UNIVERSITI PUTRA MALAYSIA

SUPPORT VECTOR MACHINE AND ITS APPLICATIONS FOR LINEAR AND NONLINEAR REGRESSION IN THE PRESENCE OF OUTLIERS OF HIGH DIMENSIONAL DATA

WALEED DHHAN SLEABI

SUPPORT VECTOR MACHINE AND ITS APPLICATIONS FOR LINEAR AND NONLINEAR REGRESSION IN THE PRESENCE OF OUTLIERS OF HIGH DIMENSIONAL DATA

By

WALEED DHHAN SLEABI

Thesis Submitted to the School of Graduated Studies, Universiti Putra Malaysia, in Fulfillment of the Requirements for the Degree of Doctor of Philosophy

September 2016

DEDICATIONS

- *To the spirit of my father since he passed away when I was at Malaysia 13 March 2013 and I still have remembered his words " my son, do not worry about anything , I pray for you all the time".*

- *To my respectful mother, who has taught me a lot on the meaning of persistency in life.*

- *To my beloved wife for all her contribution, patience and understanding throughout my doctoral studies. She incredibly supported me and made it all possible for me.*

- *To my kids, Mohammed Sadeq, and Hayderali who were accompanying me in all different parts of my study and their love have always been my greatest inspiration.*

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfillment of the requirements for the degree of Doctor of Philosophy

SUPPORT VECTOR MACHINE AND ITS APPLICATIONS FOR LINEAR AND NONLINEAR REGRESSION IN THE PRESENCE OF OUTLIERS OF HIGH DIMENSIONAL DATA

By

WALEED DHHAN SLEABI

September 2016

Chairman : Md. Sohel Rana, PhD
Faculty : Science

The ordinary least squares (OLS) is reported as the most commonly used method to estimate the relationship between variables (inputs and output) in the linear regression models because of its optimal properties and ease of calculation. Unfortunately, the OLS estimator is not efficient in cases of the presence of outliers in a data set, nonlinear relationships and high dimensional problems. Thus, the search for alternatives that feature the necessary flexibility to handle them has become an urgent necessity such as nonparametric approaches. Consequently, the support vector regression (SVR) is used as an alternative to OLS.

In this thesis, at first, we consider the identification of outliers through the SVR. In regression, outliers can be classified into two different types, such as vertical outlier and leverage points (good and bad leverage points). It is very important to identify outliers and bad leverage points (BLP) because of their significant effects on estimators. Most of the parametric diagnostic measures are considered good leverage points as bad leverage points. Hence, new nonparametric techniques are proposed for identification outliers that we call the fixed parameters support vector regression methods (FP-SVR). The results of real applications and simulation studies showed that the proposed methods have advantages over classical methods to identify vertical outliers and bad leverage points.

Further, in this thesis, the GM6 version of the robust estimation methods was developed only to identify and inhibit the influence of leverage points (LB) without taking into consideration whether it is good or bad. Thus, a new class of GM-estimators based on FP-SVR technique is developed takes into account minimizing the impact of the bad leverage points only on the model, and we call it GM-SVR. The results show that the performance of the GM-SVR is the best overall, followed by GM6 for all possible combinations of size of samples and percentages of contamination.

This thesis also addresses the problem of high dimensionality in linear and nonlinear regression models. It is well known that the support vector regression has the ability to introduce sparse models (less complexity). Unfortunately, there is a potential problem: if the value of threshold is small (ε near zero), the resulting model depends on a greater number of the training data points, thus making the solution more complexity (non-sparse).Therefore, the single index support vector regression (SI-SVR) model is proposed which combines the flexibility of the nonparametric model and the high accuracy of the parametric model. The real and simulation studies pointed out that the proposed method has the ability to address the problem of high dimensionality.

This thesis also explores the problem of high dimensionality when the number of predictors p larger than the sample size n. Although, we have proposed the SI-SVR to solve the problem of high dimensionality but this model does not have the ability to modeling examples with rank deficient. Furthermore, the efficiency of the resulting SI-SVR model can be decreased and less accurate predictions will be produced when unnecessary predictors are included in the model. Hence, a new method is suggested to overcome this issue using the Elastic Net technique for selecting significant variables which we call the elastic net single index support vector regression (ENSI-SVR). The comparison results show that the ENSI-SVR is an efficient method in dealing with sparse data to achieve dimension reduction which allows applying the SI-SVR easily.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Doktor Falsafah

MESIN VEKTOR SOKONGAN DAN PENGGUNAAN UNTUK REGRESI LINEAR DAN BUKAN LINEAR DENGAN KEHADIRAN TITIK TERPENCIL DAN DATA BERDIMENSI TINGGI

Oleh

WALEED DHHAN SLEABI

September 2016

Pengerusi : Md. Sohel Rana, PhD
Fakulti : Sains

Kaedah biasa kuasa dua terkecil (OLS) dilaporkan sebagai kaedah yang paling biasa digunakan untuk menganggarkan hubungan antara pembolehubah (input dan output) dalam model regresi linear kerana sifat-sifat yang optimum dan memudahkan dalam pengiraan. Malangnya, penganggar OLS tidak berkesan dalam kehadiran titik terpencil, hubungan tidak linear dan masalah dimensi yang tinggi. Oleh itu, mencari alternatif yang fleksibil telah menjadi satu keperluan yang segera bagi mengendalikan titik terpencil seperti kaedah tidak berparameter. Oleh itu, regresi vektor sokongan (SVR) digunakan sebagai alternatif kepada OLS.

Dalam tesis ini, pada mulanya, kami mengambil kira mengenalpasti titik terpencil melalui SVR . Dalam regresi, titik terpencil boleh diklasifikasikan kepada dua jenis yang berbeza, seperti titik terpencil menegak dan titik tuasan (titik tuasan baik dan buruk). Sangat penting untuk mengenalpasti titik terpencil dan titik tuasan buruk (BLP), ini kerana ianya mempunyai kesan yang signifikan ke atas penganggar. Kebanyakan pengukur diagnostik berparameter mempertimbangkan titik tuasan baik sebagai titik tuasan buruk. Oleh itu, teknik tidak berparameter baru dicadangkan untuk mengenalpasti titik luaran dimana ianya dinamakan Parameter Tetap Menyokong Kaedah Regresi Vektor (FP-SVR). Keputusan bagi aplikasi sebenar dan kajian simulasi menunjukkan kaedah yang dicadangkan mempunyai kebaikan berbanding dengan kaedah yang sedia ada bagi mengenal pasti titik terpencil menegak dan titik tuasan buruk .

Selanjutnya, dalam tesis ini, versi GM6 kaedah anggaran teguh telah dicadangkan hanya untuk mengenalpasti dan menghalang pengaruh titik tuasan tanpa mengambil kira sama ada ianya baik atau buruk. Oleh itu, kaedah baharu penganggar GM berdasarkan parameter tetap menyokong teknik regresi vektor dibangunkan dengan mengambil kira minimumkan kesan daripada titik tuasan buruk hanya kepada model, dan kami menamakannya GM-SVR. Keputusan menunjukkan prestasi GM-SVR adalah yang paling baik untuk keseluruhan, diikuti dengan GM6 untuk kesemua kemungkinan kombinasi saiz sampel dan peratusan data yang tercemar.

Tesis ini juga menangani masalah dimensi tinggi dalam model regresi linear dan tidak linear. Adalah diketahui umum bahawa regresi sokongan vektor mempunyai keupayaan untuk memperkenalkan model jarang (kurang rumit). Malangnya, wujudnya masalah berpotensi: jika nilai ambang adalah kecil (ε hampir sifar), model yang terhasil bergantung kepada bilangan yang lebih besar daripada titik data latihan, dengan itu membuat penyelesaian menjadi lebih kompleks (tidak jarang). Oleh itu, model Tunggal Sokongan Indeks Vektor Regresi (SI-SVR) dicadangkan bagi menggabungkan fleksibiliti model tidak berparameter dan ketepatan yang tinggi bagi model berparameter. Kajian sebenar dan simulasi menunjukkan kaedah yang dicadangkan mempunyai keupayaan untuk menangani masalah dimensi tinggi.

Tesis ini juga meneroka masalah dimensi tinggi apabila bilangan peramal p lebih besar daripada saiz sampel n. Walaupunbagaimanapun, kami mencadangkan SI-SVR untuk menyelesaikan masalah dimensi tinggi tetapi model ini tidak mempunyai keupayaan untuk model contoh dengan susunan kekurangan. Tambahan pula, kecekapan model SI-SVR yang terhasil boleh menurun dan ramalan kurang tepat akan dihasilkan apabila peramal yang tidak perlu dimasukkan dalam model. Oleh itu, satu kaedah baru dicadangkan untuk mengatasi isu ini dengan menggunakan teknik bersih anjal untuk memilih pembolehubah penting yang kami namakan Sokongan Bersih Indeks Tunggal Vektor Regresi Elastik (ENSI-SVR). Keputusan pembandingan mendapati ENSI-SVR merupakan kaedah yang berupaya dalam berurusan dengan data jarang untuk mencapai pengurangan dimensi yang membolehkan penggunaan SI-SVR dengan mudah.

ACKNOWLEDGEMENTS

First and foremost, I would like to give thanks to my God, who has provided me His strength and grace throughout my doctoral pursue.

Heartfelt appreciation also goes to my committee chairperson, Dr. Sohel Rana for his constant inspiration, efficient guidance, and constructive feedback rendered. I am deeply honored to have the opportunity to complete my degree under his supervision.

I would also like to thank my internal co-supervisors, Prof. Dr. Habshah Midi and Associate Prof. Dr. Ibragimov Gafurjan for all their support and guidance provided.

I would like to express my sincere appreciation and deepest gratitude to the Ministry of Municipalities and Public Works (MMPW) for providing me scholarship and also to all my colleagues whom are always willing to help me.

My special thanks go to my beloved wife, for standing by with me patiently with her never ending encouragement, prayers and support throughout my doctoral pursue.

Also, I would like to thank my children; MD. Sadeq, and Hayderali for giving me the happiness during my study.

I certify that a Thesis Examination Committee has met on 9 September 2016 to conduct the final examination of Waleed Dhhan Sleabi on his thesis entitled "Support Vector Machine and its Applications for Linear and Nonlinear Regression in the Presence of Outliers of High Dimensional Data" in accordance with the Universities and University Colleges Act 1971 and the Constitution of the Universiti Putra Malaysia [P.U.(A) 106] 15 March 1998. The Committee recommends that the student be awarded the Doctor of Philosophy.

Members of the Thesis Examination Committee were as follows:

Mohd Rizam bin Abu Bakar, PhD
Associate Professor
Faculty of Science
Universiti Putra Malaysia
(Chairman)

Noor Akma binti Ibrahim, PhD
Professor
Faculty of Science
Universiti Putra Malaysia
(Internal Examiner)

Norazan Mohamed Ramli, PhD
Senior Lecturer
Universiti Teknologi MARA
Malaysia
(External Examiner)

Muhammad Hanif Mian, PhD
Professor
Lahore University of Management Science
Pakistan
(External Examiner)

NOR AINI AB. SHUKOR, PhD
Professor and Deputy Dean
School of Graduate Studies
Universiti Putra Malaysia

Date: 3 November 2016

This thesis was submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfillment of the requirement for the degree of Doctor of Philosophy. The members of the Supervisory Committee were as follows:

Md. Sohel Rana, PhD
Senior Lecturer
Faculty of Science
Universiti Putra Malaysia
(Chairman)

Habshah bt Midi, PhD
Professor
Faculty of Science
Universiti Putra Malaysia
(Member)

Ibragimov Gafurjan, PhD
Associate Professor
Faculty of Science
Universiti Putra Malaysia
(Member)

BUJANG BIN KIM HUAT, PhD
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

Declaration by graduate student

I hereby confirm that:

- this thesis is my original work;
- quotations, illustrations and citations have been duly referenced;
- this thesis has not been submitted previously or concurrently for any other degree at any other institutions;
- intellectual property from the thesis and copyright of thesis are fully-owned by Universiti Putra Malaysia, as according to the Universiti Putra Malaysia (Research) Rules 2012;
- written permission must be obtained from supervisor and the office of Deputy Vice-Chancellor (Research and Innovation) before the thesis is published (in the form of written, printed or in electronic form) including books, journals, modules, proceedings, popular writings, seminar papers, manuscripts, posters, reports, lecture notes, learning modules or any other materials as stated in the Universiti Putra Malaysia (Research) Rules 2012;
- there is no plagiarism or data falsification/ fabrication in the thesis, and scholarly integrity is upheld as according to the Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2012-2013) and the Universiti Putra Malaysia (Research) Rules 2012. The thesis has undergone plagiarism software.

Signature: ______________________ Date: ______________

Name and Matric No.: Waleed Dhhan Sleabi, GS37423

Declaration by Members of Supervisory Committee

This is to confirm that:

- the research conducted and the writing of this thesis was under our supervision;
- supervision responsibilities as stated in the Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2012-2013) are adhered to.

Signature: ______________________________
Name of Chairman of Supervisory Committee: Md. Sohel Rana, PhD

Signature: ______________________________
Name of Member of Supervisory Committee: Habshah bt Midi, PhD

Signature: ______________________________
Name of Member of Supervisory Committee: Ibragimov Gafurjan, PhD

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

LIST OF ABBREVIATIONS

BIF	Bounded Influence Function
BLP	Bad Leverage Points
BLUE	Best Linear Unbiased Estimators
BP	Breakdown Point
CDE	Chaos Differential Evolution
DE	Differential Evolution
DOF	Degrees Of Freedom
EGO	Algorithm of Global Optimization
EN	Elastic Net Method
ENSI	Elastic Net Single Index
FP	Fixed Parameters
GM	Generalized M estimators
GS	Grid Search Procedure
HLP	High Leverage Point
IF	Influence Function
iid	Independent Identically Distributed
IRLS	Iteratively Reweighted Least Squares
KKT	Karush Kuhn Tucker Conditions
LASSO	Least Absolute Shrinkage Selection Operator
LAV	Least Absolute Values
LMS	Least Median of Squares
LP	Leverage Point
LTS	Least Trimmed Squares
MAD	Median Absolute Deviation
MCD	Minimum Covariance Determinant
MD	Mahalanobis Distance

MSE	Mean Square Error
MSVR	Modified Support Vector Regression
MVE	Minimum Volume Ellipsoid
NID	Normal Independent Distributed
NNR	Neural Network Regression
OLS	Ordinary Least Squares
OLSC	Ordinary Least Squares for clean data
PCA	Principal Component Analysis
PSO	Particle Swarm Optimization
RBF	Radial Basis Function
RMD	Robust Mahalanobis Distance
SE	Standard Error
SIM	Single Index Model
SLS	Semi-parametric Least Squares
SLT	Statistical Learning Theory
SRM	Structural Risk Minimization
SSVR	Standard Support Vector Regression
SV	Support Vector
SVC	Support Vector Classification
SVM	Support Vector Machine
SVR	Support Vector Regression
VAR	Variance of Residuals
WLS	Weighted Least Squares
WSLS	Weighted Semi-parametric Least Squares

CHAPTER 1

INTRODUCTION

1.1. Introduction and Background of the Study

Regression analysis is a statistical process which aims to explore the functional relationship between two or more variables so that, a dependent variable (output) can be predicted from one or more of independent variables (input) (Kutner et al., 2005). Regression analysis estimates the conditional expectation of the response variable given the explanatory variables. In other words, it estimates the average value of the dependent variable when the independent variables are fixed. This estimation can be done by using the proper technique for the phenomenon or the data set under study such as the ordinary least squares method. The ordinary least squares method (OLS) is classified as one of the prevalent estimation techniques in the regression analysis. Further, the OLS is the most popular estimation method in the linear regression community due to its superior properties and ease of computation, provided that the Gaussian Markov assumptions are met. In addition, the OLS estimator is the best linear unbiased estimator (BLUE), when the random errors are independent identically distributed (iid) normal. Unfortunately, the assumptions of the linear relationship between the variables and the normal distribution of the error term are violated in the most of the real life applications. Furthermore, the OLS estimator is not robust against unusual data points which often appear in real life applications. In other words, the OLS estimator has very low breakdown point which is equal to $1/n$ (Maronna et al., 2006), where n is the sample size. That is, even one point (abnormal) could change the estimate of least squares dramatically in the wrong direction (Rousseeow and Leroy, 1987; Kamruzzaman and Imon, 2002; Maronna et al., 2006).

The assumption of the normal distribution of the error term is violated in the presence of one or more outlier observations. Belsley et al. (1980) reported that the outliers are those points either alone or together with several other points have the largest influence on the computed values of different estimates. Hawkins (1980) defined an outlier observation as the observation that deviates so much from the other observations as to arouse suspicions which it was generated by a various mechanism. Muñoz-Garcia et al. (1990) defined the outlier observation as "An outlier is an observation which being atypical and/or erroneous deviates decidedly from the general behavior of experimental data with respect to the criteria which is to be analyzed on it". Barnett and

Lewis (1994) defined outlier points as those points that are markedly far from the majority of points in a data set. In general, there are several classes of outliers in the regression problems. Observations that are outlying in the Y-direction are expressed as outliers or vertical outliers. In contrast, the observations which are outlying in the X-direction are called high leverage points (HLP). However, there is an urgent need in the regression analysis to find out whether HLP have much impact on the fitting of a model or not (Belsley et al., 1980; Rousseeow and Leroy, 1987).

The other serious problems that affect the predicted model in addition to outliers and the non-linearity relationship among variables are problems of high-dimensional and sparse (p is larger than the number of observations n). The curse of high dimensionality refers to how certain algorithms such as algorithms in numerical analysis, sampling, combinatorics, machine learning and data mining that may perform poorly in high-dimensional data. The common theme of these problems is that when the dimensionality increases, the volume of the space increases so fast that the available data become sparse. This sparsity is problematic for any method that requires statistical significance. In high dimensional data, a matrix related to some algorithms may become singular and some additional information such as regularization, Bayesian prior and others need to be added to obtain standard solution.

Recently, several procedures which deal with these problems separately are available. However, there are not extensive studies reported in the literature which takes into consideration the presence of the non-linearity, outliers and high dimensional problems (full or less than full rank) simultaneously. As a result, the search for alternatives that feature the necessary flexibility to handle these issues has become an urgent necessity such as nonparametric methods especially learning machines.

1.2 Importance and Motivation of the Study

Nonparametric regression technique is a form of statistical regression analysis in which there is no a predetermined form of the predictor but it is constructed based on the information derived directly from the data. Whereas the classical regression statistical techniques stand upon a strict assumption in terms of they assume that the underlying probability distribution of the data is known and the relationship among the variables takes a linear form. However, in real applications, often we confront with distribution-free regression problems with a non-linear relationship between input and output variables (Ukil, 2007). One nonparametric method which is not requiring knowledge of the underlying probability distribution of the data, as well as its ability to deal with non-linear relationship is the support vector machine. Support vector machine (SVM) is one of the comparatively new and promising techniques for learning separating functions in classification problems (SVC) or for performing function estimation in regression problems (SVR).

Support vector machine was initially applied for classification tasks (Cortes and Vapnik 1995), but shortly, the formulation was extended to deal with regression problems (Smola, and Vapnik 1997; Vapnik 1995). The advantages of support vector machine are its ability to modeling the non-linear relationships by employing kernel trick and its excellent generalization ability on the real applications of the classification and regression problems while it is still capable of producing sparse model (not all observations are needed to find the optimal model) (Ceperic et al. 2014). The common formulation of support vector machine for regression is Vapnik's ε-tube SV regression (ε-SVR) (Smola, and Vapnik 1997). The ε-SVR produces predictive model depends only on a subset of the training points whereas it ignores any points within the threshold ε. This step reveals the potential problem: if the value of threshold ε is small, then the resulting model depends on a greater number of the overall training points, thus making the resulting solution non-sparse, as demonstrated in Guo et al. (2010).

Both of the parametric and nonparametric regression techniques are affected by the presence of single or multiple enormous points in a data (the parametric methods certainly are most influenced than nonparametric methods). Many researchers reported that the real data sets mostly contain unusual points ranging from 1% to 10% (Hampel et al. 1986; Wilcox, 2005). Outliers and HLP have a great effect on the values of various estimates, which leads to misleading conclusions result in wrong decisions. Hence, it is necessary to detect those unusual observations and removing them before embarking on building the predictive model (Cook, 1977) or orientation of the robust methods (Huber, 1973) which minimize the impact of outliers instead of removing them completely from the data. It is worth mentioning that the choose one of these methods is up to the researcher.

There are several parametric methods used for detecting single or multiple outliers and HLP. Unfortunately, they are not successful to identify multiple abnormal points in the data sets due to the effects of masking and swamping problems (Rousseeuw and Leroy, 1987). On the other hand, these methods can not deal with less than full rank data. To address this problem some researchers explored the use of non-parametric methods for outlier detection in cases both of full rank and less than full rank. Jordaan and Smits (2004) suggested using standard support vector regression (SSVR) for outlier detection. The idea of this technique is by running the SV regression model many times and detects points which are suspected as outliers. Nishiguchi et al. (2010) pointed out that some problems arise when applying it with real applications. It requires high computational costs for multiple outliers in the data because detection of an outlier requires a number of iterations of the calculation; the trial and error is used for accurate detection, since it is not clear how to identify the outlier threshold value. To remedy this problem, Nishiguchi et al. (2010) developed the modified support vector regression (MSVR) technique for outlier detection by employing new trade-off parameter (μ), which is successful in identifying outliers and HLP. Nonetheless, the MSVR approach is suitable for few outliers in the data, since one iteration is required to detect one outlier. Consequently, computational costs become close to those arising from the standard SVM regression method in case of presence multiple outliers. Further, there is no clear rule for choosing the value of threshold parameter, although it comes with fixed value of this parameter. The shortcoming of these methods has inspired us to develop new techniques to improve the performance of standard SVM regression for outlier detection, which we call the fixed parameters support vector regression (FP-SVR). The proposed two methods are expected to achieve accurate detection of outliers and HLP (only bad leverage points) with fixed parameters during one iteration.

This thesis also concerned on the use of robust methods to address the problem of the presence of outliers and bad leverage points (BLP) in multiple linear regression models. As we mentioned previously the OLS estimator is seriously affected by the presence of outliers. One of the most common alternative techniques to OLS of addressing the presence of outliers is the robust regression procedure (Hampel, 1974). There are many robust regression methods in the literature, such as the least absolute values (LAV), the M-estimator, generalized M-estimator (GM1-estimator), the least median of squares (LMS), the S estimator, the least trimmed squares (LTS), the MM estimator and new class of GM-estimator (GM6) proposed by Coakley and Hettmansperger (1993). Yohai and Zamar (1988) firmly recommended that one of the goals of robust regression technique is to achieve: (a) a high breakdown point of nearly 50%, (b) a bounded influence function and (c) a high efficiency, simultaneously. According to this recommendation, only GM6 method achieves the three conditions, (a), (b) and (c) simultaneously. Regrettably, this method considers the good leverage points to be bad leverage points, which means that its efficiency tends to decrease with the presence of “good” leverage points. This limitation has inspired us to develop a new class of GM-estimators based on a fixed parameters support vector regression techniques that have been proven in Chapter 3, takes into account minimizing the impact of the bad leverage points only on the model, and we call it GM-SVR.

This thesis also addresses the problem of high dimensionality in linear and nonlinear regression models. It should be noted that the sparsity feature (less complexity), which is characterized by the SVR model by itself is not sufficient to ensure good generalization to the model in addition to the problem of non-sparse that accompany the small threshold, ε near zero (Ceperic et al. 2014). It is well known the support vector regression is a fully nonparametric approach, which makes it a flexible but at the same time it is suffering from precision decrease when increasing the covariates which is called the curse of the high-dimensionality (Härdle et al. 2004). For this reason, the alternative is used to cope with this drawback. One of the common techniques to improve generalization accuracy and overcome the curse of the high dimensional problem is the single index model. Ichimura (1993) suggested a semi-parametric model which combines between the flexibility of the nonparametric model and the high accuracy of the parametric model called single index model. This model summarizes the covariates within a single variable called index. To the best of our knowledge, there is no existing research in literature which used SVR to evaluate the unknown link function of the single index model. This inspires us to propose a new technique that uses the SVR model to estimate the unknown link function of the single index model namely the single index support vector regression (SI-SVR).

It should be stated that the SI-SVR model does not have the ability to modeling the rank deficient data. Furthermore, the efficiency of the resulting model could be declined, and less accurate predictions will be produced when unnecessary predictors are included in the model (Tibshirani, 1996; Hastie et al., 2009). This requires development of a new method to overcome this issue. This can be done by employing the concept of variables selection to achieve the possibility of modeling by single index model which we call the elastic net single index support vector regression (ENSI-SVR).

1.3 Research Objectives

The main goal of this thesis is to investigate the high dimensionality problems for linear and nonlinear regression models in the presence of outliers (outlying in coordinates X and Y). The classical estimation methods such as the ordinary least squares (OLS) method are not robust against outliers. Moreover, they can not evaluate the nonlinear relationships and the difficulty to meet all the assumptions for high-dimensional data. The foremost objectives of our research can be outlined systematically as follows:

1. To propose new improved diagnostic methods for the identification of multiple outliers based on two types of kernel functions.

2. To formulate a new robust estimation method to remedy the presence of outliers in the data for the linear regression model.
3. To propose a new semi-parametric method to cope the curse of high dimensionality combines between the high precision of parametric methods and the flexibility of nonparametric methods.
4. To develop the elastic net penalty approach for selecting variables in a single index support vector regression model to overcome the curse of high dimensionality when the number of predictors, p is larger than sample size n.

1.4 Scope and Limitation of the Study

The linear and nonlinear regression models are widely used in many areas of studies such as bioinformatics, economics, financial predictions and social sciences. In the real situation, these regression models have many practical uses. However, the most applications of the linear regression models are evaluated using the OLS method because of the ease of computation and its optimal properties when the underlying assumptions are met. In reality, the OLS estimator is not resistant to outlying samples; even one outlier can destroy the OLS estimator. The alternative procedures which used to address this issue are detection methods and robust statistical methods. Flexible techniques are suggested to the identification of outliers and HLP such as SSVR and MSVR in cases of full and less than full rank data. Nonetheless, these existing methods basically focus only on the identification of leverage points without taking into consideration their classification into good and bad leverage points. It is very important to detect and classify the good and bad leverage points, as only bad leverage points are responsible for the misleading conclusion about the fitting of the regression model. On the other hand, many robust statistical estimation techniques are suggested such as LMS-estimator, LTS-estimator, M-estimator, GM1-estimator, MM-estimator, and GM6-estimator. However, some of these methods are not robust against leverage points and some methods are considered the good leverage points as bad leverage points.

The other technique of statistical modeling is the nonparametric procedure which used to evaluate the nonlinear relationships and high dimensional problems including when the number of predictors p much greater than sample size n. One of the most effective methods in the nonparametric machine learning community is the support vector machine (Frohlich and Zell, 2005). However, the ability of the SVM model to evaluate the high dimensional problems is decreased because of the resulting model is non-sparse when the threshold is small. Furthermore, the generalization performance of SVM depends heavily on the right selection of the hyper-parameters C and ε, so the major issue for practitioners attempting to apply SVM is how to set these parameter values to guarantee a good generalization performance for a training data set. It should be noted all calculations have been implemented using R software.

1.5 Overview of the Thesis

In accordance with the objectives and the scope of the study, the contents of this thesis are structured in the eight chapters. The thesis chapters are organized so that the study objectives are apparent and are conducted in the sequence outlined.

Chapter Two: This chapter briefly presents the literature review of the least squares estimation method and the violations of its underlying assumptions such as the departure of normality and the presence of outliers. The literature review of the support vector machine for regression and its basic idea to employ the kernel trick during the estimation process are highlighted. The outliers, and leverage points and their diagnostics methods are also discussed. Moreover, basic concepts of robust linear regression and some important existing robust regression methods are also reviewed. Bootstrapping methods are also briefly discussed. In this chapter, the main idea of the single index model and its estimation methods are also discussed. Finally, the concept of variable selection and some of penalization methods are also briefly highlighted.

Chapter Three: This chapter discusses the existing SSVR and MSVR which are developed by Jordaan and Smits (2004) and Nishiguchi et al. (2010). The new proposed methods (FP-SVR) for the identification of multiple vertical outliers and bad leverage points are presented in this chapter. The steps for proposed FP-SVR methods and its algorithm are also highlighted. Finally, some real and simulation studies are discussed to evaluate the performance of the proposed methods.

Chapter Four: This chapter deals with the development of the GM-estimator based on FP-SVR (denoted by GM-SVR) for data having outliers and bad leverage points. Two Monte Carlo simulation studies and two numerical examples are carried out to assess the performance of the proposed method.

Chapter Five: In this chapter, we present the proposed semi-parametric model to address the high dimensional problem, namely the single-index support vector regression (denoted by SI-SVR).The new proposed technique is useful to get rid the so-called the curse of high dimensionality. In this respect, two types of data are considered, the linear and nonlinear relationships. The numerical and simulation examples are also discussed to assess our proposed method.

Chapter Six: In this chapter, the concept of variable selection is utilized to achieve non-singular predictive matrix when the number of predictors p larger than sample size n. Then, the proposed model, namely the elastic net single-index support vector regression (denoted by ENSI-SVR) can be used to remedy the curse of high dimensionality. The semi-parametric proposed model combines the high accuracy of parametric methods and the flexibility of nonparametric methods. A Monte Carlo simulation studies and numerical example are given to assess the performance of the proposed method.

Chapter Seven: This chapter provides the summary and detailed discussions of the thesis conclusions. Areas for future research are also recommended.

REFERENCES

Algamal, Z. Y., & Lee, M. H. (2015). Adjusted adaptive lasso in high-dimensional poisson regression model. *Modern Applied Science,* 9(4), pp170-177.

Alguraibawi, M., Midi, H., & Imon, A. (2015). A new robust diagnostic plot for classifying good and bad high leverage points in a multiple linear regression model. *Mathematical Problems in Engineering.* Volume 2015, Article ID 279472, pp 1-12

An, C., & Nguyen, T. Q. (2008). Statistical learning based intra prediction in H. 264. *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on,* 2800-2803.

Andersen, R. (2008). *Modern methods for robust regression*. Sage. Los Angeles.

Anderson, C., & Schumacker, R. E. (2003). A comparison of five robust regression methods with ordinary least squares regression: Relative efficiency, bias, and test of the null hypothesis. *Understanding Statistics: Statistical Issues in Psychology, Education, and the Social Sciences, 2*(2), 79-103.

Andreou, P. C., Charalambous, C., & Martzoukos, S. H. (2009). European option pricing by using the support vector regression approach. *Artificial neural Networks–ICANN* (pp. 874-883) Springer.

Andrews, D. F. (1974). A robust method for multiple linear regression. *Technometrics, 16*(4), 523-531.

Armstrong, R. D., & Kung, M. T. (1978). Algorithm AS 132: Least absolute value estimates for a simple linear regression problem. *Applied Statistics,* , 363-366.

Bagheri, A., Midi, H., Ganjali, M., & Eftekhari, S. (2010). A comparison of various influential points diagnostic methods and robust regression approaches: Reanalysis of interstitial lung disease data. *Applied Mathematical Sciences, 4*(28), 1367-1386.

Bao, Y., Lu, Y., & Zhang, J. (2004). Forecasting stock price by SVMs regression. *Artificial intelligence: Methodology, systems, and applications* (pp. 295-303) Springer.

Barnett, V., & Lewis, T. (1994). *Outliers in statistical data*. Wiley. New York.

Beaton, A. E., & Tukey, J. W. (1974). The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics, 16*(2), 147-185.

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*. John & Wiley, New York.

Ben-Hur, A., Ong, C. S., Sonnenburg, S., Schölkopf, B., & Rätsch, G. (2008). Support vector machines and kernels for computational biology. *PLoS Comput Biol, 4*(10), e1000173.

Bermingham, M., Pong-Wong, R., Spiliopoulou, A., Hayward, C., Rudan, I., Campbell, H., Navarro, P. (2015). Application of high-dimensional feature selection: Evaluation for genomic prediction in man. *Scientific Reports, 5(10312),* pp 1-12

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, 144-152.

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*, Cambridge University Press. Cambridge.

Calafiore, G. C. (2000). Outliers robustness in multivariate orthogonal regression. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, 30*(6), 674-679.

Ceperic, V., Gielen, G., & Baric, A. (2014). Sparse ε-tube support vector regression by active learning. *Soft Computing, 18*(6), 1113-1126.

Chan, W., Chan, C., Cheung, K., & Harris, C. (2001). On the modelling of nonlinear dynamic systems using support vector neural networks. *Engineering Applications of Artificial Intelligence, 14*(2), 105-113.

Chapelle, O., & Vapnik, V. (1999). Model selection for support vector machines. *Thirteenth Annual Neural Information Processing Systems (NIPS),* Denver, USA, pp 230-236.

Chen, D. S., & Jain, R. C. (1994). A robust backpropagation learning algorithm for function approximation. *Neural Networks, IEEE Transactions on, 5*(3), 467-479.

Cherkassky, V., & Ma, Y. (2004). Practical selection of SVM parameters and noise estimation for SVM regression. *Neural Networks, 17*(1), 113-126.

Cherkassky, V., & Mulier, F. M. (1998). *Learning from data: Concepts, theory, and methods* John Wiley & Sons.

Cherkassky, V., & Mulier, F. M. (2007). *Learning from data: Concepts, theory, and methods* John Wiley & Sons.

Chuang, C., Su, S., Jeng, J., & Hsiao, C. (2002). Robust support vector regression networks for function approximation with outliers. *Neural Networks, IEEE Transactions on, 13*(6), pp 1322-1330.

Coakley, C. W., & Hettmansperger, T. P. (1993). A bounded influence, high breakdown, efficient regression estimator. *Journal of the American Statistical Association, 88*(423), 872-880.

Colliez, J., Dufrenois, F., & Hamad, D. (2006). Robust regression and outlier detection with SVR: Application to optic flow estimation. The 17th British Machine Vision Association (*BMVC),* Edinburgh, UK, pp 1229-1238.

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics,* 19(1), pp 15-18.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273-297.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods* Cambridge University press, Cambridge.

De Brabanter, K., De Brabanter, J., Suykens, J. A., & De Moor, B. (2010). Optimized fixed-size kernel models for large data sets. *Computational Statistics & Data Analysis, 54*(6), 1484-1504.

Dhhan, W., Rana, S., & Midi, H. (2015). Non-sparse ϵ-insensitive support vector regression for outlier detection. *Journal of Applied Statistics, 42*(8), 1723-1739.

Draper, N. R., Smith, H., & Pownell, E. (1966). *Applied regression analysis* Wiley New York.

Efron, B. (1992). *Bootstrap methods: Another look at the jackknife*. Springer, New York, pp 569-593.

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

Figueiredo, M. A. (2003). Adaptive sparseness for supervised learning. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 25*(9), 1150-1159.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, 33*(1), 1-22.

Frohlich, H., & Zell, A. (2005). Efficient parameter selection for support vector machines in classification and regression via model-based global optimization. *Neural Networks, 2005. IJCNN'05. Proceedings. 2005 IEEE International Joint Conference on, , 3* 1431-1436.

Gandhi, A. B., Joshi, J. B., Jayaraman, V. K., & Kulkarni, B. D. (2007). Development of support vector regression (SVR)-based correlation for prediction of overall gas hold-up in bubble column reactors for various gas–liquid systems. *Chemical Engineering Science, 62*(24), 7078-7089.

Gervini, D., & Yohai, V. J. (2002). A class of robust and fully efficient regression estimators. *Annals of Statistics.* 30(2), 583-616.

Groß, J. (2003). *Linear regression*. Springer, Heidelberg, Germany.

Guo, B., Gunn, S. R., Damper, R. I., & Nelson, J. D. (2008). Customizing kernel functions for SVM-based hyperspectral image classification. *Image Processing, IEEE Transactions on, 17*(4), 622-629.

Guo, G., Zhang, J., & Zhang, G. (2010). A method to sparsify the solution of support vector regression. *Neural Computing and Applications, 19*(1), 115-122.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research, 3*, 1157-1182.

Hadi, A. S. (1992). A new measure of overall potential influence in linear regression. *Computational Statistics & Data Analysis, 14*(1), 1-27.

Hampel, F., Ronchetti, E., Rousseeuw, P., & Stahel, W. (1986). Robust statistics, J. *Wiley& Sons, New York,*

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association, 69*(346), 383-393.

Härdle, W. K., Hoffmann, L., & Moro, R. (2011). Learning machines supporting bankruptcy prediction. Statistical tools for finance and insurance. Springer, Heidelberg, pp. 225-250.

Härdle, W., Werwatz, A., Müller, M., & Sperlich, S. (2004). *Nonparametric and semiparametric models.*Springer, New York.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). Unsupervised learning. In The elements of statistical learning. Springer, New York, pp. 485-585.

Hawkins, D. M. (1980). *Identification of outliers*. Chapman and Hall, London.

Heinz, G., Peterson, L. J., Johnson, R. W., & Kerk, C. J. (2003). Exploring relationships in body dimensions. *Journal of Statistics Education, 11*(2), pp 225-233.

Hekimoğlu, S., & Erenoglu, R. C. (2013). A new GM-estimate with high breakdown point. *Acta Geodaetica Et Geophysica, 48*(4), 419-437.

Hill, R. W. (1977). Robust regression when there are outliers in the carriers, unpublished PhD thesis. Harvard University.

Hoerl, A. E., & Kennard, R. W. (1970a). Ridge regression: Applications to nonorthogonal problems. *Technometrics, 12*(1), 69-82.

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

Horowitz, J. L. (2009). *Semiparametric and nonparametric methods in econometrics*. Springer, New York.

Hsu, C., Chang, C., & Lin, C. (2003). A practical guide to support vector classification technical report department of computer science and information engineering. *National Taiwan University, Taipei,*

Hu, Y., Gramacy, R. B., & Lian, H. (2013). Bayesian quantile regression for single-index models. *Statistics and Computing, 23*(4), 437-454.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics, 35*(1), pp 73-101.

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and monte carlo. *The Annals of Statistics,*1(5) , pp 799-821.

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

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

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics, 58*(1), 71-120.

Jackson, D. A., & Chen, Y. (2004). Robust principal component analysis and outlier detection with ecological data. *Environmetrics, 15*(2), 129-139.

Jordaan, E. M., & Smits, G. F. (2004). Robust outlier detection using SVM regression. *Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on, , 3* 2017-2022.

Kamruzzaman, M., & Imon, A. (2002). High leverage point: Another source of multicollinearity. *Pakistan Journal of Statistics-All Series-, 18*(3), pp 435-448.

Keerthi, S. S., Shevade, S. K., Bhattacharyya, C., & Murthy, K. R. K. (2001). Improvements to platt's SMO algorithm for SVM classifier design. *Neural Computation, 13*(3), 637-649.

Kuo, T., & Yajima, Y. (2010). Ranking and selecting terms for text categorization via SVM discriminate boundary. *International Journal of Intelligent Systems, 25*(2), 137-154.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models*. McGraw-Hill Irwin, New York.

Kwok, J. T. (2001). Linear dependency between ε and the input noise in ε-support vector regression. *Artificial neural Networks—ICANN 2001* , Vienna, Austria, pp. 405-410.

Lahiri, S. K., & Ghanta, K. C. (2009). Support vector regression with parameter tuning assisted by differential evolution technique: Study on pressure drop of slurry flow in pipeline. *Korean Journal of Chemical Engineering, 26*(5), 1175-1185.

Lee, Y., & Mangasarian, O. L. (2001). SSVM: A smooth support vector machine for classification. *Computational Optimization and Applications, 20*(1), 5-22.

Li, X., & Kong, J. (2014). Application of GA–SVM method with parameter optimization for landslide development prediction. *Natural Hazards and Earth System Science, 14*(3), 525-533.

Liang, W., Zhang, L., & Wang, M. (2011). The chaos differential evolution optimization algorithm and its application to support vector regression machine. *Journal of Software, 6*(7), 1297-1304.

Liebmann, B., Friedl, A., & Varmuza, K. (2009). Determination of glucose and ethanol in bioethanol production by near infrared spectroscopy and chemometrics. Analytica Chimica Acta, 642(1), 171-178.

Liu, J. N., & Hu, Y. (2013). Support vector regression with kernel mahalanobis measure for financial forecast: In *Time series analysis, modeling and applications*. Springer, Heidelberg, pp. 215-227.

Lu, C., Lee, T., & Chiu, C. (2009). Financial time series forecasting using independent component analysis and support vector regression. *Decision Support Systems, 47*(2), 115-125.

Maronna, R. A., Martin, R. D., & Yohai, V. J. (2006). *Robust statistics*. John Wiley Chichester.

Mattera, D., & Haykin, S. (1999). *Support vector machines for dynamic reconstruction of a chaotic system: In Advances in Kernel Methods,* MIT Press, Cambridge, pp 211-241.

Mejía-Guevara, I., & Kuri-Morales, Á. (2007). Evolutionary feature and parameter selection in support vector regression. *MICAI 2007: Advances in artificial intelligence* (pp. 399-408) Springer.

Mickey, M. R., Jean Dunn, O., & Clark, V. (1967). Note on the use of stepwise regression in detecting outliers. *Computers and Biomedical Research, 1*(2), 105-111.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2015). *Introduction to linear regression analysis* John Wiley & Sons.

Muñoz-Garcia, J., Moreno-Rebollo, J., & Pascual-Acosta, A. (1990). Outliers: A formal approach. *International Statistical Review/Revue Internationale De Statistique, 58(3)*, pp 215-226.

Nakayama, H., & Yun, Y. (2006). Support vector regression based on goal programming and multi-objective programming. In the 2006 IEEE International Joint Conference on Neural Network Proceedings, Vancouver, BC, Canada.

Nishiguchi, J., Kaseda, C., Nakayama, H., Arakawa, M., & Yun, Y. (2010). Modified support vector regression in outlier detection. *Neural Networks (IJCNN), the 2010 International Joint Conference on,* 1-5.

Pell, R. J. (2000). Multiple outlier detection for multivariate calibration using robust statistical techniques. *Chemometrics and Intelligent Laboratory Systems, 52*(1), 87-104.

Peng, H., & Huang, T. (2011). Penalized least squares for single index models. *Journal of Statistical Planning and Inference, 141*(4), 1362-1379.

Rahmatullah Imon, A. (2005). *Identifying multiple influential observations in linear regression. Journal of Applied Statistics*, 32(9), 929-946.

Rojo-Álvarez, J. L., Martínez-Ramón, M., Figueiras-Vidal, A. R., García-Armada, A., & Artés-Rodríguez, A. (2003). A robust support vector algorithm for nonparametric spectral analysis. IEEE Signal Processing Letters 10(11), pp 320-323.

Rosseuw, P., & Van Zomeren, B. (1990). Unmasking multivariate outliers and leverage points (with discussion). *J.Amer.Statist.Assoc, 85*, 633-651.

Roth, V. (2004). The generalized LASSO. *Neural Networks, IEEE Transactions on, 15*(1), 16-28.

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

Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. *Mathematical Statistics and Applications, 8*, 283-297.

Rousseeuw, P. J., & Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association, 88*(424), 1273-1283.

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection*. New York: John Wiley & Sons.

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

Sato, J. R., Costafreda, S., Morettin, P. A., & Brammer, M. J. (2008). Measuring time series predictability using support vector regression. *Communications in Statistics—Simulation and Computation®, 37*(6), 1183-1197.

Schölkopf, B., Burges, C. J., & Smola, A. J. (1999). *Advances in kernel methods: Support vector learning* MIT press, Cambridge.

Scholkopf, B., & Smola, A. (2002). Learning with kernels, MIT Press, Boston.

Shaowu, Z., Lianghong, W., Xiaofang, Y., & Wen, T. (2007). Parameters selection of SVM for function approximation based on differential evolution. *International Conference on Intelligent Systems and Knowledge Engineering 2007,*

She, Y., & Owen, A. B. (2011). Outlier detection using nonconvex penalized regression. *Journal of the American Statistical Association, 106*(494), pp 626-639.

Simpson, J. R. (1995). New methods and comparative evaluations for robust and biased-robust regression estimation, unpublished PhD thesis, Arizona State University.

Smets, K., Verdonk, B., & Jordaan, E. M. (2007). Evaluation of performance measures for SVR hyperparameter selection. *In IEEE 2007 International Joint Conference on Neural Networks,* Orlando, Florida, USA.

Smola, A., Murata, N., Schölkopf, B., & Muller, K. (1998). *Asymptotically optimal choice of ε-loss for support vector machines.* In the ICANN 98. Springer London, UK.

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing, 14*(3), 199-222.

Smola, A., & Vapnik, V. (1997). Support vector regression machines. *Advances in Neural Information Processing Systems, 9*, 155-161.

Smola, S. (1998). B.: A tutorial on support vector regression. NeuroCOLT technical. *C] Report NC-TR-98-030, Royal Holloway College, University of London, UK,*

Stamey, T. A., Kabalin, J. N., McNeal, J. E., Johnstone, I. M., Freiha, F., Redwine, E. A., & Yang, N. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. II. radical prostatectomy treated patients. *The Journal of Urology, 141*(5), 1076-1083.

Stromberg, A. J., Hössjer, O., & Hawkins, D. M. (2000). The least trimmed differences regression estimator and alternatives. *Journal of the American Statistical Association, 95*(451), 853-864.

Suykens, J. A. (2001). Nonlinear modelling and support vector machines. In the 18th IEEE *Instrumentation and Measurement Technology Conference (IMTC) 2001,* Budapest, Hungary.

Suykens, J. A., De Brabanter, J., Lukas, L., & Vandewalle, J. (2002). Weighted least squares support vector machines: Robustness and sparse approximation. *Neurocomputing, 48*(1), 85-105.

Tezcan, J., & Cheng, Q. (2012). Support vector regression for estimating earthquake response spectra. *Bulletin of Earthquake Engineering, 10*(4), 1205-1219.

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

Tipping, M. E. (2001). Sparse bayesian learning and the relevance vector machine. *The Journal of Machine Learning Research, 1*, 211-244.

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

Ukil, A. (2007). *Intelligent systems and signal processing in power engineering,* Springer, Heidelberg.

Üstün, B., Melssen, W., Oudenhuijzen, M., & Buydens, L. (2005). Determination of optimal support vector regression parameters by genetic algorithms and simplex optimization. *Analytica Chimica Acta, 544*(1), 292-305.

Üstün, B. (2003). A comparison of support vector machines and partial least squares regression on spectral data. *Department of Analytical Chemistry, Radboud University Nijmegen, Unpublished Master's Thesis,*

Üstün, B., Melssen, W. J., & Buydens, L. M. (2006). Facilitating the application of support vector regression by using a universal pearson VII function based kernel. *Chemometrics and Intelligent Laboratory Systems, 81*(1), 29-40.

Vanderbei, R. J. (1999). LOQO user's manual—version 3.10. *Optimization Methods and Software, 11*(1-4), 485-514.

Vapnik, V. (1995). *The nature of statistical learning theory*, 1st ed. Springer, New York.

Vapnik, V. (2000). *The nature of statistical learning theory*, 2nd ed. Springer, New York.

Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE Transactions on Neural Networks, 10*(5), pp 988-999.

Vapnik, V. N., & Vapnik, V. (1998). *Statistical learning theory* Wiley New York.

Vapnik, V., Golowich, S. E., & Smola, A. (1996). Support vector method for function approximation, regression estimation, and signal processing. *Advances in Neural Information Processing Systems, vol 9. MIT Press* , p 281-287.

Velleman, P. F., & Welsch, R. E. (1981). Efficient computing of regression diagnostics. *The American Statistician, 35*(4), 234-242.

Wang, W., Xu, Z., Lu, W., & Zhang, X. (2003). Determination of the spread parameter in the gaussian kernel for classification and regression. *Neurocomputing, 55*(3), 643-663.

Wang, X., Yang, C., Qin, B., & Gui, W. (2005). Parameter selection of support vector regression based on hybrid optimization algorithm and its application. *Journal of Control Theory and Applications, 3*(4), 371-376.

Weisberg, S. (2005). *Applied linear regression* John Wiley & Sons, Hoboken New Jersey.

Wilcox Rand, R. (2005). *Introduction to robust estimation and hypothesis testing*, Elsevier academic Press, New York.

Williams, G. (2011). *Data mining with rattle and R: The art of excavating data for knowledge discovery*, Springer, New York.

Williams, G., Hawkins, S., Gu, L., Baxter, R., & He, H. (2002). *A comparative study of RNN for outlier detection in data mining*. Data Mining, IEEE International Conference on IEEE Computer Society, Maebashi, Japan.

Wu, T. Z., Yu, K., & Yu, Y. (2010). Single-index quantile regression. *Journal of Multivariate Analysis, 101*(7), 1607-1621.

Yang, H., Huang, K., Chan, L., King, I., & Lyu, M. R. (2004). Outliers treatment in support vector regression for financial time series prediction. In the *11th International Conference* on *Neural Information Processing, ICONIP 2004, Calcutta, India*.

Yatchew, A. (2003). *Semiparametric regression for the applied econometrician* Cambridge University Press, Cambridge.

Yohai, V. J. (1987). *High breakdown-point and high efficiency robust estimates for regression.* The Annals of Statistics,15(2) , 642-656.

Yohai, V. J., & Zamar, R. H. (1988). High breakdown-point estimates of regression by means of the minimization of an efficient scale. *Journal of the American Statistical Association, 83*(402), 406-413.

Yu, C., Chen, K., & Yao, W. (2015). Outlier detection and robust mixture modeling using nonconvex penalized likelihood. *Journal of Statistical Planning and Inference, 164*, 27-38.

Zhou, X., & Ma, Y. (2013). A study on SMO algorithm for solving ϵ-SVR with non-PSD kernels. *Communications in Statistics-Simulation and Computation, 42*(10), 2175-2196.

Zhu, G., Liu, S., & Yu, J. (2002). Support vector machine and its applications to function approximation. *JOURNAL-EAST CHINA UNIVERSITY OF SCIENCE AND TECHNOLOGY, 28*(5), 555-559.

Zong, Q., Liu, W., & Dou, L. (2006). Parameters selection for SVR based on PSO. The Sixth World Congress on Intelligent Control and Automation, WCICA 2006, Dalian, China, pp 2811-2814.

Zou, H., & 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.