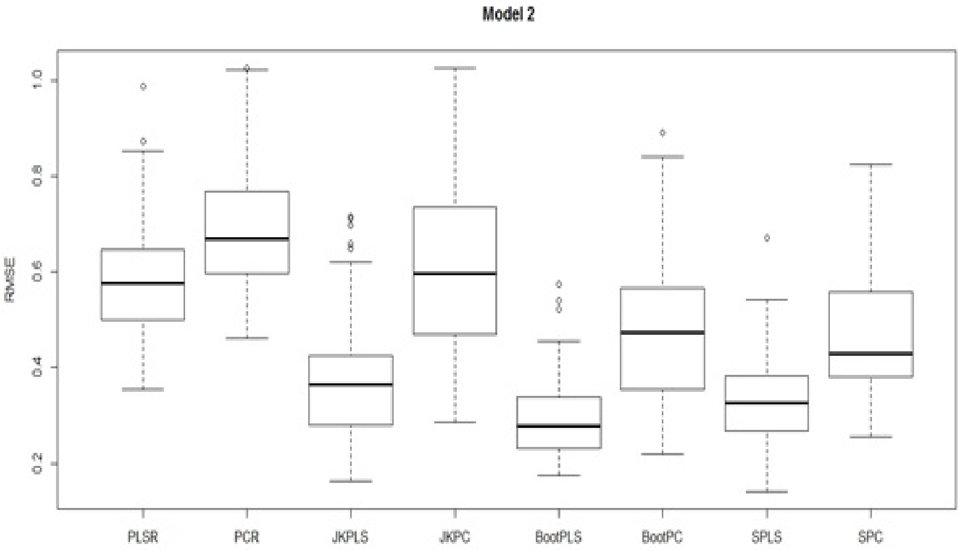

Figure 3.4: Comparison between the predictive ability of PLSR, PCR, JKPLS, JKPC, BootPLS, BootPC, SPLS, and SPC for Model 2.

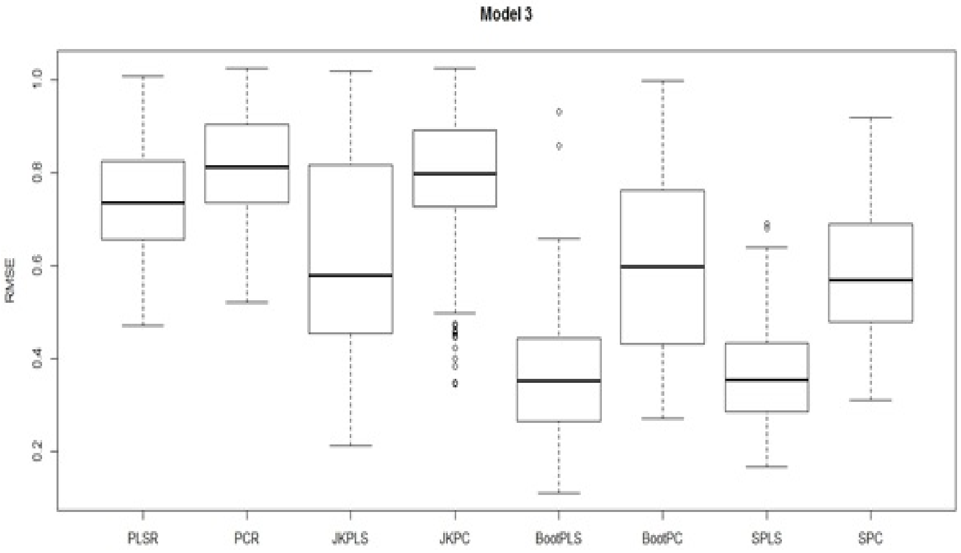

Figure 3.5: Comparison between the predictive ability of PLSR, PCR, JKPLS, JKPC, BootPLS, BootPC, SPLS, and SPC for Model 3.

considered VS methods for PCR.

The average value of true positive rate (TPR), the average value of false negative rate (FNR) and the average value of the reduced predictors (RP) value obtained by each VS method across the $G = 100$ independent generated data sets of the models 1, 2 and 3 are summarized in Table 3.1. Higher TPR and smaller FNR are desired, which shows that the VS methods perform well concerning the job of variable selection.

For dimension reduction, the VS methods on PLSR perform a considerable reduction in the number of predictor variables in all models. JKPLS, BootPLS, and SPLS, for all the models, have the highest average reduction in the number of predictors.

Since models 1 and 2 contain highly correlated variables, the average value of reduced predictors (RP) by the VS methods on PCR is higher for these models, compared with the RP obtained by using the VS methods on PCR for model 3. For model 3, which contains variables from each correlation block with partly large coefficients compared with models 1 and 2, the RP by the VS methods on PLSR is higher than the RP obtained by applying the VS methods on PCR.

For the accuracy of variable selection, for models 1 and 2, JKPC, SPLS, and SPC show the highest TPR but their FNR is too large compared with other methods. BootPC with a similar TPR as JKPC shows smaller FNR. For model 3, JKPC, BootPC, and SPC show the highest TPR while their FNR is too large compared with JKPLS, BootPLS, and SPLS (applying JK, sparse formulation based methods and Boot approaches on PLSR). This can be caused by the fact that, using the mentioned approaches on PCR cannot perform well in dimension reduction of variables compared with applying these approaches on PLSR, and therefore give higher TPR.

For all the models, Boot approach on both PCR and PLSR by performing a proper average reduction in the number of predictors can show a moderate average rate of truly selected variables while having the smallest (best) average rate of falsely selected variables compared with other VS methods.

Table 3.1: TPR, FNR and RP values for each variable selection method for models.

Results of the methods																		
	JKPLS			**JKPC**			**BootPLS**			**BootPC**			**SPLS**			**SPC**		
	TPR	FNR	RP	TPR	FNR	RP	TPR	FNR	RP	TPR	FNR	RP	TPR	FNR	RP	TPR	FNR	RP
Model 1	0.44	0.13	66	0.60	0.47	41	0.53	0.11	68	0.59	0.29	54	0.66	0.42	44	0.67	0.62	30
Model 2	0.45	0.15	83	0.55	0.41	58	0.52	0.12	86	0.55	0.25	73	0.62	0.41	58	0.69	0.66	38
Model 3	0.49	0.42	57	0.83	0.81	19	0.54	0.13	85	0.64	0.49	50	0.55	0.29	70	0.66	0.65	30

Table 3.2: $RMSE$, $RPred$ values and the number of selected predictors $(Pred)$ for each method on yarn data.

Method	**JKPLS**	**JKPC**	**BootPLS**	**BootPC**	**SPLS**	**SPC**
$RMSE$	0.0069	0.0081	0.0054	0.006	0.0178	0.0207
$RPred(Pred)$	183(85)	224(44)	195(73)	235(33)	84(184)	4(264)

3.7 Application to Real Data Set

In this section, we use the yarn real data set to further investigation the performance of all the aforementioned VS methods in terms of the $RMSE$ value and their ability of dimension reduction. The yarn data set which is available in R-package pls (Mevik and Wehrens (2007)) consists of 28 near-infrared spectra (NIR) of PET yarns, measured at 268 wavelengths, as predictors, and density as the response variable (density) (Swierenga *et al.* (1999)).

We standardize the data and fix the optimum number of components based on the lowest $RMSE$ value of cross-validation, and use the global minimum $RMSE$ value for the JK based on the $L = 7$ fold cross-validation and for the Boot-based on $B = 100$ bootstrap samples from the data as an initial value. For SPC, λ_1 was used such that the number of nonzero loadings varied in a rather wide range and, for SPLS, the tuning parameter was chosen by cross-validation (Zou *et al.* (2006), Chun and Keleș (2010), Chung *et al.* (2013)).

The optimal number of components obtained by applying PCR and PLSR on yarn data is equal to 7 and 6, respectively. We show the $RMSE$ values, the number of selected predictors $(Pred)$, and the number of reduced predictors $(RPred)$ obtained by each method in Table 3.2. The $RMSE$ value obtained by PCR and PLSR on *yarn* data set is 0.0351 and 0.0271, respectively. Table 3.2 shows that all of the VS methods on PCR and PLSR give a reduction in the $RMSE$ value. The JK and Boot approaches have a similar performance in reducing the $RMSE$ value, while Boot approach performs better in reducing the number of predictors. Both JK and Boot approaches on PCR and PLSR give a considerable reduction in both the $RMSE$ value and the number of predictors compared with SPC and SPLS. SPLS performs better than SPC by producing more reduction in the number of predictors. SPC does not show much reduction in the number of predictors. SPC for constructing the sparse loadings of at least one of the PCs (the first PC) uses 264 predictors, and therefore it can shrink only a few number of loadings to zero for yarn data.

3.8 Discussion and Conclusions

PCR and PLSR are two well-known multivariate regression methods in dealing with high-dimensional data. Although these methods can handle multicollinearity among predictors in data and reduce the data dimension, it is still necessary to eliminate the irrelevant predictors. In this thesis, we introduced a bootstrap based approach for performing variable selection for PCR and PLSR (i.e., BootPC and BootPLS).

We conducted a simulation study to compare the performance of BootPC and BootPLS with other variable selection methods for PCR and PLSR (JKPC, JKPLS, and sparse formulation based approaches for PCR and PLSR) in terms of predictive ability, dimension reduction and accuracy of variable selection. The analysis of the methods on the simulation data sets for all the presented models showed that using the variable selection methods on PCR and PLSR considerably reduced the $RMSE$ value. The VS methods on PLSR gave more reduction in the $RMSE$ value in all models. For the dimension reduction, the VS methods on PLSR in all models performed a considerable reduction in the number of variables.

VS methods on PCR in models with highly correlated variables reduced more predictors. In all models, Boot approach on both PCR and PLSR could perform favorably as well as other considered approached in terms of predictive ability (the $RMSE$ value).

For all the models, Boot approach on both PCR and PLSR by performing a proper average reduction in the number of variables showed a moderate average rate of truly selected variables while having the smallest (best) average rate of falsely selected variables among other considered VS methods.

By analyzing the methods on the real data set, we observed that the VS methods on PCR and PLSR reduced the $RMSE$ value. Boot and JK approaches showed more reduction in the $RMSE$ value and the number of predictors compared with sparse formulation based methods. Boot approach on PCR and PLSR performed favorably in reducing both the $RMSE$ value and the number of predictors compared with other considered approaches.

Chapter 4

Outlier Detection and Robust Variable Selection

4.1 Introduction

Many multivariate data sets contain atypical observations such as outliers, that is, data points that deviate from the usual assumptions and/or from the pattern suggested by the majority of the data. Outliers are more likely to appear in data sets with large number of observations and/or variables, and often they do not show up by simple visual investigation.

Classical regression methods are based on least squares fitting which can be strongly affected by outliers. When the data contains nasty outliers, typically two things happen

- the regression estimates differ substantially from the "right" answer, defined here as the estimates we would have obtained without the outliers (Hubert *et al.* (2008));
- the resulting fitted model does not allow to identify the outliers by means of different diagnostics methods.

The effect of the first consequence is fairly well-known although the size of the effect is often underestimated. But the second consequence is less well-known. Common intuition says that outliers must stick out from the classical fitted model, and indeed some of them may do so. But the most harmful types of outliers, especially if there are several of them, may affect the estimated model so much "in their direction" that they are now well-fitted by it (Hubert *et al.* (2008)).

Thus, by understanding the aforementioned effects the following two problems are equivalent

- robust regression: find a robust regression fit similar to the fit that we would have obtained without outliers;
- outlier detection: identify all the outliers that matter.

Another approach to dealing with outliers is robust regression, which tries to find the estimators that are resistant or at least not strongly influenced by the outliers. Thus, we can hope to find true outliers by studying the residuals of a robust regression. In this field many different ideas have been proposed, including least trimmed squares (LTS) (Rousseeuw (1984)), least median of squares (LMS) (Rousseeuw (1984)), M-estimators (Huber and Ronchetti (2009), Huber (1973)), GM-estimators or bounded-influence estimators (Krasker and Welsch (1982)), and S-estimators (Rousseeuw and Yohai (1984)).

Robust regression and outlier diagnostic methods end up being very similar. Both involve trying to find outliers and trying to estimate coefficients in a manner that is not overly influenced by outliers. The difference is in the order in which these two steps are performed. When using diagnostics we look for the outliers first and then, once they have been removed, we use LS on this "clean" data set for better estimates. Robust regression instead looks to find better robust estimates first and, given these estimates, we can discover the outliers by analyzing the residuals.

We aim to perform (robust) variable selection in linear regression methods, and the aforementioned ideas (Robust regression and outlier detection) have been incorporated in some of the current robust variable selection methods. Robust variable selection is variable selection as described in the previous chapters, except these methods are designed to work in the situation where the observations (in both $\boldsymbol{X}$ and $\mathbf{y}$) may contain outliers and/or the error distribution could exhibit non-Gaussian behavior.

A small proportion of outliers in the data may largely influence likelihood-type model selection methods such as AIC (Akaike (1970)), Mallows' C_p (Mallows (1973)). Under slight data contamination, variance inflation factor (VIF) criterion (Lin *et al.* (2011)) may lead to a completely different and improper selected model. One solution considered to this robust variable selection problem was to *robustify* the previously proposed variable selection criteria. One part of *robustifying* the previously suggested criteria is to use robust regression estimates. Hence, when there are contaminations in data, we need a robust variable selection method that is resistant to outliers in order to select variables reliably.

Recently, robust variable selection methods have received more attention in the literature. There are various robust variable selection approaches that are based on robustifying classical selection criteria, namely robust AIC (Ronchetti (1985)), robust C_p (Ronchetti and Staudte (1994)), robust selection criteria based on stratified bootstrap (Müller and Welsh (2005)), and robust final prediction error (Maronna *et al.* (2006)). Another robust approach, that is an added variable t-test in the context of regression based on the forward search procedure for variable selection, has been proposed by Atkinson and Riani (2002). McCann and Welsch (2007) proposed to add a dummy variable identity matrix to the design matrix for performing robust variable selection using elemental set sampling. Also, for generalized linear models, a robust selection criterion has been proposed (Cantoni and Ronchetti (2001)). Salibian-Barrera and Van Aelst (2008) used the fast and robust bootstrap to achieve a faster model selection method based on bootstrap, which makes it feasible to consider larger numbers of predictors. Yao and Wang (2013) imposed L_1 penalty in robust minimum average variance estimation (MAVE) (Čížek and Härdle (2006)) to achieve a robust variable selection method. They investigated their method only in the presence of outliers in the data.

Most of the robust model selection methods need to fit a large number of submodels robustly. When the number of predictors is large, then it is computationally more efficient to use variable selection methods that sequence the predictors according to their importance, such as forward selection (Weisberg (2014)).

The least angle regression (LARS) algorithm proposed by Efron *et al.* (2004) is a modified version of forward stagewise procedure. It is a powerful and computationally efficient procedure to sequence the predictor variables for least squares regression. LARS is based on the pairwise correlation between the predictors and the response variable, and therefore is not robust to the presence of a small amount of contamination in data.

Variance inflation factor (VIF) regression proposed by Lin *et al.* (2011) also inherits the spirit of a variation of forward stagewise regression. VIF regression selects those predictor variables among other available predictor variables that can reduce a statistically sufficient part of the variance in the predictive model. VIF regression approximates the partial correlation of each candidate variable with response variable by correcting the marginal correlations.

Khan *et al.* (2007) proposed robust LARS (RLARS) which replaces the means, variances and correlations of the variables inside LARS by (their robust counterparts) medians, median absolute deviations (MAD) and robust pairwise correlation estimates.

Dupuis and Victoria-Feser (2013) proposed robust version of VIF (RobVIF) regression, which robustly sequence the predictor variables. They used a robust weighted slope parameter to calculate the robust VIF selection criterion.

Our contribution in this thesis is to propose an algorithm which combines RLARS with least trimmed squares (LTS) regression (Rousseeuw (1984)), that is a highly robust regression method (Rousseeuw and Van Driessen (2006)), and then perform it on jack-knife (JK) subsets (Efron (1982)) to detect outliers. The merit of using JK subsets is to find the regression model with optimal predictive ability as measured by the MAD of the prediction errors obtained by cross-validation. This optimal regression model is devoted to do outlier detection in data. Then, the detected outliers are removed and standard LARS is performed on the clean data to obtain robust sequenced predictor variables in order of importance. Therefore, we obtain an outlier detection and robust variable selection method simultaneously.

The rest of this chapter is organized as follows. First, in Section 4.2 we will provide the methodology of two robust variable selection methods, robust LARS (RLARS), and robust variance inflation factor (RobVIF), as the two important counterparts of the proposed method. In Section 4.3, we will explain our strategy for outlier detection and robust variable selection as well as the proposed algorithm, i.e., jack-knife robust least angle regression (JKRLARS). In Section 4.4, we will conduct different simulation studies to evaluate and compare the performance of JKRLARS with LARS, RLARS and RobVIF. In Section 4.5, we will conduct a real data comparison. Finally, in Section 4.6, we will present the conclusions.

4.2 Robust Variable Selection Methods

In this chapter, we aim to select the most relevant variables as predictors to enter the linear regression model in data sets with large number of predictors containing contaminations such as outliers and/or leverage points. Thus, in this section we briefly review two robust variable selection methods as the two counterparts to the proposed method.

4.2.1 Robust Least Angle Regression

As mentioned in Section 2.4.1 of Chapter 2, the FS procedure enters variables in small steps into the regression model to prevent correlated predictors from being excluded from the top of the sequence. However, this method often becomes time-consuming due to the fact that often a large number of small steps are taken in the direction of the same variable. LARS solves this problem by analytically determining the optimal step size for each variable.
Khan *et al.* (2007) showed that the LARS resulting sequence of the predictors can be derived from the correlation matrix of the data (without the observations themselves) and proposed robust least angle regression (RLARS) by replacing the means, variances and correlations of the data with their robust counterparts. As robust measures for mean and variance they proposed to use computationally fast measures such as median and median absolute deviation (MAD), respectively. They introduced a fast robust pairwise correlation estimator based on bivariate winsorization (a generalization of univariate winsorization as introduced in Huber and Ronchetti (2009)).
Huber introduced the idea of one-dimensional winsorization of the data, and suggested that classical correlation coefficients be calculated from the transformed data (Huber (2011)). Alqallaf *et al.* (2002) examined this approach for the estimation of individual elements of a large-dimension correlation matrix. For n univariate observations $\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_n$, the transformation is given by $u_i = \psi_c((x_i - med(x_i))/MAD(x_i)),\ i = 1, 2, \ldots, n$, where the Huber score function $\psi_c(x)$ is defined as $\psi_c(x) = min\{max\{-c, x\},\ c\}$, with c a tuning constant chosen by the user. This one-dimensional winsorization approach is very fast to compute but unfortunately it does not take into account the orientation of the bivariate data. It only brings the outlying observations to the boundary of a $2c \times 2c$ square, as shown in Figure 4.1. It can be clearly seen in this plot that the univariate approach does not resolve the effect of the obvious outliers at the bottom right which are shrunken to the corner $(2, -2)$, and thus are left almost

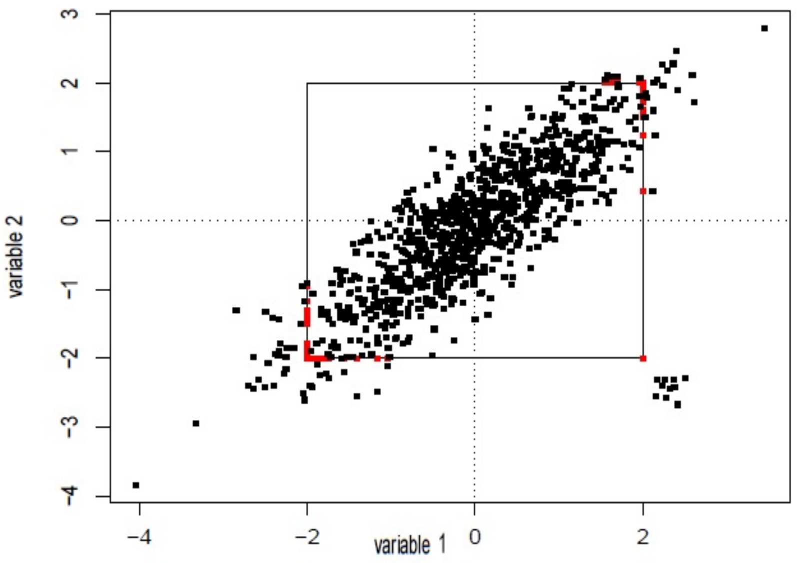

Figure 4.1: Limitation of separate univariate winsorizations ($c = 2$). The correlation outliers (red points) are only shrunken to the boundary of the square(Khan *et al.* (2007)).

unchanged. Bivariate winsorization, which is based on an initial tolerance ellipse for the majority of the data, can remedy this problem. After robustly standardizing the data, bivariate winsorization is obtained on the basis of an initial robust bivariate correlation matrix $\boldsymbol{R}_0$ and a corresponding tolerance ellipse. For instance, consider the Mahalonobis distance $D(\boldsymbol{x})$, with $\boldsymbol{x} = (x_1, x_2)^T \in \mathbb{R}^2$ based on initial bivariate correlation matrix $\boldsymbol{R}_0$ and set the tuning constant equal to the 95% quantile of the χ^2_2 distribution which is $c = 5.99$. The outliers by using the bivariate transformation $\boldsymbol{u} = \min(\sqrt{c/D(\boldsymbol{x})}, 1)\boldsymbol{x}$ with $\boldsymbol{x} = (x_1, x_2)^T$ are shrunken to the boundary of the 95% tolerance ellipse and therefore the resulting correlation estimate will be less affected by the outliers. Thus, a more robust correlation estimate is given.

An essential part of the bivariate winsorization procedure is choosing a proper initial correlation matrix $\boldsymbol{R}_0$. Khan *et al.* (2007) proposed *adjusted winsorization* as initial estimator.

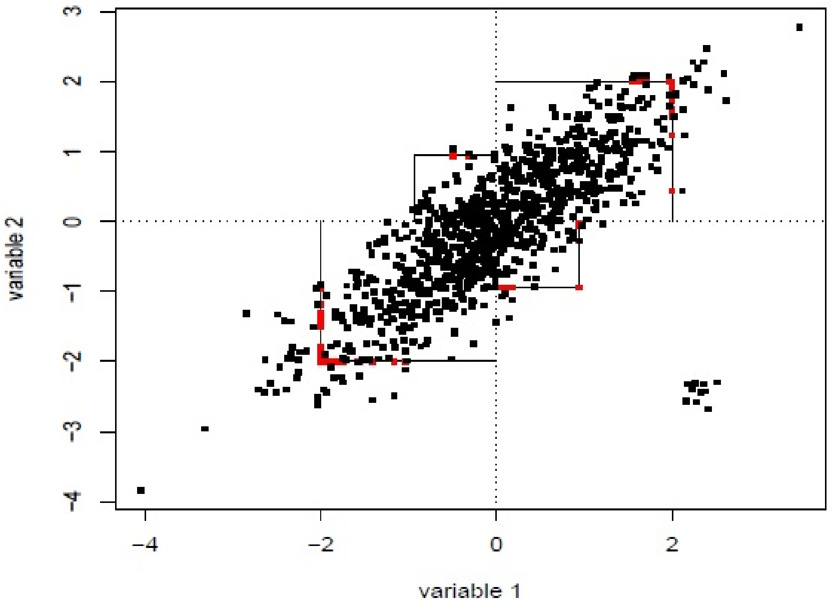

Figure 4.2: Adjusted winsorization (for initial estimate $\boldsymbol{R}_0$) with $c_1 = 2,\ c_2 = \sqrt{h}c_1$ (Khan *et al.* (2007)).

Adjusted winsorization uses univariate winsorization with different tuning constants for different quadrants: a tuning constant c_1 for the two quadrants that contain the majority of the standardized data, and a smaller tuning constant c_2 for the other two quadrants. For instance, $c_1 = 2$, or $c_1 = 2.5$, and $c_2 = \sqrt{h}c_1$ where $h = n_2/n_1$, with n_1 the number of observations in the major quadrants and $n_2 = n - n_1$ (Khan *et al.* (2007) used $c_1 = 2$). The initial correlation matrix $\boldsymbol{R}_0$ is obtained by computing the classical correlation matrix of the adjusted winsorized data. It can be seen from Figure 4.2 that the adjusted winsorization shrinks bivariate outliers to the boundary of the smaller square.

Khan *et al.* (2007) also introduced bootstrap robust least angle regression (BRLARS) by sequencing the predictor variables on bootstrap samples to obtain more stable sequences which have also been used in random forests (Breiman (2001)) (see for example Hastie *et al.* (2009)). They generate B bootstrap samples from the original data and apply RLARS on each of these samples to sequence the predictors. Then, they ranked the predictor variables according to their average rank over the bootstrap samples. The relevant predictors with the smallest average rank construct the reduced set.

4.2.2 Robust Variance Inflation Factor (RobVIF) Regression

Dupuis and Victoria-Feser (2013) proposed the robust VIF (RobVIF) regression which is a robust version of VIF regression proposed by Lin *et al.* (2011). They used a robust weighted slope parameter to calculate the robust VIF selection criterion. Let $\mathbf{X}_S$ be the design matrix that contains the selected variables at stage S, and $\tilde{\mathbf{X}}_S = [\mathbf{X}_S \quad \mathbf{z}_j]$ with $\mathbf{z}_j$ the new potential candidate predictor to be evaluated for inclusion. Consider the following models

$$\mathbf{y} = \mathbf{X}_S\boldsymbol{\beta}_S + \mathbf{z}_j\beta_j + \mathbf{e}_{step}, \quad \mathbf{e}_{step} \sim N(\mathbf{0}, \sigma^2_{step}\mathbf{I}), \tag{4.1}$$

$$\mathbf{r}_S = \mathbf{z}_j\gamma_j + \mathbf{e}_{stage}, \quad \mathbf{e}_{stage} \sim N(\mathbf{0}, \sigma^2_{stage}\mathbf{I}), \tag{4.2}$$

where $\mathbf{r}_S$ are the residuals of the projection of $\mathbf{y}$ on $\mathbf{X}_S$, $\boldsymbol{\beta}_S$ and β_j are slope parameters, γ_j is the slope parameter of the fit of $\mathbf{z}_j$ on the residuals $\mathbf{r}_S$, $\mathbf{e}_{step}$ and $\mathbf{e}_{stage}$ are the errors. Lin *et al.* (2011) showed that when least squares are used to estimate, $\hat{\gamma} = \rho^2\hat{\beta}_j$ where $\rho^2 = \mathbf{z}_j^T(\mathbf{I} - \mathbf{X}_S(\mathbf{X}_S^T\mathbf{X}_S)^{-1}\mathbf{X}_S^T)\mathbf{z}_j$ (see Section 2.4.2 in Chapter 2). Dupuis and Victoria-Feser (2013) calculated the robust weighted slope estimators $\hat{\beta}_j^w$ using Tukey's redescending biweight weights (Huber and Ronchetti (2009)), and then they computed an approximate robust test statistic in order to compare it with an adapted quantile to decide whether or not to add $\mathbf{z}_j$ to the current set of predictors. Let $\mathbf{X}_S^w = diag(\sqrt{w_{iS}^0})\mathbf{X}_S$ be the weighted design matrix with v columns ($v-1$ predictors) at stage S, $\mathbf{y}^w = diag(\sqrt{w_{iS}^0})\mathbf{y}$ the weighted response variable at stage

S, and $\mathbf{z}_j^w = diag(\sqrt{w_{ij}})\mathbf{z}_j$ the new candidate predictor to be considered to enter the current set at stage $S+1$ (see details in Dupuis and Victoria-Feser (2013) for calculation of weights w_{iS}^0 and w_{ij}). Then $\hat{\beta}_j^w$, the robust weighted estimator of β_j, in (4.1) is obtained. Let

$$\rho^w = (\mathbf{z}_j^{wT}\mathbf{z}_j^w)^{-1}(\mathbf{z}_j^{wT}\mathbf{z}_j^w - \mathbf{z}_j^{wT}\mathbf{H}_S^w\mathbf{z}_j^w),$$

with $\mathbf{H}_S^w = \mathbf{X}_S^w(\mathbf{X}_S^{wT}\mathbf{X}_S^w)^{-1}\mathbf{X}_S^{wT}$, then

$$\hat{\beta}_j^w = (\rho^w)^{-1}\hat{\gamma}_j^w$$

with $\hat{\gamma}_j^w = (\mathbf{z}_j^{wT}\mathbf{z}_j^w)^{-1}\mathbf{z}_j^{wT}\mathbf{r}_S^w$ the weighted estimator of the fit of $\mathbf{z}_j^w$ on the weighted residuals $\mathbf{r}_S^w$ (4.2). Since computing ρ using all the data is computationally expensive, Dupuis and Victoria-Feser (2013) used a subsampling approach followed by Lin *et al.* (2011) to estimate ρ^w on a randomly chosen subsample of size g. Then the approximate robust test statistic T_w based on $\hat{\gamma}_j^w$ by comparing the expected value of the estimated variance of $\hat{\beta}_j^w$ and $\hat{\gamma}_j^w$ is given by

$$T_w = (\rho^w)^{-1/2}\frac{\hat{\gamma}_j^w}{\sqrt{\hat{\sigma}^2/n(1/n\sum_i z_{ij}^{w2})^{-1}e_c^{-1}}}, \tag{4.3}$$

with $\hat{\sigma}^2$ a robust mean squared error for the model with $\mathbf{r}_S^w$ as response and $\mathbf{z}_j^w$ as predictor variable (that is (4.2)), where z_{ij}^w denotes the element of $\mathbf{z}_j^w$. Then T_w is compared with an adapted quantile as a decision rule to decide whether or not to add the new predictor variable (to calculate e_c and for more details see Dupuis and Victoria-Feser (2013)).

4.3 Jack-knife Robust LARS (JKRLARS) Algorithm

As mentioned before, since LARS is based on the pairwise correlation between the predictors and the response variable, therefore it is not robust to the presence of a small amount of contamination in data. We propose a method for obtaining the robust sequenced predictor variables for LARS. In this section we explain our strategy of outlier detection and robust variable selection. We seek to detect outliers in the data and at the same time specify a robust sequence of the predictor variables in order of their importance. Hence, we propose an approach that can perform outlier detection and robust variable selection simultaneously. To generate jack-knife subsets, first the data is randomly partitioned into L non-overlapping equally-sized subsets. Then, each subset is retained once and RLARS is applied to the remaining subsets to find L sequence predictor variables in their importance. The most appropriate predictor variables relevant to these L RLARS sequences are used to fit a robust regression model. We use the least trimmed squares (LTS) regression which is a highly robust and computationally efficient robust regression method (Rousseeuw and Van Driessen (2006)). Denote the vector of squared residuals by $\mathbf{r}^2(\boldsymbol{\beta}) = (r_1^2, \ldots, r_n^2)^T$ with $r_i^2 = (y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2, i = 1, \ldots, n$. Then the LTS estimator is defined as

$$\hat{\boldsymbol{\beta}}_{LTS} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{h} (\mathbf{r}^2(\boldsymbol{\beta}))_{i:n}, \tag{4.4}$$

where $(\mathbf{r}^2(\boldsymbol{\beta}))_{1:n} \leq \cdots \leq (\mathbf{r}^2(\boldsymbol{\beta}))_{n:n}$ are the order statistics of the squared residuals and $h \leq n$. For $h = [(n+p+1)/2]$, where p is the number of predictors (here, $[a]$ denotes the integer part of a) the LTS breakdown point equals 50% whereas for greater h its breakdown point is $(n-h)/n$. The usual choice $h \approx 0.75n$ yields the LTS breakdown point of 25% (Hubert *et al.* (2008)). Hence, LTS regression seeks to find the subset of h observations whose least squares fit gives the smallest sum of squared residuals. Based on the LTS regression model, predicted values of the observations are calculated by cross-validation. The median absolute deviation (MAD) values of the prediction errors for each LTS regression model are calculated. Then, the optimal LTS regression model which yields to minimum MAD value is selected. To identify outliers, the standardized prediction errors of this optimal model are computed. The detected outliers are left out, and LARS is applied to the clean data to find an improved sequence of predictor variables. The goals of identifying outliers in data and robustly sequencing predictor variables simultaneously reflect the specifications of the proposed approach.

Consider a data set $(y_i, \mathbf{x}_i), i = 1, \dots, n$, where $y_i \in \mathbb{R}$, and $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T \in \mathbb{R}^p$, are the response and predictor values respectively. Let $J = \{1, \dots, p\}$ and $I = \{1, \dots, n\}$ be the set of indices for the candidate predictors and observations, respectively, and $q \ll p$ the length of the most relevant predictor variables returned by RLARS. Then the JKRLARS algorithm proceeds as follows:

Step 1. The observations $(y_i, \mathbf{x}_i), i = 1, \dots, n$ are partitioned into L randomized non-overlapping equally-sized JK subsets $I_f \subset I = \{1, \dots, n\}$, with $f = 1, \dots, L$. Clearly, I_f contains the indices of the observations in f-th subset, with $|I_f| \approx \frac{n}{L}$;

Step 2. With the f-th subset left out, RLARS is applied to the set $(y_i, \mathbf{x}_i), i \notin I_f$, of $L - 1$ subsets to find a sequence of predictor variables $(\boldsymbol{x}_j{}^{(f)})_{j \in J_f}$ with $J_f \subset J = \{1, \dots, p\}$ such that $|J_f| \ll p$, where p is the number of predictor variables. Clearly, J_f contains the indices of the $|J_f| = q$ most relevant predictor variables returned by RLARS;

Step 3. The predictors $\boldsymbol{x}_j^{(f)}, j \in J_f$, are used as predictor variables for LTS regression. We thus consider the regression model

$$y_i = \mathbf{x}_i^{(f)^T} \boldsymbol{\beta}^{(f)} + e_i, \tag{4.5}$$

where $\mathbf{x}_i^{(f)}$ denotes the i-th observation of the predictors $\boldsymbol{x}_j^{(f)}, j \in J_f$ for the i-th observation and $\boldsymbol{\beta}^{(f)}$ is estimated by using LTS;

Step 4. To evaluate the prediction performance of each LTS regression fit, we perform L-fold cross-validation. In order to detect outliers, we compute the predicted values $\hat{y}_i$ which are defined as

$$\hat{y}_i = \mathbf{x}_i^{(f)^T} \hat{\boldsymbol{\beta}}^{(f)}, i \in I_f, \tag{4.6}$$

where $\hat{\boldsymbol{\beta}}^{(f)}$ denotes the LTS estimates of the regression coefficients with the f-th subset left out as obtained in the previous step. Thus, predicted values for all the observations are obtained. For each LTS regression the prediction errors $\text{PE}_i = y_i - \hat{y}_i$, and the corresponding MAD value of the prediction errors as a measure of the model's predictive ability are calculated. The minimum MAD value over all LTS models is obtained to find the LTS model with optimal predictive ability. Then, the corresponding standardized prediction errors of this optimal model are used to detect outliers. The standardized prediction errors are defined by $\frac{\text{PE}_i}{\hat{\sigma}}, i = 1, \ldots, n$ where $\hat{\sigma} = c_{h,n}\sqrt{\frac{1}{h}\sum_{i=1}^{h}(\mathbf{r}^2(\boldsymbol{\beta}))_{i:n}}$, and $c_{h,n}$ makes $\hat{\sigma}$ consistent and unbiased at Gaussian error distribution (Pison *et al.* (2002)). It should be mentioned that the LTS scale estimate $\hat{\sigma}$ is itself highly robust and can therefore be used to identify outliers by $\frac{\text{PE}_i}{\hat{\sigma}}$. As in Hubert *et al.* (2008) we define the set of the indices of outlying observations as $I_{Out} = \left\{ i \in I : |\frac{\text{PE}_i}{\hat{\sigma}}| > \sqrt{\chi^2_{1,0.975}} \right\}$;

Step 5. The detected outliers $(y_i, \mathbf{x}_i), i \in I_{Out}$, are removed (or given weight zero) and LARS is applied to the clean data $(y_i, \mathbf{x}_i), i \in I^c_{Out}$, with I^c_{Out} the complement of I_{Out}. The predictors $\boldsymbol{x}_j, j \in J_{final}$ with $J_{final} \subset J$ from the candidate predictors are obtained as the robust version of LARS sequenced predictor variables.

4.4 Simulation Study

We conduct a simulation study to investigate and compare the performance of JKRLARS with its counterparts. We perform all the methods in R Core Team (2014). We use package *lars* (Hastie and Efron (2013)) to perform LARS and we perform RLARS (Khan *et al.* (2007)) and RobVIF (Dupuis and Victoria-Feser (2013)) using the codes available on the authors' website. We consider $h \approx 0.75n$ in order to guarantee a breakdown point of 25% for LTS inside JKRLARS. The number of $L = 10$ subsets is considered for the JKRLARS algorithm. For RobVIF, we consider the same subsample size of 200 with the same initial values for wealth and pay-out equal to 0.5 and 0.05, respectively similar to Dupuis and Victoria-Feser (2013).

We consider a simulation setting similar to Khan *et al.* (2007), which is based on the design of Frank and Friedman (1993).

The linear model is created as

$$\mathbf{y} = \mathbf{l}_1 + ... + \mathbf{l}_k + \sigma\mathbf{e}, \tag{4.7}$$

with $k = 6$ latent independent standard normal variables $\mathbf{l}_1, \mathbf{l}_2, \ldots, \mathbf{l}_k$ and an independent standard normal variable $\mathbf{e}$. The value of σ is chosen such that the signal to noise ratio is equal to 3, that is, $\sigma = \sqrt{k}/3$. Let $\mathbf{e}_1, ..., \mathbf{e}_p$ be independent standard normal variables, then the set of p predictors is created as

$$\begin{gathered}
\boldsymbol{x}_j = \mathbf{l}_j + \tau\mathbf{e}_j, \qquad j = 1, ..., k \\
\boldsymbol{x}_{k+1} = \mathbf{l}_1 + \delta\mathbf{e}_{k+1}, \\
\boldsymbol{x}_{k+2} = \mathbf{l}_1 + \delta\mathbf{e}_{k+2}, \\
\boldsymbol{x}_{k+3} = \mathbf{l}_2 + \delta\mathbf{e}_{k+3}, \\
\boldsymbol{x}_{k+4} = \mathbf{l}_2 + \delta\mathbf{e}_{k+4}, \\
. \\
. \\
. \\
\boldsymbol{x}_{3k-1} = \mathbf{l}_k + \delta\mathbf{e}_{3k-1}, \\
\boldsymbol{x}_{3k} = \mathbf{l}_k + \delta\mathbf{e}_{3k}, \\
\text{and } \boldsymbol{x}_j = \mathbf{e}_j, \qquad j = 3k + 1, ..., p,
\end{gathered}$$

with $\delta = 5$ and $\tau = 0.4$ so that target predictor variables $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ are formed by low noise perturbations of the latent variables. Variables $\boldsymbol{x}_{k+1}, \ldots, \boldsymbol{x}_{3k}$ are noise predictor variables that are correlated with the latent variables, and variables $\boldsymbol{x}_{3k+1}, \ldots, \boldsymbol{x}_p$ are independent noise predictor variables.

We consider five different sampling distributions,

(a) $e \sim (1-a)N(0,1) + aN(0,1)/U(0,1)$, symmetric, slash contamination;

(b) $e \sim$ Cauchy$(0,1)$, heavy-tailed cauchy contamination;

(c) $e \sim (1-a)N(0,1) + aN(20,1)$, asymmetric, shifted normal contamination;

(d) same as (a), with high leverage X values, $X \sim N(50,1)$;

(e) same as (b), with high leverage X values, $X \sim N(50,1)$;

where $a = 0.1$ denotes the fraction of contamination in the data.

We generate 200 independent data sets of size $n = 150$ with $p = 50$ predictors from the above five simulation scenarios and each time we perform all the aforementioned methods on the same data set.

4.4.1 Performance measurements

We evaluate the performance of the JKRLARS concerning outlier detection by the true positive rate (TPR) and false negative rate (FNR). A true positive is an observation that is contaminated in the data and is also detected as outlier. Analogously, a false negative is an observation that is contaminated in the data, but is labeled as a regular observation. Denote the set of the indices of the regular observations in the data by $I_R \subset I = \{1, \ldots, n\}$ and the set of the indices of the contaminated observations in the data by $I_R^c \subset I = \{1, \ldots, n\}$. The TPR and FNR can then be defined as

$$TPR = \frac{|\{i : i \in I_{Out} \wedge i \in I_R^c\}|}{|I_R^c|}; \tag{4.8}$$

$$FNR = \frac{|\{i : i \in I_{Out}^c \wedge i \in I_R^c\}|}{|I_R|}. \tag{4.9}$$

Higher TPR and smaller FNR are desired, thus showing that JKRLARS performs well concerning outlier detection.

In order to compare the performance of JKRLARS with its counterparts, we plot the recall curves, that is, to plot the average number of target variables t_m (target variables refer to $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ in Section 4.4) included in the first m sequenced predictor variables entering the model over the 200 independent data sets as a function of varying m. With k number of target variables, good performance is achieved when the method can find the k target variables in the first t_k sequenced predictor variables, with t_k equal or close to k.

4.4.2 Simulation Study Results

In this section we present and discuss the results of the five presented simulation scenarios. We perform LARS, RLARS, RobVIF and JKRLARS on each of the 200 independent data sets.

First, we investigate how JKRLARS performs to detect outliers in the data. Table 4.1 shows the results for TPR and FNR averaged over the 200 data sets for the 5 types of considered contamination. From Table 4.1 we can see that the TPR and FNR of the outlier detection procedure is almost perfect in scenarios (c), (d), and (e). High leverage points and clear outliers can thus be detected with high accuracy. At a first look TPR and FNR for scenarios (a) and (b) seem much worse, but we should bear in mind that in these cases the observations are contaminated by producing errors from the long-tailed slash and Cauchy distributions, respectively. Not all errors produced from these distributions will lie in the tails of the distribution. Thus, in these scenarios not all of the 10% fraction of contaminated observations will be actual outliers. Considering the cut-off value $\sqrt{\chi^2_{1,0.975}}$ for identifying outliers, it can easily be checked that only 35.2% of the generated slash errors and 26.7% of the generated Cauchy errors are expected to produce outlying observations. Comparing the TPR with these fractions, we see that the outlier detection procedure still performs reasonably well in these difficult scenarios. Therefore, JKRLARS does a good job of outlier detection considering both TPR and FNR in all scenarios.

We performed LARS, RLARS, RobVIF and JKRLARS on all 200 data sets in each scenario to select the predictor variables in order of their importance. Figure 1 displays the recall curves for comparing LARS, RLARS, RobVIF and JKRLARS in terms of sequencing the predictor variables in each simulation scenario. For each sequence of predictor variables we determine the number t_m of target variables included in the first m sequenced variables entering the model with m ranging from 1 to 25.

In scenarios a, b and c, JKRLARS shows the same excellent performance as RLARS and RobVIF in sequencing the $k = 6$ target variables at the top of the sequence (Figures 4.3 (a)-(c)). In the high leverage scenarios (d) and (e), RobVIF fails in selecting the target variables. In the slash with high leverage scenario RobVIF can select none of the target variables (Figure 4.3 (d)), and in the Cauchy with high leverage scenario it can select only one of the target variables (Figure 4.3 (e)). Figures 4.3 (d) and (e) show that in the high leverage scenarios RLARS has much more problems in picking up the target variables in the beginning, while JKRLARS succeeds much better in picking up most of the relevant variables in the beginning of the sequence. In particular, in models with (less than) 10 predictors JKRLARS captures 5 of the 6 relevant variables, while RLARS needs models with up to 15 predictors to include at least 5 of the 6 relevant variables.

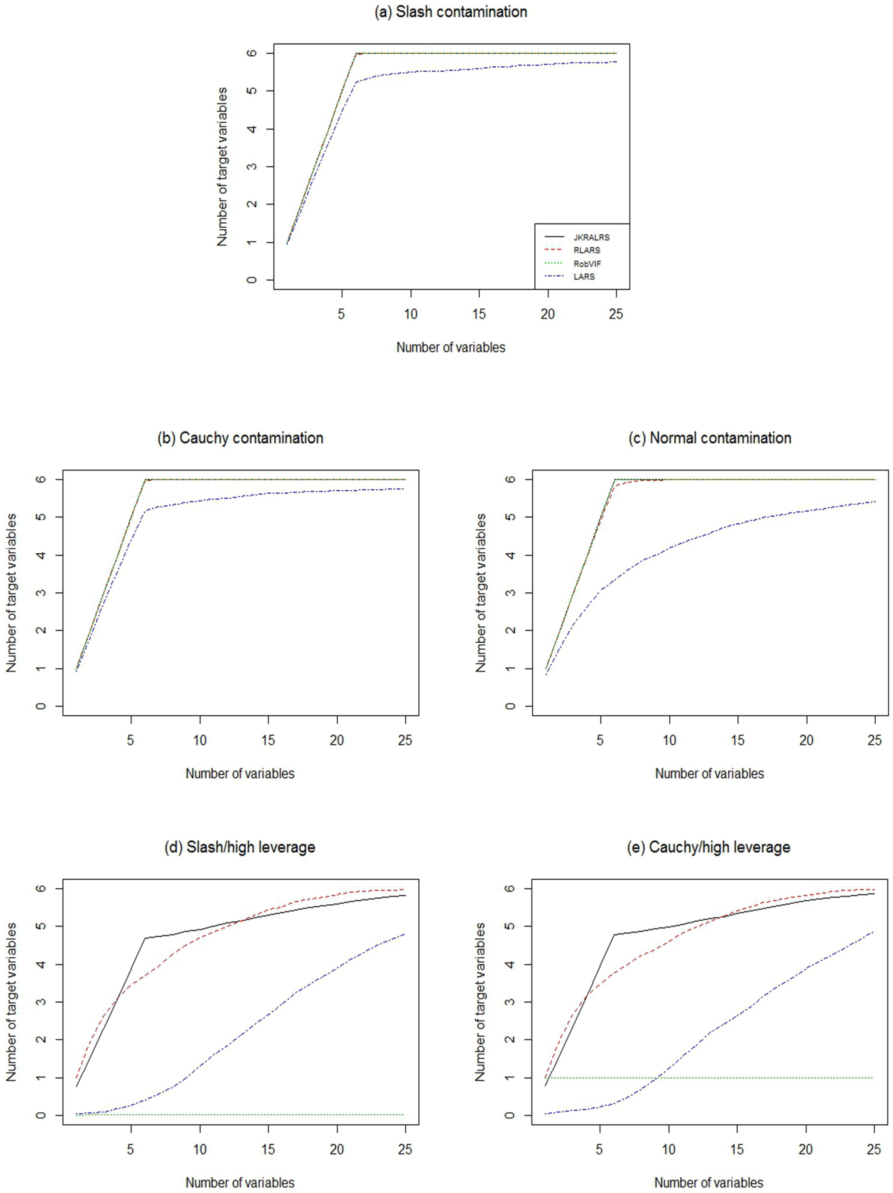

Figure 4.3: Average number of target variables t_m versus m for LARS, RLARS, RobVIF, and JKRLARS for scenarios (a)-(e). The lines shown in all plots follow the legend of figure (a).

Table 4.1: The true positive rate (TPR) and the false negative rate (FNR) averaged over 200 simulation runs, are reported for JKRLARS.

Case	**a**	**b**	**c**	**d**	**e**
TPR	0.25	0.20	1	0.96	0.97
FNR	0.08	0.08	0	0.004	0.003

4.5 Application to Real Data

We evaluate and compare the performance of LARS, RLARS, BRLARS, RobVIF and JKRLARS on the 1990 US Census data. In order to investigate the selected variables obtained by each method, we measure their median absolute prediction error (MAPE) of the optimal number of selected variables as measured by 10 test subsets, i.e., the data is partitioned into 10 roughly equally-sized subsets and the MAPE of the methods is calculated on each test subset. We repeat this process 10 times and each time we use the same test subset for all the methods. The optimal number of variables for each method is obtained by considering its learning curve. For this purpose, we plot the number of variables in the model versus the MAPE as measured by 10 test subsets. The optimal number of variables can be chosen as the point where the learning curve does not show a considerable slope anymore.

The original Census data contains 22784 observations and 139 variables and can be downloaded at http://www.cs.toronto.edu/~delve/data/census-house/desc.html. More details on how the data was obtained can be found in Dupuis and Victoria-Feser (2011). After removing the collinear predictors, and those are all zeros or almost zeros, the data contains 51 predictors and the response variable, that is the average price asked for the housing unit. As mentioned in the Section 4.4, in order to guarantee a breakdown point of 25% for LTS inside JKRLARS we consider $h \approx 0.75n$.

The number of $L = 10$ subsets and 10 bootstrapped samples is considered for the JKRLARS and BRLARS methods, respectively. For RobVIF, we consider the same subsample size of 200 with the same initial value for wealth and pay-out equal to 0.5 and 0.05, respectively similar to Dupuis and Victoria-Feser (2013). As it can be seen from the learning curves in Figure 4.4, the optimal number of selected variables for LARS, RLARS, BRLARS, RobVIF, and JKRLARS is 8, 8, 8, 9, and 8, respectively. Figure 4.5 shows the box plots of the MAPE values of the optimal number of selected variables over 10 test subsets with 10 times replication for each method. Box plots in Figure 4.5 suggest that JKRLARS outperforms its robust counterparts while giving a better performance of LARS in the presence of outliers in data.

4.6 Discussion and Conclusions

Since outlier detection and variable selection in the presence of contaminations such as outliers and/or leverage points in the data are inseparable problems, therefore we need a robust method with the ability of outlier detection and variable selection, simultaneously. We proposed an outlier detection and robust variable selection method by combining RLARS with LTS regression as a highly robust regression method on jack-knife subsets. The merit of using these subsets is to find the regression model with optimal predictive ability as measured by the MAD of the prediction errors obtained by cross-validation. This optimal regression model is devoted to the job of outlier detection. Then, removing the detected outliers standard LARS is performed on the clean data to obtain robust sequenced predictor variables in order of their importance.

The results of performing this method on contaminated simulation data sets showed that JKRLARS does a good job of outlier detection. Also, concerning robust variable selection, it performs well in sequencing the predictor variables robustly for the different data configurations containing outliers and leverage points. Thus, a robust

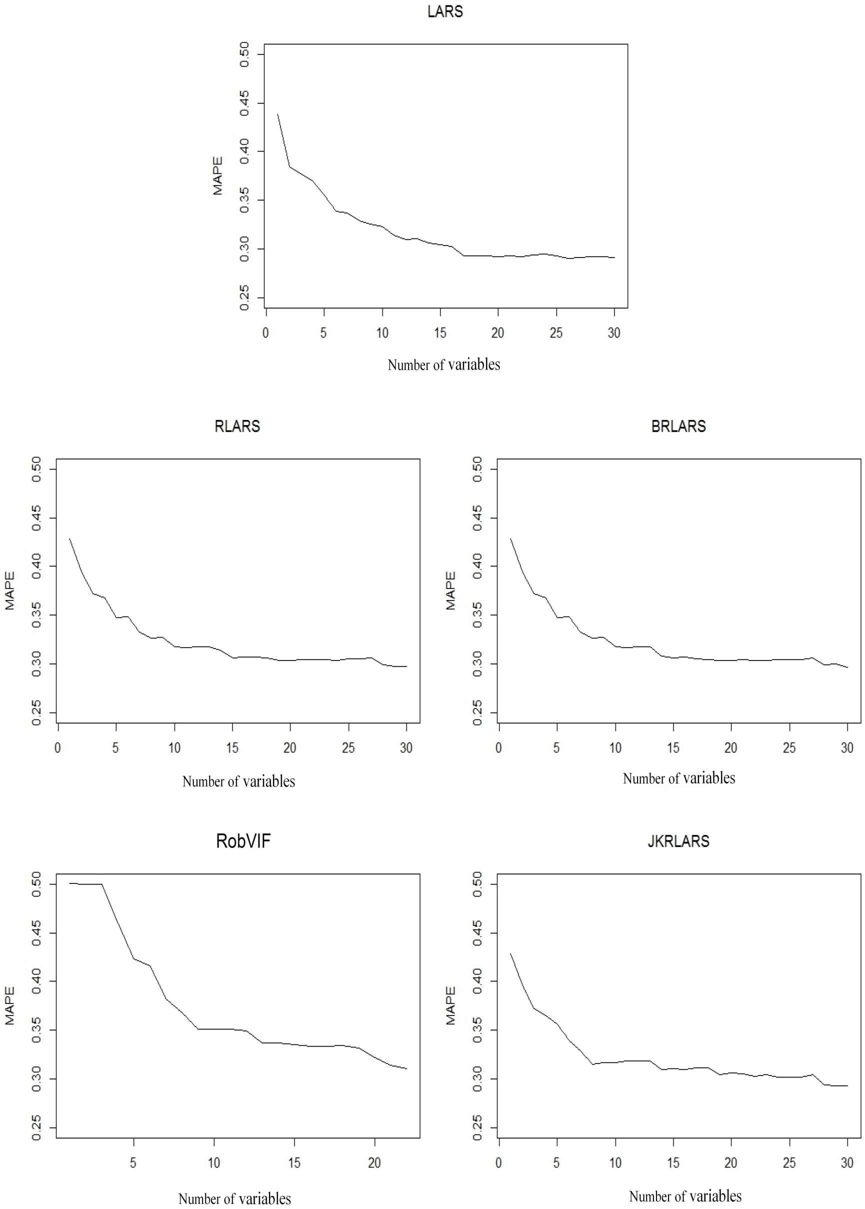

Figure 4.4: Learning curves for LARS, RLARS, BRLARS, RobVIF, and JKRLARS on Census data. Each learning curve suggests the optimal number of predictors for each method.

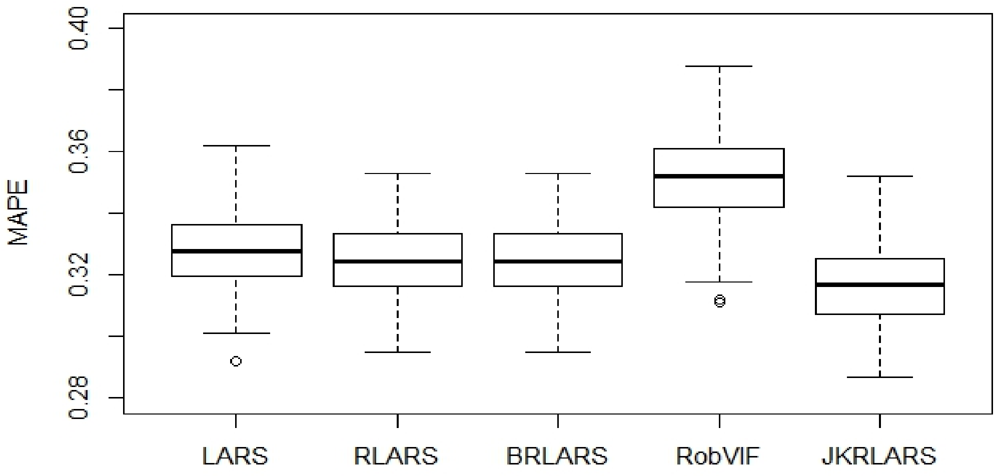

Figure 4.5: Box plots of MAPE values of the optimal number of selected variables over 10 test subsets with 10 times replication for each method.

version of LARS sequenced predictor variables is yielded by JKRLARS. JKRLARS not only performs as good as its counterparts, which are RLARS and RobVIF, in robustly sequencing the predictor variables in the presented simulation scenarios containing outliers, but also outperforms RLARS in situations with high leverage points (RobVIF fails in robustly sequencing the predictor variables in these situations).

Finally, through real data set, we confirm that JKRLARS outperforms LARS and the other considered robust variable selection methods (RLARS, BRLARS, RobVIF).

Chapter 5

Conclusions and Future Work

In this chapter, the main ideas of this thesis and the main results obtained are summarized. Further developments are also addressed.

5.1 Conclusions

The main subject of this thesis focuses on the development of regression methods for performing variable selection and their application to data sets with a large number of predictors. Data sets with a large number of predictors usually contain outliers, and linear regression methods are highly sensitive to the existence of outliers in the data. Outlier detection and variable selection when the data is contaminated with outliers are not two separable problems. Therefore, according to the main theme of this thesis the focus is also on the development of regression methods to robustly select variables in the presence of outliers in data.

In Chapter 3, we have addressed that PLSR builds the components by modeling the relations between predictors and response variable in contrast to PCR, which creates the components by modeling the linear combinations of the predictors that explain most of their variation. Therefore, PLSR is more useful for prediction rather than explanation with less number of components in comparison with PCR. We introduced a procedure (i.e., BootPLSR and BootPCR) for selecting the significant predictors in the PLSR and PCR models. This procedure is a bootstrap-based technique for significance tests of regression coefficients to select the significant predictors. We applied the proposed method to simulated data and compared its performance with jack-knife based technique and sparse formulation based methods for PLSR and PCR. Although none of the methods could perform better than each other throughout simulation studies, the proposed method could perform favorably as well as other considered methods in terms of predictive ability. For all the simulation models, the proposed method showed a moderate average rate of truly selected variables, while having the smallest (best) average rate of falsely selected variables compared with other considered variable selection methods (i.e., JKPC, JKPLS, SPC, and SPLS). We also applied the aforementioned methods to real data and compared their performance concerning predictive ability ($RMSE$ value) and dimension reduction of the variables. The proposed method performed favorably in reducing the $RMSE$ value (in the other words, showed a better predictive ability) , as well as the number of predictors compared with other considered methods.

In Chapter 4 we mentioned that LARS is a time-efficient procedure to sequence predictors in order of their importance and that it is not resistant to the presence of outliers in data. Therefore, LARS sequenced predictors can be easily affected by a small fraction of outliers in data. A robust version of LARS has already been introduced by replacing the means, variances and correlations of variables inside LARS with their robust counterparts.

We introduced a procedure for performing outlier detection and robust variable selection by combining RLARS with LTS regression as a highly robust regression method on jack-knife subsets. By using these subsets, the regression model with optimal predictive ability, as measured by the MAD of the prediction errors obtained by cross-validation, is found. This optimal regression model is devoted to detect outliers. Then, after removing the detected outliers LARS is performed on the clean data to obtain robust sequenced predictor variables in order of their importance.

The results of applying the proposed method to simulated data showed that it performed favorably in outlier detection and robust variable selection. In simulation studies with outliers, JKRLARS could perform the job of robust variable selection as well as its counterparts (i.e., RLARS and RobVIF). Also, it could outperform RLARS in other simulation studies containing both outliers and leverage points, while RobVIF failed to robustly sequence the predictor variables in these situations.

We also evaluated and compared the performance of the proposed method with LARS and the aforementioned robust variable selection methods on real data. Concerning robust variable selection, the proposed method showed a better performance, as measured by the MAPE of the optimal number of selected variables achieved by each method on test subsets, compared to LARS and other considered robust variable selection methods.

5.2 Future Developments

There are some possibilities that could be extended from this thesis.

In this work we mainly focused on variable selection and outlier detection in data sets with a large number of predictors. We are also thinking about the possibility of developing the discussed algorithm to enable principal components regression (PCR) and partial least squares regression (PLSR) to perform variable selection in generalized linear models in order to consider data from different distributions with a large number of predictors.

Given that least absolute shrinkage and selection operator (LASSO) estimator can be computed using the least angle regression (LARS) algorithm, the discussed algorithm may be extended to obtain a robust version of LASSO in order to detect outliers and robustly select variables in data sets with a large number of predictors.

In some circumstances, the data is measured in space and time, and therefore it is important to develop spatial/temporal methods to select variables in both spatial and temporal dimensions for high-dimensional data by combining statistical regression techniques.

Bibliography

Akaike, H. (1970). Statistical predictor identification, *Annals of the Institute of Statistical Mathematics* **22**(1): 203–217.

Akaike, H. (1974). A new look at the statistical model identification, *Automatic Control, IEEE Transactions on* **19**(6): 716–723.

Alqallaf, F. A., Konis, K. P., Martin, R. D. and Zamar, R. H. (2002). Scalable robust covariance and correlation estimates for data mining, *Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining*, ACM, pp. 14–23.

Atkinson, A. C. and Riani, M. (2002). Forward search added-variable t-tests and the effect of masked outliers on model selection, *Biometrika* **89**(4): 939–946.

Aucott, L. S., Garthwaite, P. H. and Currall, J. (2000). Regression methods for high dimensional multicollinear data, *Communications in Statistics-Simulation and Computation* **29**(4): 1021–1037.

Boulesteix, A. L. and Strimmer, K. (2007). Partial least squares: a versatile tool for the analysis of high-dimensional genomic data, *Briefings in Bioinformatics* **8**(1): 32–44.

Breiman, L. (1996). Bagging predictors, *Machine Learning* **24**(2): 123–140.

Breiman, L. (2001). Random forests, *Machine Learning* **45**(1): 5–32.

Bühlmann, P. and Van de Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*, Springer.

Burden, F. R., Brereton, R. G. and Walsh, P. T. (1997). Cross-validatory selection of test and validation sets in multivariate calibration and neural networks as applied to spectroscopy, *Analyst* **122**(10): 1015–1022.

Candes, E. and Tao, T. (2007). The dantzig selector: Statistical estimation when p is much larger than n, *The Annals of Statistics* **35**(6): 2313–2351.

Cantoni, E. and Ronchetti, E. (2001). Robust inference for generalized linear models, *Journal of the American Statistical Association* **96**(455): 1022–1030.

Chun, H. and Keleș, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72**(1): 3–25.

Chung, D., Chun, H. and Keles, S. (2013). *SPLS: Sparse Partial Least Squares (SPLS) Regression and Classification*. R package version 2.2-1.
URL: *http://CRAN.R-project.org/package=spls*

Čížek, P. and Härdle, W. (2006). Robust estimation of dimension reduction space, *Computational Statistics & Data Analysis* **51**(2): 545–555.

De Jong, S. (1993a). PLS fits closer than PCR, *Journal of Chemometrics* **7**(6): 551–557.

De Jong, S. (1993b). SIMPLS: an alternative approach to partial least squares regression, *Chemometrics and Intelligent Laboratory Systems* **18**(3): 251–263.

Dupuis, D. J. and Victoria-Feser, M.-P. (2011). Fast robust model selection in large datasets, *Journal of the American Statistical Association* **106**(493): 203–212.

Dupuis, D. J. and Victoria-Feser, M.-P. (2013). Robust VIF regression with application to variable selection in large data sets, *The Annals of Applied Statistics* **7**(1): 319–341.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*, SIAM NSF-CBMS.

Efron, B. (2004). The estimation of prediction error, *Journal of the American Statistical Association* **99**(467): 619–642.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression, *The Annals of Statistics* **32**(2): 407–499.

Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap*, CRC press.

Engelen, S., Hubert, M., Branden, K. V. and Verboven, S. (2004). Robust PCR and robust PLSR: a comparative study, *Theory and applications of recent robust methods*, Springer, pp. 105–117.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association* **96**(456): 1348–1360.

Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space, *Statistica Sinica* **20**(1): 101–148.

Foster, D. P. and Stine, R. A. (2008). α-investing: a procedure for sequential control of expected false discoveries, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**(2): 429–444.

Frank, I. E. (1987). Intermediate least squares regression method, *Chemometrics and Intelligent Laboratory Systems* **1**(3): 233–242.

Frank, L. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools, *Technometrics* **35**(2): 109–135.

Furnival, G. M. and Wilson, R. W. (1974). Regressions by leaps and bounds, *Technometrics* **16**(4): 499–511.

Garthwaite, P. H. (1994). An interpretation of partial least squares, *Journal of the American Statistical Association* **89**(425): 122–127.

Geladi, P. and Kowalski, B. R. (1986). Partial least-squares regression: a tutorial, *Analytica Chimica Acta* **185**: 1–17.

Gray, H. L. and Schucany, W. R. (1972). *The generalized jackknife statistic*, Marcel Dekker New York.

Hastie, T. and Efron, B. (2013). *lars: Least Angle Regression, Lasso and Forward Stagewise*. R package version 1.2.
URL: *http://CRAN.R-project.org/package=lars*

Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Friedman, J. and Tibshirani, R. (2009). *The elements of statistical learning*, Springer.

Helland, I. S. (1988). On the structure of partial least squares regression, *Communications in Statistics-Simulation and Computation* **17**(2): 581–607.

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

Hinkley, D. V. (1977). Jackknifing in unbalanced situations, *Technometrics* **19**(3): 285–292.

Hinkley, D. and WET, B. C. (1984). Improvements of jackknife confidence limit methods, *Biometrika* **71**(2): 331–339.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics* **12**(1): 55–67.

Huang, X., Pan, W., Park, S., Han, X., Miller, L. W. and Hall, J. (2004). Modeling the relationship between LVAD support time and gene expression changes in the human heart by penalized partial least squares, *Bioinformatics* **20**(6): 888–894.

Huber, P. J. (1973). Robust regression: asymptotics, conjectures and Monte Carlo, *The Annals of Statistics* **1**(5): 799–821.

Huber, P. J. (2011). *Robust statistics*, Springer.

Huber, P. J. and Ronchetti, E. M. (2009). *Robust Statistics*, Wiley, New York.

Hubert, M., Rousseeuw, P. J. and Van Aelst, S. (2008). High-breakdown robust multivariate methods, *Statistical Science* **23**(1): 92–119.

Hubert, M. and Verboven, S. (2003). A robust PCR method for high-dimensional regressors, *Journal of Chemometrics* **17**(8-9): 438–452.

Hulland, J. (1999). Use of partial least squares PLS in strategic management research: a review of four recent studies, *Strategic Management Journal* **20**(2): 195–204.

Johnson, R. A. and Wichern, D. W. (2002). *Applied multivariate statistical analysis*, Prentice Hall.

Jolliffe, I. (2005). *Principal component analysis*, Wiley Online Library.

Jolliffe, I. T. (1982). Note on the use of principal components in regression, *Applied Statistics* **31**(3): 300–303.

Jolliffe, I. T. (1995). Rotation of principal components: choice of normalization constraints, *Journal of Applied Statistics* **22**(1): 29–35.

Jolliffe, I. T., Trendafilov, N. T. and Uddin, M. (2003). A modified principal component technique based on the LASSO, *Journal of Computational and Graphical Statistics* **12**(3): 531–547.

Jouan-Rimbaud, D., Massart, D. L., Leardi, R. and De Noord, O. E. (1995). Genetic algorithms as a tool for wavelength selection in multivariate calibration, *Analytical Chemistry* **67**(23): 4295–4301.

Kadane, J. B. and Lazar, N. A. (2004). Methods and criteria for model selection, *Journal of the American Statistical Association* **99**(465): 279–290.

Khan, J. A., Van Aelst, S. and Zamar, R. H. (2007). Robust linear model selection based on least angle regression, *Journal of the American Statistical Association* **102**(480): 1289–1299.

Kramer, R. (1998). *Chemometric techniques for quantitative analysis*, CRC Press.

Krasker, W. S. and Welsch, R. E. (1982). Efficient bounded-influence regression estimation, *Journal of the American Statistical Association* **77**(379): 595–604.

Lin, D., Foster, D. P. and Ungar, L. H. (2011). VIF regression: A fast regression algorithm for large data, *Journal of the American Statistical Association* **106**(493): 232–247.

Luo, W. (2008). *Spatial/Temporal Modelling of Crop Disease Data using High-Dimensional Regression*, PhD thesis, The University of Leeds.

Mallows, C. L. (1973). Some comments on c_p, *Technometrics* **15**(4): 661–675.

Mallows, C. L. (1995). More comments on c_p, *Technometrics* **37**(4): 362–372.

Manne, R. (1987). Analysis of two partial-least-squares algorithms for multivariate calibration, *Chemometrics and Intelligent Laboratory Systems* **2**(1): 187–197.

Maronna, R. A., Martin, R. D. and Yohai, V. J. (2006). *Robust Statistics: Theory and Methods*, John Wiley & Sons.

Marquaridt, D. W. (1970). Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation, *Technometrics* **12**(3): 591–612.

Martens, H. and Martens, M. (2000). Modified jack-knife estimation of parameter uncertainty in bilinear modelling by partial least squares regression (PLSR), *Food Quality and Preference* **11**(1): 5–16.

Martens, H. and Næs, T. (1989). *Multivariate calibration*, John Wiley & Sons.

Massy, W. F. (1965). Principal components regression in exploratory statistical research, *Journal of the American Statistical Association* **60**(309): 234–256.

McCann, L. and Welsch, R. E. (2007). Robust variable selection using least angle regression and elemental set sampling, *Computational Statistics & Data Analysis* **52**(1): 249–257.

Meier, L. and Bühlmann, P. (2007). Smoothing l_1-penalized estimators for high-dimensional time-course data, *Electronic Journal of Statistics* **1**: 597–615.

Mevik, B. H. and Wehrens, R. (2007). The PLS package: principal component and partial least squares regression in r, *Journal of Statistical Software* **18**(2): 1–24.

Mevik, B. H., Wehrens, R. and Liland, K. H. (2013). *pls: Partial Least Squares and Principal Component regression*. R package version 2.4-3.
URL: *http://CRAN.R-project.org/package=pls*

Miller, A. (2012). *Subset selection in regression*, CRC Press.

Miller, R. G. (1964). A trustworthy jackknife, *The Annals of Mathematical Statistics* **35**(4): 1594–1605.

Miller, R. G. (1974). The jackknife-a review, *Biometrika* **61**(1): 1–15.

Müller, S. and Welsh, A. (2005). Outlier robust model selection in linear regression, *Journal of the American Statistical Association* **100**(472): 1297–1310.

Næs, T., Isaksson, T., Fearn, T. and Davies, T. (2002). *A user-friendly guide to multivariate calibration and classification*, NIR Publications Chichester.

Pison, G., Van Aelst, S. and Willems, G. (2002). Small sample corrections for LTS and MCD, *Metrika* **55**(1-2): 111–123.

Quenouille, M. H. (1949). Approximate tests of correlation in time-series, *Journal of the Royal Statistical Society. Series B (Methodological)* **11**(1): 68–84.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *http://www.R-project.org/*

Reeds, J. A. (1978). Jackknifing maximum likelihood estimates, *The Annals of Statistics* **6**(4): 727–739.

Ronchetti, E. (1985). Robust model selection in regression, *Statistics & Probability Letters* **3**(1): 21–23.

Ronchetti, E. and Staudte, R. G. (1994). A robust version of mallows' c_p, *Journal of the American Statistical Association* **89**(426): 550–559.

Rosipal, R. and Krämer, N. (2006). Overview and recent advances in partial least squares, *Subspace, Latent Structure and Feature Selection*, Springer, pp. 34–51.

Rousseeuw, P. J. (1984). Least median of squares regression, *Journal of the American Statistical Association* **79**(388): 871–880.

Rousseeuw, P. J. and Van Driessen, K. (2006). Computing LTS regression for large data sets, *Data Mining and Knowledge Discovery* **12**(1): 29–45.

Rousseeuw, P. and Yohai, V. (1984). Robust regression by means of S-estimators, *Robust and nonlinear time series analysis*, Springer, pp. 256–272.

Salibian-Barrera, M. and Van Aelst, S. (2008). Robust model selection using fast and robust bootstrap, *Computational Statistics & Data Analysis* **52**(12): 5121–5135.

Schumann, S., Nolte, L. P. and Zheng, G. (2013). Comparison of partial least squares regression and principal component regression for pelvic shape prediction, *Journal of Biomechanics* **46**(1): 197–199.

Sen, P. K. (1988). Functional jackknifing: rationality and general asymptotics, *The Annals of Statistics* **16**(1): 450–469.

Shahriari, S., Faria, S. and Gonçalves, A. M. (2014). Variable selection methods in high-dimensional regression-a simulation study, *Communications in Statistics-Simulation and Computation* . (accepted).

Shahriari, S., Faria, S., Gonçalves, A. M. and Van Aelst, S. (2014). Outlier detection and robust variable selection for least angle regression, *Computational Science and Its Applications–ICCSA 2014*, Springer, pp. 512–522.

Shao, J. (1991). Consistency of jackknife variance estimators, *Statistics: A Journal of Theoretical and Applied Statistics* **22**(1): 49–57.

Shao, J. (1993). Linear model selection by cross-validation, *Journal of the American Statistical Association* **88**(422): 486–494.

Shao, J. and Wu, C. J. (1989). A general theory for jackknife variance estimation, *The Annals of Statistics* **17**(3): 1176–1197.

Stone, M. and Brooks, R. J. (1990). Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression, *Journal of the Royal Statistical Society, Series B (Methodological)* **52**(2): 237–269.

Sugiura, N. (1978). Further analysts of the data by Akaike's information criterion and the finite corrections: Further analysts of the data by Akaike's, *Communications in Statistics - Theory and Methods* **7**(1): 13–26.

Swierenga, H., De Weijer, A., Van Wijk, R. and Buydens, L. (1999). Strategy for constructing robust multivariate calibration models, *Chemometrics and Intelligent Laboratory Systems* **49**(1): 1–17.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO, *Journal of the Royal Statistical Society. Series B (Methodological)* **58**(1): 267–288.

Trevor, H., Robert, T. and Jerome, F. (2001). The elements of statistical learning: data mining, inference and prediction, *New York: Springer-Verlag* **1**(8): 371–406.

Tukey, J. W. (1958). Bias and confidence in not-quite large samples, *Annals of Mathematical Statistics*, Vol. 29, Inst Mathematical Statistics, pp. 614–614.

Vines, S. (2000). Simple principal components, *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **49**(4): 441–451.

Wehrens, R. and Van der Linden, W. E. (1997). Bootstrapping principal component regression models, *Journal of Chemometrics* **11**(2): 157–171.

Weisberg, S. (2014). *Applied linear regression*, John Wiley & Sons.

Wentzell, P. D. and Vega Montoto, L. (2003). 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.

Westad, F. and Martens, H. (2000). Variable selection in near infrared spectroscopy based on significance testing in partial least squares regression, *Journal of Near Infrared Spectroscopy* **8**(2): 117–124.

Wold, H. (1966). Estimation of principal components and related models by iterative least squares, *Multivariate analysis* **1**: 391–420.

Wold, H. (1975a). *Path models with latent variables: The NIPALS approach*, Acad. Press.

Wold, H. (1975b). Soft modeling by latent variables: the nonlinear iterative partial least squares approach, *Perspectives in Probability and Statistics, Papers in Honour of MS Bartlett* : 520–540.

Wold, S., Ruhe, A., Wold, H. and Dunn, III, W. J. (1984). The collinearity problem in linear regression. the partial least squares (PLS) approach to generalized inverses, *SIAM Journal on Scientific and Statistical Computing* **5**(3): 735–743.

Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis, *The Annals of Statistics* **14**(4): 1261–1295.

Yao, W. and Wang, Q. (2013). Robust variable selection through MAVE, *Computational Statistics & Data Analysis* **63**: 42–49.

Zhou, J., Foster, D. P., Stine, R. A. and Ungar, L. H. (2006). Streamwise feature selection, *The Journal of Machine Learning Research* **7**: 1861–1885.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**(2): 301–320.

Zou, H. and Hastie, T. (2012). *elasticnet: Elastic-Net for Sparse Estimation and Sparse PCA*. R package version 1.1.
URL: *http://CRAN.R-project.org/package=elasticnet*

Zou, H., Hastie, T. and Tibshirani, R. (2006). Sparse principal component analysis, *Journal of Computational and Graphical Statistics* **15**(2): 265–286.

1.6