UNIVERSITI PUTRA MALAYSIA

ROBUST VARIABLE SELECTION METHODS FOR LARGE- SCALE DATA IN THE PRESENCE OF MULTICOLLINEARITY, AUTOCORRELATED ERRORS AND OUTLIERS

HASSAN S. URAIBI

IPM 2016 5

ROBUST VARIABLE SELECTION METHODS FOR LARGE- SCALE DATA IN THE PRESENCE OF MULTICOLLINEARITY, AUTOCORRELATED ERRORS AND OUTLIERS

By

HASSAN S. URAIBI

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

June 2016

DEDICATION

I would like to dedicate this dissertation work to

- *My respectful father and mother, who have taught me a lot on the meaning of persistency in life.*

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

- *My daughters and sons, Sura, Shahad, Iman, Fatima, Zainab, Adyian, Ali and Mohemmed, who were accompanying me in all different parts of my study and their love have always been my greatest inspiration.*

- *My dear friends Dr.Hakim Al-jibory, Dr.Ali Taqi , Dr.Zakariya Y. Algamal , Dr.Rahim Alhamzawi, Dr.Ali Jawad Alkinani and Dr.Tahir Reisan for their valuable support.*

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirement for the Degree of Doctor of Philosophy

ROBUST VARIABLE SELECTION METHODS FOR LARGE- SCALE DATA IN THE PRESENCE OF MULTICOLLINEARITY, AUTOCORRELATED ERRORS AND OUTLIERS

By

HASSAN S. URAIBI

June 2016

Chairman : Professor Habshah Midi, PhD
Faculty : Science

The robust correlation coefficient based on robust multivariate location and scatter matrix such as Fast Minimum Covariance Determinant (Fast MCD) is not feasible option for high dimensional data due to its time consuming procedure. To overcome this problem, robust adjusted Winsorization correlation (Adj.Winso.cor) is put forward. Unfortunately, the Adj.Winso.cor yields very poor results in the presence of multivariate outliers. Hence, we propose robust multivariate correlation matrix based on Reweighted Fast Consistent and High breakdown (RFCH) estimator. The findings show that the RFCH.cor is more robust than the Adj.Winso.cor in the presence of multivariate outliers.

Forward selection (FS) is very effective variable selection procedure for selecting a parsimonious subset of covariates from a large number of candidate covariates. However, FS is not robust to outliers. Robust forward selection method (FS.Winso) based on partial correlations which is derived from Maronna's bivariate M-estimator of scatter matrix and adjusted Winsorization pairwise correlation are introduced in a literatures to overcome the problem of outliers. We develop Robust Forward Selection algorithm based on RFCH correlation coefficient (RFS.RFCH) because FS.Winso is not robust to multivariate outliers. The results of our study indicate that the RFS.RFCH is more efficient than the FS and FS.Winso.

The existing Robust-LARS based on Winsorization correlation (RLARS-Winsor) has some drawbacks whereby it is not robust in the presence of multivariate outliers. Hence, Robust-LARS (RLARS-RFCH) based on $\sqrt{n}$ consistent multivariate (RFCH) correlation matrix is developed. The proposed method is computationally efficient and its performance outperformed the RLARS-Winsor

The algorithm of all possible subsets is greedy and it is inefficient and unstable in the presence of autocorrelated errors and outliers. To overcome the instability selection problem, a stability selection approach is put forward to enhance the performance of single-split variable selection method. Unfortunately, the classical stability selection procedure is very sensitive to outliers and serially correlated errors. The stability

procedure based on RFCH estimator is therefore developed. The results of the study show that our propose Robust Multi Split based on RFCH successfully and consistently select the correct variables in the final model.

Thus far, there is no variable selection procedure in literature that deal with the problem of high magnitude of multicollinearity in the presence of outliers. Hence, Robust Non-Grouped variable selection(RNGVS.RFCH) in the presence of high multicollinearity problem and outliers is developed. The results signify that our proposed RNGVS.RFCH method able to correctly select the important variables in the final model.

Not much research is focused on the problem of large data in the presence of outliers and autocorrelated errors. In this situation, the existing Elastic-Net and RE-Net methods are not capable of selecting the important variables in the final model. Thus, a new method that we call before and after elastic-net (BAE-Net) regression is proposed. The Reweighted Multivariate Normal (RMVN) algorithm is incorporated in the algorithm of the BAE-Net. The BAE-Net is found to do a credible job in selecting the correct important variables in the final model.

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

KAEDAH TEGUH PEMILIHAN PEMBOLEHUBAH BAGI DATA BERSKALA BESAR DENGAN KEHADIRAN MULTIKOLINEARAN, RALAT BERAUTOKORELASI DAN TITIK TERPENCIL

Oleh

HASSAN S. URAIBI

Jun 2016

Pengerusi : Profesor Habshah Midi, PhD
Fakulti : Sains

Pekali korelasi teguh berdasarkan lokasi multivariat teguh dan matrik serakan seperti Penentu Kovarians Minimum Pantas (*Fast-MCD*) tidak dapat dilaksanakan bagi data berdimensi tinggi disebabkan tatacaranya mengambil masa yang panjang. Untuk mengatasi masalah ini, korelasi *Winsorization Terlaras* teguh (*Adj.Winso.cor*) diketengahkan. Malangnya, *Adj.Winso.cor* memberikan keputusan yang lemah dengan kehadiran titik terpencil multivariat teguh. Oleh itu, kami mencadangkan matriks korelasi multivariat teguh berdasarkan Penganggar Berpemberat Konsisten Laju dan Titik Musnah Tinggi (*RFCH*). Hasil kajian menunjukkan bahawa *RFCH.cor* adalah lebih teguh daripada *Adj.Winso.cor* dengan kehadiran titik terpencil multivariat.

Pemilihan hadapan (FS) adalah tatacara pemilihan pembolehubah yang sangat berkesan bagi memilih subset kovariat parsimonius daripada sejumlah besar kovariat. Walaubagaimanapun FS tidak teguh terhadap titik terpencil. Kaedah pemilihan teguh hadapan (*FS.Winso*) berasaskan korelasi separa yang terhasil daripada serakan matrik penganggar-*M bivariate Maronna* dan korelasi *Winsorization* terlaras diperkenalkan dalam literature bagi mengatasi masalah titik terpencil. Kami bangunkan tatacara pemilihan teguh hadapan berasaskan pekali korelasi *RFCH (RFS.RFCH*) kerana *FS.Winso* tidak teguh terhadap titik terpencil multivariat. Keputusan kajian kami menunjukkan *RFS.RFCH* adalah lebih cekap berbanding FS dan *FS.Winso*.

Kaedah tersedia LARS-Teguh berasaskan korelasi *Winsorization* (RLARS-Winsor) mempunyai kelemahan dimana ianya tidak teguh dengan kehadiran titik terpencil multivariat. Oleh itu, LARS-Teguh (RLARS-RFCH) berasaskan matrik korelasi multivariat konsisten $\sqrt{n}$ (RFCH) dibangunkan. Kaedah yang dicadangkan berkomputasi efisien dan prestasinya menandingi *RLARS-Winsor*.

Tatacara semua kemungkinan subset adalah tamak dan tidak efisien dan tidak stabil dengan kehadiran ralat berautokorelasi dan titik terpencil. Untuk mengatasi masalah pemilihan yang tidak stabil, tatacara pemilihan stabil diketengahkan bagi meningkatkan prestasi kaedah pemilihan pembolehubah pecahan tunggal. Malangnya, kaedah pemilihan stabil klasik sangat peka terhadap titik terpencil dan siri ralat berkorelasi.

Oleh yang demikian, tatacara stabil berasaskan penganggar RFCH dibangunkan. Keputusan kajian menunjukkan bahawa kaedah Teguh Pecahan berganda yang kami bangunkan berasaskan RFCH berjaya dan secara konsisten memilih pembolehubah yang betul ke dalam model akhir.

Setakat ini, tiada tatacara pemilihan pembolehubah dalam literatur yang mengendalikan masalah multikolinearan paras tinggi dengan kehadiran titik terpencil. Oleh itu, pemilihan teguh pembolehubah tidak berkumpulan (*RNGVS.RFCH*) dengan kehadiran multikolinearan paras tinggi dan titik terpencil, di bangunkan. Keputusan kajian menunjukkan kaedah *RNGVS.RFCH* yang di cadangkan berupaya memilih dengan betul pembolehubah penting kedalam model akhir.

Tidak banyak penyelidikan menumpukan masalah data besar dengan kehadiran titik terpencil dan ralat berautokorelasi. Dalam keadaan ini, kaedah tersedia *Elastic-Net* dan *RE-Net* tidak berupaya memilih pembolehubah penting kedalam model akhir. Oleh itu, kaedah baru yang kami namakan regresi sebelum dan selepas elastic-net (BAE-Net) dicadangkan. Tatacara Multivariat Normal Berpemberat (RMVN) di gabungkan dalam tatacara BAE-Net. Kaedah BAE-Net didapati menunjukkan prestasi yang baik dalam memilih dengan betul pembolehubah penting kedalam model akhir.

ACKNOWLEDGEMENTS

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

Firstly, I would like to express my sincere gratitude to my supervisor Prof. Dr. Habshah Midi for the continuous support of my Ph.D study and related research, for her patience, motivation, and immense knowledge. Her guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better supervisor and mentor for my Ph.D study.

Besides my supervisor, I would like to thank the rest of my thesis committee: Dr. Md. Sohel Rana and Dr.Mohd Bakri Adam for their insightful comments and encouragement, but also for the hard question which inspired me to widen my research from various perspectives.

A special word of thanks to Professor Dr. Rahmatullah Imon, who is a professor of statistics from the Department of Mathematical Sciences, Ball State University, U.S.A, for his valuable time in sharing with me some of the insightful ideas during his visit to UPM campus.

My sincere thanks also goes to Prof. Dr. Stefan Van Aelst, Department of Applied Mathematics, Computer Science and Statistics,Ghent University, Belgium, Prof.Dr. Claudio Agostinelli, Dipartimento di Scienze Ambientali, Informatica e Statistica, Universita 'Ca' Foscari di Venezia,Venezia, Italy, and Dr. Fabian Schroeder, Glasergasse,Vienna, Austria who gave me an opportunity to access R code of their articles and technical reports which are interesting.

I would like to extend my gratitute to all the wonderful people such as Mohammed, Taha, Waleed, Shafie, Balqish and others. Their presence have indeed enriched my journey in completing my doctoral pursue. Appreciation also extended to all members of Graduate School and Faculty of Science, who have worked hard in creating a conducive environment for all post graduate students. I am glad to be able to graduate from this institution.

My sincere regards to Dr.Adel Alsaedy, Mussab Alshemary, Hassan Alshemary, Dhiah Aljiwary, Dr.Ali Jwa, Dr.Raheem Alhamzawi , Zakaryia Algamal , Dr.Tahir Reisan and all my siblings, who have continuously encourage me not to loose heart in all that I am pursuing, both mentally and also spiritually.

Last but not least, my special thanks go to my beloved wife, for standing by me patiently with her never ending encouragement, prayers and support throughout my doctoral pursue.

I certify that a Thesis Examination Committee has met on 14 June 2016 to conduct the final examination of Hassan S. Uraibi on his thesis entitled "Robust Variable Selection Methods for Large-Scale Data in The Presence of Multicollinearity, Autocorrelated Errors and Outliers" 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 b Abu Bakar, PhD
Associated Professor
Faculty of Sciences
Universiti Putra Malaysia
(Chairman)

Noor Akma bt Ibrahim, PhD
Professor
Faculty of Sciences
Universiti Putra Malaysia
(Internal Examiner)

Abdul Ghapor bin Hussin, PhD
Professor
Faculty of Science and Technology of Defense
Universiti Pertahanan Nasional Malaysia
(External Examiner)

A.H.M. Rahmatullah Imon, PhD
Professor
Ball State University
United States
(External Examiner)

ZULKARNAIN ZAINAL, PhD
Professor and Deputy Dean
School of Graduate Studies
Universiti Putra Malaysia

Date: 26 July 2016

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

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

Mohd Bakri Adam, PhD
Associate Professor
Faculty of Science
Universiti Putra Malaysia
(Member)

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

BUJANG 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 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 detection software.

Signature:________________________ Date:________________________

Name and Matric No.: Hassan S. Uraibi, GS 33238

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: Professor Dr. Habshah Midi

Signature: ________________________

Name of Member of Supervisory Committee: Associate Professor Mohd Bakri Adam

Signature: ________________________

Name of Member of Supervisory Committee: Dr. Md. Sohel Rana

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

LIST OF ABBREVIATIONS

AIC	Akaike Information Criterion
API	Air Pollution Index
AR	Autoregressive
BAE-Net	Before and After Elastic-Net
BIC	Bayesian Information Criterion
BP	Breakdown Point
Cp	Mallow's Crirerion
D.W	Durbin Watson
DGK	Robust Mahalanobis Distance based on the Minimum Volume Ellipsoid
DRGP	Diagnostic Robust Generalized Potential
FS	Forward Selection
FSA	Feasible Solution Algorithm
GLS	Generalized Least Square
IF	Influence Function
LARS	Least Angle Regression
Lasso	Least Absolute Shrinkage and Selection Operator
LP	Leverage Point
LTS	Least Trimmed Squares
MAD	Median Absolute Deviation
MADN	Normalized Median Absolute Deviation
MB	Median Ball
MCC	Maximizing the Contribution of Covariates
MCD	Minimum Covariance Determinant
MD	Mahalanobis Distance
MSE	Mean Square Errors
MVE	Minimum Volume Ellipsoid
OGK	Orthogonalized Gnanadesikan-Kettenring
OLS	Ordinary Least Squares
ORMSPE	Optimal RMSPE
RCOPW	Robust Cochrane-Orcutt Prais-Winsten Residuals
RCS	Robust Context-Sensitive
RE	Relative Effeicincy

RFCH	Reweighted Fast Consistent and High breakdown estimator
RFS	Robust Forward Selection
RLARS	Robust LARS
RMD	Robust Mahalanobis Distance
RMSPE	Root of Mean Square Prediction Errors
RMVN	Reweighted Multivariate Normal
RNGVS	Robust Non-Grouped Variable Selection
RPE	Robust Prediction Error
RSDE	Residual StanDard Error
SD	Standard Deviation
SEM	Standard Error Multiplier
SSR	Sum of Squares Residuals
V.O	Vertical Outlier
VIF	Variance Inflation Factor

CHAPTER 1

INTRODUCTION

1.1 Introduction and Background of the Study

The process of collecting large data has become an easy issue as a result of the fantastic growth in computer and networking technologies in the recent years. The collected data not only concerned the sample size, but also concerned the possibility of selecting large number of variables under study. This situation may give rise to a problem of curse dimensionality which forms a major challenge to variable selection researchers.

The curse of dimensionality refers to how certain algorithms such as algorithms in numerical analysis, sampling, combinatorics, machine learning, data mining and variable selection, 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 algoritms may become singular and some additional information such as regularization, Bayesian prior and others need to be added to obtain standard solution.

In the traditional statistical inference, the estimates of the population parameters can be substantially refined as the sample size increases toward infinity. A traditional requirement of estimators is consistency, that is, the convergence to the unknown true value of the parameter. High dimensional data is another setting of statistical problems, in which the dimension of variables p increases along with the sample size n so that the ratio p/n tends to a constant. It was called the “increasing dimension asymptotics” or “the Kolmogorov asymptotics (Aivasian et al., 1989). This procedure is allowing to analyze effects of inaccuracies accumulation in estimating a great number of parameters.

The curse of dimensionality is not a problem of high-dimensional data, but a combined problem of data and the algorithm being employed create a problem. It arises when the algorithm does not scale well to high-dimensional data, typically due to extensive amount of time or memory that is exponential in the number of dimensions of the data.

In the last decade, variable selection for high dimensional data has attracted much attention to researchers. High dimensional data can be classified into three cases, whereby the first case refers to the situation when the number of observations (n) is more than or equals to ten times the number of predictors (p), where $p \geq 10$. We call this case as large scale data in which the traditional approach of using Least

Squares (LS) is not appropriate due to time consuming. The second case of high dimensional data is when $p = n$. In this case, the algebric method is more suitable than the LS method. Finally, high dimensional data is also refers to a situation when $p > n$ in which the solution of the LS cannot be uniqe. Nonetheless, in this thesis we only focused for the case of large scale data.

The panelized methods which are introduced to overcome the problem of curse dimensionality, can also be used to analyze data when $p < n$, or $p \approx n$. This can be done because the line of LS is flexible and hence the penalty terms tend to reduce the overfitted problem (James et al., 2013). The problem becomes more complicated than the curse of dimensionality when outliers, multicollinearity and serial correlated residuals are present in the original data.

Khan et al. (2007a) pointed out that when the robust fit takes 0.001 cpu second, the all subsets regression need $2^d \times 0.001/(3600 \times 24 \times 365)$ years to select the final model where d is the number of candidate predictors. As a results of this new challenge, reducing the time of computation has become important target of modern variable selection methods.

The geometric interpretation of standardized data assists in introducing the concept of orthogonal design to variable selection methods, such that the cosine of specific angle equals to a value of regression coefficient, in which it is equivalent to the value of correlation between a covariate and a response variable. This concept has become indispensable in the modern variable selection method the last ten years, and it is considered faster than those that are based on original observations. Unfortunately, the classical and modern methods performed very poorly in the presence of outliers, multicollinearity and serialy correlated errors.

Multicollinearity problem may be present even though the magnitude of correlations between explanatory variables are small (Alley ,1987). The problem becomes more serious when the degree of correlation increases and resulting in a large standard error of regression coefficients. Consequently the t statistics become small which makes the regression coefficients not significant (Schroeder et al., 1986). Hence, more attention should be given in the field of data collection, to make decision, whether the selection method should be grouped or not. Determining the relevance of variable selection method rely on the interest of scientific field of applied science. Some researchers of chemical research ignore the highly correlated covariates, but this is not statistically proven because removing one covariate may affect the significant explanatory power of a model. On the other hand, research on gene expression considers grouped variable selection whereby the highly correlated covariates (genes) that share the same traits as one group. In this situation, whereby selecting one gene substantially needs select others (grouped variable selection).

In the traditional statistical approach when two covariates are correlated, one of the covariate should be dropped. The problem with this approach is to determine which one of the two covariates that need to be removed from the model. To deal with this problem, variable selection procedure should be considered because it has the ability to determine the important variable to be included in the final model.

1.2 Importance and Motivation of the Study.

It is well known that the Pearson's correlation is sensitive to outliers. As such it is important to use robust correlations as an altervative to the classical correlation. However, robust correlation creates problem of computational burden when its formulation is based on multivariate location and scatter matrix, such as Fast Minimum Covariance Determinant (FMCD) method of Rousseeuw and Van Driessen (1999). Khan et al. (2007a,b) pointed out that FMCD algorithm is not fast enough for any type of high dimensional data. Olive and Hawkins (2010) showed that FMCD is only outlier's diagnostic method since it is not known whether or not it is consistent. Hence, the construction of robust correlation based on FMCD gives rise to computational burden and it is infeasible option. Khan et al. (2007 b) proposed pairwise robust correlation which is called adjusted-Winsorized correlation estimate to solve the computational burden when bivariate outliers are present in a data. Unfortunately, this type of robust correlation is affected in the presence of multivariate outliers. Bivariate and high dimensional outliers refer to the existence of outliers in two variables /predictors and more than two predictors / variables, respectively.

This problem has motivated us to propose a new robust multivariate correlation matrix based on Reweighted Fast Consistent and High breakdown point (RFCH) location and dispersion estimator introduced by Olive and Hawkins (2010). To the best of our knowledge research on the RFCH correlations has not been considered in the literature. This is the first attempt to develop such robust correlation to overcome the problem of multivariate outliers and computational burden.

The Forward Selection (FS) is a commonly used method in variable selection. However, this method is very sensitive to the presence of outliers. To remedy this problem Khan et al. (2007a) and Khan et al. (2007b) developed robust forward selection based on adjusted winsorization (FS.Winso). They used Maronna's M estimate of the multivariate location and scatter matrix to formulate pairwise correlation. Subsequently, FS.Winso is developed. Unfortunately, such bivariate correlation is resistant only to bivariate outliers but not to multivariate outliers. However, outliers often exist in more than two variables (predictors). Moreover, FS.Winso is greedy algorithm due the original forward selection which is a greedy (Guyon and Elisseeff, 2003). In another word, the algorithm of FS.Winso does not consider all variables before making decision which variables to be included in the final model. This is due to the nature of the algorithm where it will stop when the next variable enters is not significant. The shortcomings of the FS.Winso has inspired us to develop new Robust Forward Selection based on $\sqrt{n}$ consistent

Reweighted Fast Consistent High breakdown estimator which is robust not only to bivariate but also to multivariate outliers.

This thesis also addresses another variable selection technique that deals with large number of covariates, using Least Angle Regression Selection (LARS) [see Efron et al. ,2004; Zou ,2006; Khan et al., 2007b; Agostinelli and Salibian-Barrera, 2010]. They noted that fitting all possible subsets and using stepwise selection procedure is not practical because it is very time consuming algorithm. Moreover, such methods suffer from correlated predictors. One solution to this problem is by employing LARS in the variable selection procedure. However, the classical LARS is very sensitive to the presence of outliers because it is based on classical correlation matrix.

Khan et al. (2007b) proposed robust LARS based on robust bivariate winsorization correlations. As already mentioned, this bivariate correlation is not resistant to multivariate outliers. This issue has encovrageed us to develop a robust LARS based on RFCH correlation matrix which is known to be $\sqrt{n}$ consistent estimator.

Splitting data into two parts is common in data analysis. Wasserman and Roeder (2009) proposed single-split data approach for variable selection. Nonetheless, this approach does not guarantee reproducible result due to arbitrarily splitting the data. In order to enhance the performance of single split variable selection, stability selection or multisplit approach is put forward (Meinshausen and Buhlman,2010; Shah and Samworth,2013). The weakness of this procedure is that, it is very sensitive to outliers. Additionally, this method cannot remedy the problem of serially correlated errors in a model. However, to the best of our knowledge, no research has been done to rectify the problem of outliers and serially correlated errors in multisplit variable selection approach. The gap in the literature regarding this issues has motivated us to take up the challenge to propose robust stability selection procedure for autocorrelated errors and in the presence of outlier.

Multicollinearity adds a new complication to variable selection technique especially when the degree of collinearity between variables is high (> 0.90). Mantel (1970) pointed out that Forward Selection (FS) technique failed to select important variables when collinearity problem is present in a data. Tibshirani (1996), Zou (2006) and Lin et al. (2012) noted that multicollinearity problem has an adverse effect on the variable selection procedure. Yang (2013) proposed Standard Error Adjusted Adaptive lasso (SE-lasso) and two stages model selection based on lasso (NSE-lasso) to rectify high collinearity among variables in variable selection technique. Unfortunately, these methods are very sensitive to the presence of outliers. The weakness of these methods has inspired us to develop robust variable selection procedure for extremely correlated variables in the presence of outliers. To the best of our knowledge, this is indeed the first attempt to overcome the problem of high correlated variables and the existence of outliers in variable selection procedure.

Huge and massive data form is a major challenge to statistics practitioners who utilize classical statistical methods because it is now evident that such methods do not perform well in massive setting. Massive data is often related to high dimentional data when the number of predictors is more than n. High dimensional data also usually refer as those data set with large number of predictors and large sample size. As the value of p increases, the computational burden of all subsets selection increases very quickly (Khan et al., 2007a). In this situation, many traditional variable selection techniques such as forward, backward and stepwise selection are computationaly intensive, unstable and time-consuming. Penalization methods such as LASSO (Tibshirani,1996), adaptive LASSO (Zou, 2006) , Elastic-Net (Zou and Hasti,2005), LARS (Efron et al. ,2004) and Dantzig Selector (Candes and Tao, 2007) are put forward as an alternative solution. Nonetheless, these methods are not robust when both outliers and autocorrelated errors exist in a data set. As a result of this, those methods are not sufficient enough to select important variables to a model and lead to bias estimate. Hence, It will select inaccurate number of variables to be included in a model. The shortcomings of those methods have inspired us to develop another variable selection method that is able to reduce the effect of outliers and correlated errors.

1.3 Research Objectives

The main objective of this thesis is to propose robust variable selection via concentrated data. The classical variable selection methods (traditional and modern) are based on LS estimates. Unfortunately, the LS estimate is not robust in the presence of outliers. The estimators of concentrated algorithms such as RFCH and RMVN are high breakdown and $\sqrt{n}$ consistent. With some modification to some existing variable selection procedures, the RFCH correlation matrix is formulated and incorporated in the forward selection, LARS, all subsets regression, adaptive lasso and Elasti Net to establish new improved variable selection methods. The foremost objectives of our research can be outlined systematically as follows.

1. To formulate a new robust correlations based on RFCH correlations matrix that is robust against multivariate outliers.
2. To develop a new robust forward selection method based on $\sqrt{n}$ consistent correlations matrix that is robust against multivariate outliers.
3. To formulate a new robust LARS method based on RFCH correlation matrix that can remedy the problem of multivariate outliers.
4. To develop a new robust stability selection procedure for autocorrelated errors in the presence of outliers.
5. To develop a new robust non-group variable selection procedure for extremely correlated variables and in the presence of outliers.
6. To develop robust Elastic NET variable selection procedure in the presence of serially correlated errors and in the presence of outliers.

1.4 Significance of The Study

Linear regression variable selection has many practical applications and it is an important issue for many areas of studies such as gene's expression, health, business, engineering, education, medicine and social science. In research studies, the statistics practitioners often obtained many independent variables, but they are not certain which variables are important to be included in the final model. In this situation, they may employ variable selection proceudre. There are a number of traditional variable selection procedures in the literatures, such as all possible subsets , stepwise regression and recently, the penalized methods such as lasso, adaptive lasso, Elastic net and Least Angle Regression. Unfortunately, the traditional methods fall short in one or more of the variable selection goals. For instance, all subsets may become impractical option for high dimensional data due to the expensive computational cost. Small change in data may result in large changes in a subset of predictors used, that is associated with the coefficients, predictions and so on.

Although, modern variable selection methods are put forward to overcome these deficiencies, many statistics practitioners are not aware of the fact that most of these methods are based on objective function which is sensitive to outliers, affected by multicollinearity and autocorrelation problems. The problems are further complicated for high dimensional or large scale dataset. This type of data may contain some fraction of outliers, highly correlated covariates, and other violations of LS assumptions. The robust variable selection procedures which are suggested in this thesis perform well in good and contaminated data. Their excellence performances are verified by the assessments done by Monte Carlo simulation study together with some real and artificial data.

This research also pointes out that the general framework of forward selection procedure can be very useful to overcome the problem of highly correlated variables based on the sequence of correlations. Therefore, the robust partial correlation for the scaled data is very crucial before any remedial action is taken.

A credible robust variable selection procedures are suggested in this thesis to enhance the performance of robust forward selection and Elastic net for autocorrelated errors. The RFCH and RMVN estimators perform excellently well in all types of outliers scenarios.

In this research, the RFCH estimator is used to construct a robust multivariate correlation and plug-in variable selection in terms of correlation. We use the concentrated data which are formed from the last step of RFCH algorithm to obtain robust regression estimates. Similar to the last procedure, we use RMVN estimator to eliminate the effect of outliers and Elastic net is computed with controlling procedure. A novel robust variable selection is offered in this thesis, when at least two independent variables are perfectly correlated. For all these discoveries, we expect there will be a good application for researchers in the future.

1.5 Scope and Limitation of the Study

Six objectives are studied in this thesis. We propose robust procedure which is connected with either plug-in of variable selection or concentrated data before applying our proposed method. The target of the first procedure is to robustify the forward selection, while the second target is to propose concentrated variable selection in the outset by reducing the effect of outliers in the original dataset and then applying the classical variable selection methods.

1.6 Outline of the Thesis

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

Chapter 2: This chapter presents a brief literature review of the OLS estimations of linear regression parameters and the violations from least squares assumptions. A review on variable selection problems are also discussed. Moreover, basic concepts of robust regression and some important existing robust regression methods are also highlighted. Diagnostic methods of outlying observations are also reviewed. Finally, stability selection and robust variable selections methods are discussed briefly.

Chapter 3: This chapter presents the robust correlations matrix. Two approaches of robust correlation are discussed. The first is the adjusted Winsorization correlation and the second is our proposed procedure that is based on RFCH estimator. The adjusted Winsorization correlation is not resistant to multivariate outliers. The advantages of using robust correlation matrix based on RFCH is supported by the evidence from the Montle Carlo simulation and modified real data.

Chapter 4: This chapter discusses the robust forward selection based on correlations. Both approaches of robust correlation, namely the adjusted Winsorization correlation and the RFCH correlations, are considered The forward selection based on adjusted Winsorization correlation is not resistant to multivariate outliers. The advantages of using forward selection based on RFCH is supported by the evidence from the Montle Carlo simulation and modified real data.

Chapter 5: In this chapter, we propose another variables selection method that is based on RFCH correlation matrix. In the robust literature, robust LARS is proposed in 2007, and is constructed based on adjusted Winsorization correlation. We incorporated the RFCH correlation matrix in the formulation of the new robust version of LARS. A study through Monte Carlo simulation and artificial dataset are done to support our conclusion that our proposed method, LARS.RFCH is more efficient than LARS.Winso. The univariate, bivariate and multivariate outliers cannot be visualized together in one set of real data. Hence, we contaminated the original artificial dataset three times to consider three outlier scenarios (univariate, bivariate and multivariate outlier).

Chapter 6: This chapter investigates robust stability selection procedure as a solution to the problem of variable selection in the presence of autocorrelated errors and outliers. The autocorrelation problem is first remedied and then employed the RFCH estimator to obtain data set without any outlying observation. Lastly, classical stability selection on the clean unautocorrelated data is employed to produce robust selection procedure. A study through Monte Carlo simulation and real Air quality data in Malaysia support the finding that in the presence of autocorrelated errors and outliers, our proposed robust stability selection is more efficient than the existing methods.

Chapter 7: In this chapter we propose robust variable selection procedure for exteremely correlated variables in the presence of outliers. We call this procedure robust variable selection for exteremely correlated variables, for ungrouped data. Similar to the previous chapter, the RFCH is employed to clean the data. The merit and the excellent performance of our proposed method is assessed by using Monte Carlo simulation experiments and artificial data.

Chapter 8: This chapter deals with an alternative method of robust variable selection procedure using Elastic net in the presence of autocorrelation and outliers. Unlike chapters 6 and 7, the RMVN estimator is used to clean the data. We propose adjusting the robust Elastic Net estimator to solve the overfitting problem. Similar to chapter 5, the problem of autocorrelation should be first be solved before running the algorithm. The performance of our proposed method is evaluated by using Monte Carlo simulation experiments and real datasets.

Chapter 9: This chapter presents the contributions, conclusions and recommendations for future studies.

REFERENCES

Agostinelli, C. and Salibian.M. (2010). Robust model selection with lars based on s-estimators *Proceedings of COMPSTAT'2010* (pp. 69-78): Springer.

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

Alfons, A., Baaske, W. E, Filzmoser, P., Mader, W., and Wieser, Roland. (2011). Robust variable selection with application to quality of life research. *Statistical Methods and Applications, 20*(1), 65-82.

Alfons, A., Croux, C. and Gelper, S. (2013). Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *The Annals of Applied Statistics, 7*(1), 226-248.

Alkenani, A. and Yu, K. (2013). A comparative study for robust canonical correlation methods. *Journal of Statistical Computation and Simulation, 83*(4), 692-720.

Allen, D. M. (1974). The relationship between variable selection and data agumentation and a method for prediction. *Technometrics, 16*(1), 125-127.

Alley, W. M. (1987). A note on stagewise regression. *The American Statistician, 41*(2), 132-134.

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

Altman, N. and Leger, C. (1997). On the optimality of prediction-based selection criteria and the convergence rates of estimators. *Journal of the Royal Statistical Society. Series B (Methodological)*, 205-216.

Andersen, R. (Ed.) (2008). *Modern methods for robust regression*. Thousand Oaks, CA:Sage Publication, Inc.

Anderson, C. and 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.

Ann, L. H. and Midi, H.. (2011). *The effect of high leverage points on the robust autocorrelation test in multiple linear regression.* Paper presented at the Proceedings of the 11th WSEAS international conference on Applied computer science.

Armstrong, R. D. and Kung, MT. (1981). An algorithm to select the best subset for a least absolute value regression problem: DTIC Document.

Aivasian S. A., Buchstaber V. M., Yenyukov I. S., Meshalkin L. D. (1989). *Applied Statistics. Classification and Reduction of Dimensionality*. Moscow, Russian.

Atkinson, AC. (1982). Regression diagnostics, transformations and constructed variables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1-36.

Baccini, M., Biggeri, A., Grillo, P., Consonni, D., and Bertazzi, P. A. (2011). Health impact assessment of fine particle pollution at the regional level. *American journal of epidemiology*, kwr256.

Barnett, V. and Lewis, T. (1994). *Outliers in statistical data* (Vol. 3): Wiley New York.

Belsley, D. A., Kuh, E. and Welsch, RE. (1980). Regression DiagnosticsWiley. *New York*.

Belsley, D. A., Kuh, E. and Welsch, RE. (1980). Recession Diagnostics: John Wiley and Sons, New York.

Burnham, K. P. and Anderson, D. R. (2002). *Model selection and multimodel inference: a practical information-theoretic approach*: Springer Science and Business Media.

Burnham, K. P. and Anderson, D. R. (2004). Multimodel inference understanding AIC and BIC in model selection. *Sociological methods and research*, *33*(2), 261-304.

Burns, P. J. (1992). A genetic algorithm for robust regression estimation. *StatScience Technical Note*.

Candes, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, 2313-2351.

Chatfield, C. (1995). *Problem solving: a statistician's guide*: CRC Press.

Chatterjee, S, and Price, B. (1977). Selection of variables in a regression equation. *Regression Analysis by Example*, 201-203.

Chatterjee, S. and Hadi, A. (2015). *Regression analysis by example*: John Wiley and Sons.

Claeskens, G. and Hjort, N. L. (2008). *Model selection and model averaging* (Vol. 330): Cambridge University Press Cambridge.

Cochrane, D. and Orcutt, G. H. (1949). Application of least squares regression to relationships containing auto-correlated error terms. *Journal of the American Statistical Association, 44*(245), 32-61.

Cohen, J., Cohen, P., West, S.G. and Aiken, L. S. (2013). *Applied multiple regression/correlation analysis for the behavioral sciences*: Routledge.

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

Cook, R D. and Hawkins, D. M. (1990). Unmasking multivariate outliers and leverage points: comment. *Journal of the American Statistical Association*, 640-644.

Cook, R D., and Weisberg, S. (1982). Residuals and influence in regression. Chapman and Hall, New York — London

Cook, R D, Hawkins, D. M. and Weisberg, S. (1993). Exact iterative computation of the robust multivariate minimum volume ellipsoid estimator. *Statistics and probability letters, 16*(3), 213-218.

Corani, G. (2005). Air quality prediction in Milan: feed-forward neural networks, pruned neural networks and lazy learning. *Ecological Modelling, 185*(2), 513-529.

Croux, C., Haesbroeck, G., and Rousseeuw, P. J. (2002). Location adjustment for the minimum volume ellipsoid estimator. *Statistics and Computing, 12*(3), 191-200.

Croux, C. and Ruiz, G.Anne. (1996). *A fast algorithm for robust principal components based on projection pursuit.* Paper presented at the Compstat.

Devlin, S. J, Gnanadesikan, R. and Kettenring, J. R. (1981). Robust estimation of dispersion matrices and principal components. *Journal of the American Statistical Association, 76*(374), 354-362.

Dixon, W. J. (1950). Analysis of extreme values. *The Annals of Mathematical Statistics*, 488-506.

Douglas, C. M., Elizabeth, A. P. and Geoffrey, V. (1992). Introduction to linear regression analysis: Wiley Inter. Science," New York, NY.

Draper, N. R., and Smith, H. (1981). Applied regression analysis, 709 pp: John Wiley, New York.

Edgeworth, F. Y. (1887). On observations relating to several quantities. *Hermathena, 6*(13), 279-285.

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

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.

Firat Ö., A, and Wilcox, R. (2012). New results on the small-sample properties of some robust univariate estimators of location. *Communications in Statistics-Simulation and Computation, 41*(9), 1544-1556.

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.

Fox, J. (2002). Robust regression. *An R and S-Plus companion to applied regression*. Sage Publications, Inc. Thousand Oaks, CA, USA

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

Freund, Y., and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, *55*(1), 119-139.

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

Gnanadesikan, R. and Kettenring, J. R. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, 81-124.

Godish, T. and Fu, J. S. (2003). *Air quality*: CRC Press.

Greene, W. H. (2003). *Econometric analysis*: Pearson Education India.

Gujarati, D. N. and Porter, D. (2009). Basic Econometrics Mc Graw-Hill International Edition.

Habshah, M, Norazan, M. R., and Imon, AHMR. (2009). The performance of diagnostic-robust generalized potentials for the identification of multiple high leverage points in linear regression. *Journal of Applied Statistics, 36*(5), 507-520.

Hampel, F. R., Ronchetti, E. M, Rousseeuw, P. J. and Stahel, W. A. (1986). Robust statistics, J. *Wileyand Sons, New York*.

Hao, Y. (1992). *Maximum Median Likelihood and Maximum Trimmed Likelihood Estimations.* (Ph. D), Ph. D Thesis, University of Toronto, Toronto.

Hastie, T., Tibshirani, R., Friedman, J. and Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer, 27*(2), 83-85.

Hawkins, D. M. (1980). *Identification of outliers* (Vol. 11): Springer.

Hawkins, D. M. and Olive, D. J. (1999). Improved feasible solution algorithms for high breakdown estimation. *Computational statistics and data analysis, 30*(1), 1-11.

Hawkins, D. M. and Olive, D. J. (2002). Inconsistency of resampling algorithms for high-breakdown regression estimators and a new algorithm. *Journal of the American Statistical Association, 97*(457), 136-159.

Heritier, S., Cantoni, E., Copt, S. and Victoria-Feser, M. (2009). *Robust methods in Biostatistics* (Vol. 825): John Wiley and Sons.

Hesterberg, T., Choi, N. H., Meier, L, and Fraley, C. (2008). Least angle and ℓ1 penalized regression: A review. *Statistics Surveys, 2*, 61-93.

Hoaglin, D. C. and Welsch, RE. (1978). The hat matrix in regression and ANOVA. *The American Statistician, 32*(1), 17-22.

Hocking, R. and Pendleton, O. (1983). The regression dilemma. *Communications in Statistics-Theory and Methods, 12*(5), 497-527.

Hoerl, A., and Kennard, R. (1988). Ridge regression. *Encyclopedia of statistical sciences*.

Huber, P. J. (1981). *Robust Statistics*. New York: John Wiley and Sons, Inc.

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

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

Huber, P. J.and Ronchetti, EM. (1981). Robust Statistics, ser. *Wiley Series in Probability and Mathematical Statistics. New York, NY, USA: Wiley-IEEE, 52*, 54.

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

Hurvich, C. M. and Tsai, C. (1989). Regression and time series model selection in small samples. *Biometrika, 76*(2), 297-307.

Imon, AHMR. (2005). A stepwise procedure for the identification of multiple outliers and high leverage points in linear regression. *PAKISTAN JOURNAL OF STATISTICS-ALL SERIES-, 21*(1), 71.

Izenman, A. J. (2008). *Modern multivariate statistical techniques* (Vol. 1): Springer.

Khan, J. A., Van Aelst, S. and Zamar, R. H. (2007a). Building a robust linear model with forward selection and stepwise procedures. *Computational Statistics and Data Analysis, 52*(1), 239-248.

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

Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *Annals of statistics*, 1356-1378.

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

Kuha, J. (2004). AIC and BIC comparisons of assumptions and performance. *Sociological Methods and Research*, *33*(2), 188-229.

Kutner, M. H, Nachtsheim, C. and Neter, J. (2004). *Applied linear regression models*: McGraw-Hill/Irwin.

Kutner, M. H., Nachtsheim, C. J, Neter, J. and Li, W. (2005). Applied linear statistical models.

Lane, L. J, and Dietrich, D. L. (1976). Bias of selected coefficients in stepwise regression. *US Agric Res Serv Reprints of articles by ARS employees*.

Leng, C., Lin, Y. and Wahba, G. (2006). A note on the lasso and related procedures in model selection. *Statistica Sinica*, *16*(4), 1273.

Leroy, A. M, and Rousseeuw, P. J. (1987). Robust regression and outlier detection. *Wiley Series in Probability and Mathematical Statistics, New York: Wiley, 1987, 1*.

Lin, D., Foster, D. P. and Ungar, L. H. (2012). VIF regression: a fast regression algorithm for large data. *Journal of the American Statistical Association*.

Lopuhaa, H. P. (1999). Asymptotics of reweighted estimators of multivariate location and scatter. *Annals of Statistics*, 1638-1665.

Lopuhaa, H. P. and Rousseeuw, P. J. (1991). Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *The Annals of Statistics*, 229-248.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta), 2*, 49-55.

Mallows, C. L. (1973). Some comments on C p. *Technometrics, 15*(4), 661-675.

Mantel, N. (1970). Why stepdown procedures in variable selection. *Technometrics, 12*(3), 621-625.

Maronna, R., Martin, D. and Yohai, V. (2006). *Robust statistics*: John Wiley and Sons, Chichester. ISBN.

Maronna, R. A. and Zamar, R. H. (2002). Robust estimates of location and dispersion for high-dimensional datasets. *Technometrics, 44*(4).

Maronna, R. A., martin, R. D., and Yohai, V. J. (2006). *Robust Statistics: Theory and methods*: John Wiley and Sons, Ltd.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 1436-1462.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 72*(4), 417-473.

Meintanis, S. G, and Donatos, G. S (1997). A Comparative Study of Some Robust Methods For Coefficient-Estimation In Linear Regression. *Computational Statistics and Data Analysis, 23*, 525-540.

Mendenhall, W., Sincich, T. and Boudreau, N. S. (1996). *A second course in statistics: regression analysis* (Vol. 5): Prentice Hall Upper Saddle River^ eNew Jersey New Jersey.

Midi, H. (1999). Preliminary estimators for robust non-linear regression estimation. *Journal of Applied Statistics, 26*(5), 591-600.

Midi, H. and Mohammed, M. A. (2015), The Identification of Good and Bad High Leverage Points in Multiple Linear Regression Model. Mathematical Methods and Systems in Science and Engineering Proceeding, Spain.

Miller, A. (2002). *Subset selection in regression*: CRC Press.

Montgomery, D. C, Peck, E. A, and Vining, G. G. (2012). *Introduction to linear regression analysis* (Vol. 821): John Wiley and Sons.

Morrison, D F. (1983). *Applied linear statistical methods*: Prentice-Hall Englewood Cliffs, NJ.

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

Olive, D. J. (2004). A resistant estimator of multivariate location and dispersion. *Computational statistics and data analysis, 46*(1), 93-102.

Olive, D. J, and Hawkins, D. M. (2008). High breakdown multivariate estimators. *Preprint, see (www. math. siu. edu/olive/preprints. htm).*

Olive, D. J, and Hawkins, D. M. (2010). Robust multivariate location and dispersion. *Preprint, see (www. math. siu. edu/olive/preprints. htm)*.

Pearson, K. and Lee, A. (1908). On the generalised probable error in multiple normal correlation. *Biometrika, 6*(1), 59-68.

Peña, D., and Yohai, V. J. (1995). The detection of influential subsets in linear regression by using an influence matrix. *Journal of the Royal Statistical Society. Series B (Methodological)*, 145-156.

ires, J., Martins, F., Sousa, S., Alvim-Ferraz, M. and Pereira, MC. (2008). Selection and validation of parameters in multiple linear and principal component regressions. *Environmental Modelling and Software, 23*(1), 50-55.

Rana, S., Midi, H. and Imon, AHMR. (2008). A robust modification of the goldfeld-quandt test for the detection of heteroscedasticity in the presence of outliers. *Journal of Mathematics and Statistics, 4*(4), 277-283.

Riazoshams, H., Midi, H. and Sharipov, O. (2010). The performance of robust two-stage estimator in nonlinear regression with autocorrelated error. *Communications in Statistics-Simulation and Computation, 39*(6), 1251-1268.

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

Ronchetti, E., Field, C., and Blanchard, W. (1997). Robust linear model selection by cross-validation. *Journal of the American Statistical Association, 92*(439), 1017-1023.

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

Rousseeuw, P.J., and Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. New York: Wiley.

Rousseeuw, P.J., and Yohai, V. J. (1984). Robust regression by means of *S*-Estimators *Robust and Nonlinear Time Series Analysis, Lecture Notes in Statistics*. Heidelberg, Germany: Springer-Verlag.

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

Rousseeuw, P. J, and Driessen, K. V. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics, 41*(3), 212-223.

Rousseeuw, P. J, and Leroy, A. M. (2005). *Robust regression and outlier detection* (Vol. 589): John Wiley and Sons.

Rousseeuw, P. J, and Van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, *85*(411), 633-639.

Ryan, T P. (1997). *Modern regression models*: New York: Wiley.

Saithanu, K, and Mekparyup, J. (2014). Using Multiple Linear Regression To Predict PM10 Concentration In Chonburi, Thailand. *Global Journal of Pure and Applied Mathematics*, *10*(6), 835-839.

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

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461-464.

Seber, G. A. (1977). *Linear regression analysis*, J: Wiley, New York.

Sedek, J. N. M., Ramli, N. A. and Yahaya, A. S. (2006). Air quality predictions using log normal distribution functions of particulate matter in Kuala Lumpur. *Malaysian Journal of Environmental Management, 7*, 33-41.

Shah, R. D, and Samworth, R. J. (2013). Variable selection with error control: another look at stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 75*(1), 55-80.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica, 7*(2), 221-242.

Shevlyakov, G., and Smirnov, P. (2011). Robust estimation of the correlation coefficient: An attempt of survey. Austrian Journal of Statistics, 40(1and2), 147-156.

Sommer, S. and Staudte, R. G. (1995). Robust variable selection in regression in the presence of outliers and leverage points. *Australian Journal of Statistics, 37*(3), 323-336.

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.

Tam, B. N. and Neumann, C. M. (2004). A human health assessment of hazardous air pollutants in Portland, *Journal of Environmental Management, 73*(2), 131-145.

Tharmaratnam, K. and Claeskens, G. (2011). S-estimation and a robust conditional Akaike information criterion for linear mixed models. *Available at SSRN 1974883*.

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

Tukey, J. W. (1960). A survey of sampling from contaminated distributions. In I. Olkin 156 *Austrian Journal of Statistics*, Vol. 40 (2011), No. 1 and 2, 147–156

Ul-Saufie, A. Z., Yahaya, A. S., Ramli, N. A. and Hamid, H. A. (2012a). Robust regression models for predicting PM10 concentration in an industrial area. *International Journal of Engineering and Technology, 2*(3), 364-370.

Ul-Saufie, A. Z., Yahaya, A. S., Ramli, N. A., and Hamid, H. A. (2012b). Performance of Multiple Linear Regression Model for Long-term PM^ sub 10^ Concentration Prediction Based on Gaseous and Meteorological Parameters. *Journal of Applied Sciences*, *12*(14), 1488.

Uraibi, H. S, Midi, H., A Talib, B. and Yousif, J. H. (2009). Linear regression model selection based on robust bootstrapping technique. *American Journal of Applied Sciences*, *6*(6), 1191-1198.

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

Weisberg, S. (1980). Applied Linear Regression, New York: JohnWiley.

West, D. B. (2001). *Introduction to graph theory* (Vol. 2): Prentice hall, Upper Saddle River.

Woodruff, D. L. and Rocke, D. M. (1994). Computable robust estimation of multivariate location and shape in high dimension using compound estimators. *Journal of the American Statistical Association*, *89*(427), 888-896.

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

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

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68*(1), 49-67.

Zhang, J., Olive, D. J. and Ye, P. (2012). Robust covariance matrix estimation with canonical correlation analysis. *International Journal of Statistics and Probability*, *1*(2), p119.

Zhang, P. (1992). Inference after variable selection in linear regression models. *Biometrika*, *79*(4), 741-746.

Zhao, P, and Yu, B. (2006). On model selection consistency of Lasso. *The Journal of Machine Learning Research*, *7*, 2541-2563.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association, 101*(476), 1418-1429.

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.