UNIVERSITI PUTRA MALAYSIA

ROBUST DIAGNOSTICS AND VARIABLE SELECTION PROCEDURE BASED ON MODIFIED REWEIGHTED FAST CONSISTENT AND HIGH BREAKDOWN ESTIMATOR FOR HIGH DIMENSIONAL DATA

ISHAQ ABDULLAHI BABA

IPM 2022 4

ROBUST DIAGNOSTICS AND VARIABLE SELECTION PROCEDURE BASED ON MODIFIED REWEIGHTED FAST CONSISTENT AND HIGH BREAKDOWN ESTIMATOR FOR HIGH DIMENSIONAL DATA

By

ISHAQ ABDULLAHI BABA

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

January 2022

DEDICATION

This thesis is dedicated to my wife, Hauwau Babayo, and my children, Abdullahi Ishaq Baba, Abdurahman Ishaq Baba, and Mohammad Ishaq Baba.

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

ROBUST DIAGNOSTICS AND VARIABLE SELECTION PROCEDURE BASED ON MODIFIED REWEIGHTED FAST CONSISTENT AND HIGH BREAKDOWN ESTIMATOR FOR HIGH DIMENSIONAL DATA

By

ISHAQ ABDULLAHI BABA

January 2022

Chairman: Prof. Habshah Binti Midi, PhD
Faculty: Science

The reweighted fast, consistent and high breakdown (RFCH) estimator is a multivariate procedure used to estimate the robust location and scatter matrix. It is incorporated in the robust Mahalanobis distance to detect the presence of high leverage points in a dataset. The method showed excellent performance compared to its competitors. However, it cannot be applied when the sample size is less than the number of predictor variables. In addressing this problem, some robust procedures for high dimensional dataset via the RFCH algorithm are developed.

A modified reweighted fast consistent and high breakdown (MRFCH) estimator in high dimensional data based on the diagonal elements of the scatter matrix instead of its entire elements in the computation of robust Mahalanobis distance within the RFCH algorithm is developed. The proposed method inherits the robustness properties of the original RFCH estimators. Simulation results and artificial data examples showed that the proposed MRFCH is more efficient and faster than the MRCD and OGK estimators.

Outlier detection and classification are critical issues that affect prediction accuracy if not handled correctly. Mahalanobis distance (MD) measure is one of the most popular multivariate analysis tools used to detect multivariate outlying observations. However, the traditional MD based on the classical mean and covariance rarely identifies all the multivariate outliers in a given dataset, which gives rise to the masking and swamping problems. Therefore, the robust location and covariance matrix based on the MRFCH is used instead of the classical estimators to tackle these problems. The proposed algorithm has been applied to detect outliers in the high dimensional

data. The results obtained from the simulation study and real data sets indicate that the proposed method possesses high detection power with minimal misclassification error compared to the MRCD and MDP methods.

The classical correlation estimators that employ the sample mean of the dependent and independent variables are known to be affected by outliers. Therefore, the robust weighted correlation coefficient that can reduce the effect of outliers is proposed. The weights based on the RD (MRFCH) are incorporated in establishing the proposed robust correlation to solve the problems. The performance of the proposed method is illustrated using simulation study and on glass vessel data with 1920 variables, cardiomyopathy microarray data with 6319 variables, and octane data with 226 dimensions. The results show that the robust weighted correlation based on RD (MRFCH) is more powerful and efficient than the existing methods, irrespective of dimension, sample size, and contamination levels.

Sure screening-based correlation methods are popular tools used to select the most significant variables in the true model in sparse and high dimensional analysis. However, in practice, high leverage points may lead to misleading results in solving variable selection problems. Therefore, a robust sure independence screening procedure based on the weighted correlation algorithm of MRFCH for high dimensional data is developed to address this problem. The simulation study results and real data sets indicate that the proposed MRFCHCS+LAD-SCAD estimator was found to be the best method compared to other methods in this study.

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

DIAGNOSTIK TEGUH DAN PROSEDUR PEMILIHAN PEMBOLEHUBAH BERDASARKAN PENGANGGAR TERUBAHSUAI BERPEMBERAT YANG PANTAS, KONSISTEN DAN TITIK MUSNAH YANG TINGGI BAGI DATA BERDIMENSI TINGGI

Oleh

ISHAQ ABDULLAHI BABA

Januari 2022

Pengerusi: Prof. Habshah Binti Midi, PhD
Fakulti: Sains

Penganggar berpemberat yang pantas, konsisten dan titik musnah yang tinggi (RFCH) adalah prosedur multivariat yang digunakan untuk menganggar lokasi teguh dan matriks penyebaran. Ia telah digabungkan dalam jarak Mahalanobis teguh bagi mengesan kehadiran titik tuasan tinggi dalam set data. Kaedah ini telah menunjukkan prestasi cemerlang apabila dibandingkan dengan pesaingnya. Namun, ia tidak boleh diguna-pakai apabila saiz sampel adalah lebih kecil daripada pembolehubah peramal. Bagi menangani masalah ini, beberapa prosedur teguh untuk set data berdimensi tinggi melalui RFCH telah dibangunkan.

Penganggar terubahsuai berpemberat yang pantas, konsisten dan titik musnah yang tinggi (MRFCH) bagi data berdimensi tinggi dengan menggunakan unsur pepenjuru dari matriks kovarians dan bukannya unsur keseluruhannya dalam pengiraan jarak Mahalanobis teguh, telah dibangunkan. Kaedah yang dicadan- gkan mewarisi sifat keteguhan penganggar RFCH asal. Hasil simulasi dan contoh data buatan menunjukkan bahawa kaedah MRFCH yang dicadangkan ada-lah lebih cekap dan lebih pantas daripada penganggar MRCD dan OGK.

Pengesanan titik ter- pencil dan pengkelasan adalah masalah kritikal yang menjejaskan ketepatan ramalan jika ia tidak dikendalikan dengan betul. Ukuran jarak Mahalanobis (MD) adalah salah satu alat analisis multivariat yang paling popular digunakan untuk mengesan cerapan terpencil multivariat. Walau bagaimanapun, MD tradisional berdasarkan min klasik dan kovarians jarang mengenal pasti semua titik terpencil multivariat dalam set data tertentu, yang menimbulkan masalah penyamaran dan penukaran. Oleh itu, lokasi teguh dan matriks kovarians teguh berdasarkan MR-FCH digunakan sebagai gantian penganggar klasik untuk mengatasi masalah tersebut. Kaedah yang dicadangkan telah digunakan untuk mengesan titik terpencil bagi data berdi- mensi tinggi. Hasil kajian simulasi dan set data sebenar menunjukkan kaedah yang dicadangkan mempunyai kuasa pengesanan yang tinggi dengan ralat silapklasifikasi yang minimum berbanding dengan kaedah MRCD dan MDP.

Penganggar korelasi klasik yang menggunakan min sampel pembolehubah bersandar dan pembolehubah bebas adalah diketahui boleh dipengaruhi oleh titik terpencil. Oleh itu, pekali korelasi berpemberat teguh yang dapat mengurangkan kesan titik terpencil telah dicadangkan. Pemberat berdasarkan RD (MRFCH) telah digabungkan untuk membangunkan korelasi teguh bagi menyelesaikan masalah tersebut. Prestasi kaedah yang dicadangkan dipamerkan menggunakan kajian simu- lasi dan data kapal kaca dengan 1920 pembolehubah, data mikroarray kardiomiopati dengan 6319 pembolehubah, dan data oktan berdimensi 226. Hasil kajian me- nunjukkan bahawa korelasi berpemberat teguh berdasarkan RD (MRFCH) adalah lebih berkuasa dan efisien daripada kaedah yang sedia ada tanpa mengira dimensi dan tahap pencemaran.

Kaedah korelasi berdasarkan saringan pasti adalah alat popu- lar yang digunakan untuk memilih pembolehubah yang paling penting untuk di- masukkan ke dalam model sebenar dalam analisis berdimensi jarang dan tinggi. Walau bagaimanapun, secara praktik, titik tuasan tinggi boleh menghasilkan keputusan yang mengelirukan semasa menyelesaikan masalah pemilihan pembole- hubah. Oleh itu prosedur penyaringan kebebasan yang pasti berdasarkan algoritma korelasi berpemberat MRFCH bagi data berdimensi tinggi telah dibangunkan untuk men- gatasi masalah tersebut. Hasil kajian simulasi dan set data sebenar menunjukkan ba- hawa penganggar MR-FCHCS + LAD-SCAD yang dicadangkan didapati sebagai kaedah terbaik apabila dibandingkan dengan kaedah lain dalam kajian ini.

ACKNOWLEDGEMENTS

All thanks and praise be to Allah, who bless us with knowledge, wisdom, and guidance to see the success of this research work. May his blessing and peace be upon our Prophet Muhammad (SAW), his family, and companions.

I want to express my profound gratitude to my supervisor and chairman supervisory committee, Professor Dr. Habshah Bt Midi, for all the valuable, constructive comments, educative discussion, support, encouragement, and guidance during the cause of this study. Despite her tide schedules, I am impressed by her patience, cheerfulness, and sheer professionalism. I pray for Allah to continue guiding and protecting her and making her knowledge more beneficial to all humanity.

My numerous appreciations go to the members of my supervisory committee, Prof Dr. Leong Wah June and Associate Professor Dr. Ibragimov Gafurjan. Most of the time, they listen to me whenever I approach them with any academic or nonacademic issue. May Allah reward them abundantly and continue increasing their knowledge and making it beneficial to the entire world.

A similar appreciation goes to the Director Institute for Mathematical Research, all academics, nonacademic staff, and all students of Universiti Putra Malaysia for their support and contribution. I thank all of them for the excellent job they are doing.

I am incredibly grateful to all my family members, especially my wife, Hauwau Bababyo, and our children Abdullahi, Abdurrahman, and Mohammad, for their patience and understanding throughout my stay in Malaysia. May Allah continue to guide, protect and reward them all.

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 the Supervisory Committee were as follows:

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

Leong Wah June, PhD
Professor
Faculty of Science
Universiti Putra Malaysia
(Member)

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

ZALILAH MOHD SHARIFF, PhD
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date: 19 May 2022

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 Binti Midi

Signature: ______________________________
Name of
Member of
Supervisory
Committee: Professor Dr. Leong Wah June

Signature: ______________________________
Name of
Member of
Supervisory
Committee: Associate Professor Dr. Ibragimov Gafurjan

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CC	Canonical Correlation
DC-SIS	Distance-Based Correlation Screening
DetMCD	Deterministic Minimum Covariance Determinant Algorithm
GM	Generalized M-Estimators
DRGP	Diagnostic Robust Generalized Potentials
FMCD	Fast Minimum Covariance Determinant
FN	False Negatives
FP	False Positives
GCD	Generalized Cook Distance
HLP	High Leverage Points
ISIS	Iteratively Sure Independence Screening
LASSO	Least Absolute Shrinkage and Selection Operator
LS-SCAD	Least Squares Smoothly Clipped Absolute Deviation
LTS	Least Trimmed Squares
MAD	Median Absolute Deviation
MB	Median Ball
MCD	Minimum Covariance Determinant
MD	Mahalanobis Distance
MDP	Minimum Diagonal Product
MDP	Minimum Diagonal Product
MMSE	Median Mean Square Error
MRCD	Minimum Regularized Covariance Determinant

MRFCH	Modified Reweighted Fast Consistent and High Breakdown Point
MSE	Mean Square Error
MVE	Minimum Volume Ellipsoid
NIR	Near-Infrared Spectroscopy
OGK	Orthogonal Gnanadesikan Kettenring
PLS	Penalized Least Squares
PPMC	Pearson Product-Moment Correlation
QWAS	Genome Wide Association Studies
RD	Robust Distance
RFCH	Reweighted Fast Consistent and High Breakdown
ROC	Receiver Operating Characteristic
ROE	Return On Equity
RPCA	Robust Principal Component Analysis
RRCS	Robust Rank Correlation Screening
RRC-SIS	Robust Rank Correlation Screening
SCAD	Smoothly Clipped Absolute Deviation
SIS	Sure Independent Screening
WLAD	Weighted Least Absolute Deviation

CHAPTER 1

INTRODUCTION

1.1 Background of the Study

Recent innovations in science and technology have made data collection and processing an attractive topic in scientific research and industrial applications. Some familiar data sources include social media platforms, health, educational, financial, and economic sectors, to name but a few. The main concern of data analysts is the number of data points relative to the number of variables under consideration and vice versa. In practice, multivariate data appear more frequently than univariate data because most experiments pay much attention to observations' features in many cases. Thus, investigating the relationship between the dependent and independent variables is paramount in solving most real-life problems involving multivariate data.

Multivariate location and dispersion estimates have imperative uses in theoretical and applied statistical analysis. However, in the presence of outlying data points, classical estimates of mean and covariance matrices are not trustworthy. It is clear now that even a single outlier can distort the classical mean and covariance estimates, making them practically inadequate, affecting or corrupting the estimate of correlations, principal component transformations, and multivariate outlier detection based on the Mahalanobis distances.

Mahalanobis distance (MD) is one of the widely multivariate statistical tools used to measure the distances between two points with several variables. More precisely, it is widely applied to detect multivariate outlying observations in a given dataset. Besides detection, the MD has been used severally in different fields, namely: In image processing, MD is used for image segmentation, in financial analysis, MD is used to predict financial crises, and in geostatistics, MD is used to detect influential observation in multiple spatial linear models. This approach utilizes the conventional arithmetic mean and sample covariance matrix to compute distances. The principle is to assign large distances to outlying and small distances to regular observations based on the selected cutoff point criterion. The MD produces elegant estimates when the number of observations exceeds the number of variables. In contrast, for high dimensional data where the number of variables surpasses the sample size, computation of Mahalanobis distance is infeasible because of the nonexistent inverse of the covariance matrix estimates.

In the presence of contaminated points, the Mahalanobis distance based on the classical location and scatter matrix rarely identify all the multivariate outliers in a given dataset. The problem becomes more pronounced when the dimension is increasingly high. This gives rise to masking or swamping problems because classical sample mean and covariance are not robust. Robust estimators of multivariate location and scatter such as the Minimum covariance determinant (MCD) and the minimum volume ellipsoid (MVE) estimates (Rousseeuw, 1984, 1985) are designed to replace the classical mean and covariance matrix estimators because the latter are susceptible in the presence of contamination. These estimators achieved a high breakdown point, but they are computationally intensive (time-consuming). As a result, Olive and Hawkins (2010) introduced the reweighted fast consistent and high (RFCH) breakdown estimator to address the limitations of the MVE and MCD estimators. Consequently, the main shortcoming of these methods is that they are not applicable when the number of variables surpasses the sample size. Nevertheless, these techniques' theoretical and computational difficulties and many new research problems provide excellent opportunities and meaningful challenges for developing high-dimensional data analysis.

The product-moment correlation is a classical correlation method used to measure the relationship between the predictor and dependent variables. This method forms the basis of multiple linear regression analysis. In regression, the objective is to simultaneously perform estimation and variables selection since not all the effects of independent variables are significant to the response variable in most applied research. Similar to the Mahalanobis distance estimates, if the number of independent variables exceeds the sample size, fitting the model to all the independent variables will produce corrupt regression coefficient estimates, especially when the independent variables are highly correlated. A common practice to deal with this problem is to apply a sure independent screening (SIS) algorithm. SIS method is a dimension reduction procedure used to reduce dimension from relatively high to below sample size and then perform parameter estimation and variables selection simultaneously via a lower dimensional regularized least square method such as least absolute shrinkage and selection operator (LASSO) and smoothly clipped absolute deviation (SCAD) (Fan and Lv, 2008). Despite the excellent performance of this procedure, it performed poorly in the presence of an outlier since the SIS method uses a classical correlation in the screening step, and they are much affected by outlying points. Hence, there is a need to propose a robust screening methodology that can produce better estimates even in the presence of outliers. Therefore this thesis focuses on developing an extended multivariate location and dispersion estimators that build upon the reweighted fast consistent and high break down (RFCH) of Olive and Hawkins (2010) and Minimum Diagonal Product (MDP) of (Ro et al., 2015).

1.2 Motivation of the Study

It is now patent that the classical multivariate estimates of mean and covariance matrix are susceptible to outliers. As a result, it is essential to use a robust multivariate location and dispersion matrix as an alternative to the classical mean and covariance. However, robust multivariate location and dispersion matrix estimators based on multivariate location and scatter matrix, such as the Fast Minimum Covariance Determinant (FMCD) method (Rousseeuw and Driessen, 1999) are faced with the problem of a computational burden, especially for large data points. Hubert et al. (2012) pointed out that MCD is affine equivariant but not permutation invariant. Thus, they proposed a deterministic Minimum Covariance Determinant algorithm (DetMCD) faster than MCD, which does not use a random subset.

Furthermore, Olive and Hawkins (2010) proposed a reweighted fast consistent and high break down (RFCH) estimator to find reliable location and scatter estimates. Compared with FMCD and the Ortogonalized Gnanadesikan-Kettenring (OGK), the RFCH shows better error measures of scatter estimates (Zhang et al., 2012; Alkenani and Yu, 2013). The authors revealed that RFCH possesses good performance at a different level of contamination. Recently,Uraibi and Midi (2019) practically showed the performance of RFCH in terms of outlier detection using stack-loss and Hawkins Bradu Kass datasets. They also showed that RFCH is computationally very fast. However, all previously mentioned methods are not feasible when the number of variables exceeds the sample size. As a result, the traditional Mahalanobis distance, which relies on the location and scatter matrix estimates, may not be feasible. This limitation has also been deliberated by Filzmoser et al. (2008) and Chen et al. (2010), as cited in (Ro et al., 2015).

The two latest standard methods that improve the performance of the Mahalanobis distance for high dimensional operations are the Minimum Diagonal Product (MDP) estimator of Ro et al. (2015) and Minimum regularized covariance determinant (MRCD) (Boudt et al., 2020). The MDP can be applied directly when $p >> n$. In this procedure, a subset of data points are computed such that the product of the diagonal values of the sample covariance matrix is minimal. Compared with the regularized minimum covariance determinant of Fritsch et al. (2011) and the robust principal component analysis (RPCA) of Hubert et al. (2005), the MDP showed better performance but produced higher type 1 error rates in detection (Martinez et al., 2020). The MRCD is an extension of MCD developed to knock down the limitation of MCD for not being able to use for high dimensional data. In this method, the subset-based covariance in MCD is replaced by the regularized covariance defined by the -subset of the weighted average of sample covariance and a predetermined target positive definite matrix. It is also proof to produce robust location and scatter matrix estimates for high dimensional cases. Nevertheless, our investigation revealed that the method is computationally expensive and produced higher classification error in the presence of outliers. This is because the robust distance produced by MRCD relies on the MCD distance, which is no longer reliable when the dimension increases relatively to sample size n (Boudt et al., 2020). The RFCH technique introduced by Olive and Hawkins (2010) is a fast consistent and high breakdown estimator of multivariate location and scatter matrix. This method produces reliable estimates compared to MCD but cannot be applied when the dimension exceeds the sample size. The procedures' deficiencies, as mentioned earlier, motivate us to propose a modified Reweighted Fast Consistent and High breakdown point (MRFCH) location and dispersion estimator for high dimensional data. The main philosophy of the proposed method is to use the diagonal elements of the scatter matrix in place of the whole scatter matrix in the calculation of Mahalanobis distance for the computation of RFCH algorithm while preserving the robustness properties of the RFCH estimator. According to Ro et al. (2015), the fast MCD method selects asubset with the minimal determinant of it covariance matrix estimates, and it cannot be applied in high dimensional data analysis. The Minimum Diagonal Product (MDP) estimator Ro et al. (2015) objectives is to select a subset with minimal product of the covariance matrix diagonals elements.

The outlier problem is a critical issue that affects prediction accuracy if not correctly identified. For instance, outlier detection techniques and prediction estimates are susceptible to outliers in a given data set. Mahalanobis distance (MD) is one of the most popular multivariate analysis procedures to detect multivariate outlying points. It is now apparent that classical Mahalanobis distance based on the classical mean and covariance are susceptible to outlying observations; hence, detecting outliers based on this classical MD may lead to the masking and or swamping problem. Besides, due to the nonexistence of the inverse of the classical covariance matrix, the outlier detection based on the classical Mahalanobis distance for a high dimensional dataset may not be feasible (Hubert et al., 2005; Filzmoser et al., 2008). Recently, Boudt et al. (2020) introduced the MRCD estimators, which can be used in place of MCD estimators for high dimensional data analysis. They cited that Mahalanobis distances based on the MRCD estimators can be applied for outlier detection in a high dimensional dataset. However, they mentioned that the cutoff value based on the square root of the chi-square values from the RD (MRCD) is liable to the more severe masking and/ or swapping problem. Since the asymptotic distribution of Mahalanobis distances calculated based on MRCD method is different from F and chi-square distributions. For this reason,Ro et al. (2015) proposed the Minimum Diagonal Product (MDP) estimators to obtain robust Mahalanobis distances in high dimensional data and used a cutoff point based on the standard normal distribution. Our search discloses that the effectiveness of the RD based on the MRCD estimator depreciates as the number of predictor variables becomes large. This method produces high misclassification errors during the outlier detection and classification calculations. Thus, these weakness has inspired us to develop outlier detection procedure based on the MRFCH in high dimensional data by conjoining the idea of Olive and Hawkins (2010) and (Ro et al., 2015). Note that the original RFCH cannot be applied for a high dimensional dataset due to the usage of the classical covariance matrix within the original RFCH algorithm, which produces a singularity problem. Our modified (MRFCH) procedure directly substituted the covariance matrix estimates in the computation of the Mahalanobis distances by its diagonal elements to obtain the final estimate of the location and covariance matrix and used them to compute robust Mahalanobis distances. In our proposed method, robust Mahalanobis distances are calculated based on the MRFCH algorithm, and outliers are detected based on the cutoff point presented by (Midi et al., 2009; Lim and Midi, 2016). Furthermore, our proposed MRFCH is an extension of the RFCH and faster than the existing MRCD since the original RFCH has been noted to be faster than the OGK and MCD and perform excellently in detecting high leverage observations in the analysis of the linear regression model (Zhang et al., 2012; Alkenani and Yu, 2013; Uraibi and Midi, 2019).

Similarly, the Pearson correlation is a statistical technique used to investigate the relationship between the response and predictors. Although this technique is fast and straightforward, it is susceptible to the presence of contamination because its calculations involve the use of sample mean of response and predictors variables. Several authors have discussed the nonrobustness of this technique using practical examples. Abdullah (1990) developed a robust correlation coefficient based on the least median of square (LMS) estimator to remedy this problem. As an alternative, Uraibi and Midi (2019) proposed a robust multivariate correlation matrix based on the Reweighted Fast Consistent and High breakdown point (RFCH) estimator (Olive & Hawkins, 2010). The former and latter methods show substantial resistance to outlying points even though they are impractical when the independent variables surpass the sample size. Recently, Raymaekers and Rousseeuw (2021) proposed a data transformation correlation procedure for high dimensional that uses wrapping function via the MAD and one-step M estimate of location. Moreover, a comparison between the Pearson correlation, rank correlation methods, and transformation method presented in Table 2 by Raymaekers and Rousseeuw (2021) has shown that quadrant correlation has the highest breakdown point value with the lowest efficiency. On the other hand, while the wrapping correlation achieves a breakdown point lower than quadrant correlation, the Pearson correlation achieves zero breakdowns but 100 % efficiency. Thus, to achieve a higher breakdown point and efficiency with less computational running time, we propose a robust correlation based on the modified RFCH (MRFCH) that is resistant to outliers.

Variable selection procedures significantly impact scientific and knowledge discovery in high dimensional datasets. The main objective of the variable selection technique is to determine the number of predictor variables that can be included in building a regression model to increase model predictive power and improve interpretability. The curse of dimensionality is the major challenge in building an efficient working statistical model in high dimensional data analysis. The traditional all possible subsets techniques, which include forward selection, backward elimination, and stepwise regression, are often used to determine the most critical variables that influence the dependent variable. However, in most situations, they do not ensure the consistent selection, and they are computationally intensive, having the problem of a singular data matrix (Fan and Li, 2001; Breiman, 1996). To remedy these problems, the penalized least squares estimators such as the least squares Bridge estimator of Frank and Friedman (1993) and the least squares smoothly clipped absolute deviation (LS-SCAD) estimator of Fan and Li (2001) are presented. Tibshirani (1996) proposed the least square LASSO as a particular case of the Bridge regression. Furthermore, the Lasso derivative methods, including the Adaptive Lasso of Zou (2006), SEA-Lasso, and NSEA-Lasso of Qian and Yang (2013) are put forward. The exciting properties of these estimators are that they perform both estimation and variable selection simultaneously and work well for high dimensional data. Besides, none of those mentioned above penalized estimators perform well for ultrahigh dimensional data due to statistical accuracy, algorithmic stability, and computational expediency challenges. To tackle these shortcomings, Fan and Lv (2008) introduced the concept of correlation based sure independent screening (SIS), and it improved iteratively sure independence screening (ISIS) algorithms to filter out the predictor variables that have a weak correlation with the response variable. The SIS and ISIS attract the attention of researchers due to their simplicity and wide range of applications in many real-life problems. Several extensions have been proposed in the literature. For example, Hall and Miller (2009) suggested the generalized correlation ranking method. Fan et al. (2011) presented an iterative nonparametric sure independence screening for sparse additive model.

It is essential to highlight that the aforementioned correlation-based screening methods do not perform well when the classical underlying assumptions are violated. Thus, Li et al. (2012) proposed the robust rank correlation screening RRCS based on Kendall tau rank correlation to deal with heavy tailed distribution observations. Kong et al. (2017) proposed the sure screening based on canonical correlation procedure. Li et al. (2012) introduced a distance correlation based screening algorithm as discussed in (Zhong and Zhu, 2015). Ma and Zhang (2016) developed a robust model free feature screening via quantile correlation. Wang et al. (2017) and Wang et al. (2016) proposed two step robust variable screening that combined influential diagnostics procedure and the sure screening based on the distance correlation to conduct variable selection. Ahmed and Bajwa (2019) recently studied an extended correlation-based variable selection for a linear model with post-screening inference. One shortcoming of these robust correlation-screening algorithms is that they only consider the problem of heavy tailed distribution, but not outliers on $\mathbf{X}$ and $\mathbf{y}$ direction that in reality is possible, refer to Li et al. (2012) and (Kong et al., 2017). Also, an example of this scenario is given in (Arslan, 2012; Smucler and Yohai, 2017; Uraibi and Midi, 2019). They all demonstrated the effect of outlying observations on variable selection via penalized methods. Moreover, no research work has considered correlation-based screening algorithms with such problem of $\mathbf{X}$ and $\mathbf{y}$ outliers. This motivates us to propose a robust and efficient correlation-based sure independent screening procedure for sparse high dimensional regression model in the presence of outlying point via the modified Reweighted Fast Consistent and High breakdown point (RFCH) estimator.

1.3 Objective of the Study

The primary aim of this thesis work is to study and examine the behavior of the various existing robust methods in high dimensional data analysis and propose a novel procedure for computing robust location and scatter matrix, robust outlier detection, robust correlation coefficients, and robust Penalized LAD-SCAD estimator for high dimensional data via the modified reweighted consistent and high breakdown (MR-

FCH) estimator. To achieve our aims, we consider the following specific objectives:

1. To improvise the reweighted fast consistent and high breakdown (RFCH) estimator to estimate the multivariate location and dispersion matrix in high dimensional data.
2. To develop an efficient algorithm for the identification of high leverage points based on the modified RFCH for high dimensional data.
3. To formulate a new robust correlation via the modified RFCH for high dimensional data.
4. To construct Penalized LAD-SCAD and LAD-Lasso for estimation and variable selection based on robust screening method via the modified RFCH for high dimensional models.

1.4 Significant of the Study

The main goal of regression analysis is to perform estimation and variables selection simultaneously since in many real life applications, not all the effects of independent variables are significant to the response. For example, in genome wide association studies (GWAS), it is believed that a particular kind of cancer disease is only associated with a few genes functioning together.Wang et al. (2007) used the china stock dataset obtained from the China Centre for Economic Research to determine the influence of some factors on the return on equity (ROE), considered as the response variable. Based on the least absolute deviation Lasso method, their finding pointed only three variables as significant out of nine. These examples, with many others, necessitate the development of various existing sure screening based methods, especially when the dimension of variables is much larger than the sample size. The curse of dimensionality is the major challenge in building an efficient working statistical model in high dimensional data analysis. Classical estimates of correlation and dispersion matrices produce corrupt estimates when outliers are present. In most situations, even a single outlier can disfigure the classical estimates of mean, covariance, correlation, Mahalanobis distances, and variable selection. In this study, we extend the reweighted consistent and fast breakdown (RFCH) estimation (Olive and Hawkins, 2010) to higher dimensions by replacing the sample covariance matrix of the Mahalanobis distance with it diagonal elements before computing the distances. The resulting extended RFCH enjoys the robustness properties of the RFCH by Olive and Hawkins (2010), even when the dimension of variables exceeds the sample size. The performance of the extended RFCH is confirmed by simulation study for both high and low dimensions cases. Secondly, we show the use of the extended RFCH for outlier detection and classification based on simulation and real data from the gene expression data, chemometrics, and octane data set. We believe that extended RFCH is a valuable alternative to the existing high dimensional robust multivariate analysis.

Thirdly, the study suggests a robust correlation for high dimensions based on the extended RFCH estimator as an alternative to the existing correlation coefficient. Compared with the Pearson correlation, Kendall, and robust correlation Raymaekers and Rousseeuw (2021), the robust correlation based on extended RFCH shows better bias and MSE values. Thus the robust correlation via extended RFCH is a good option, especially in the presence of X and y outliers.

Finally, a robust and efficient variable selection and estimation procedure via the robust correlation based on modified RFCH shows excellent performance based on simulation and real data examples. In addition, the method is developed to solve dimension reduction problems in the presence of outliers in the variable selection algorithm.

1.5 Limitation of the Study

This thesis will not be completed without limitations. The thesis developed four new novel methods based on the RFCH estimator for a high dimensional dataset. Firstly, we showed that the diagonal elements of the covariance matrix could be used instead of the entire covariance matrix within the RFCH algorithm. Following Ro et al. (2015), incorporating the diagonal elements idea in DRGP (MVE) estimator could be excellent future research. Secondly, we compared the proposed modified RFCH (MRFCH) estimator to the MRCD, OGKQn, OGKmad, and MDP. To evaluate the performance of our develop algorithm, the MSE and time criterion was applied. Comparing the proposed method with some other estimators would be another good topic of study.

Thirdly, we compare the Pearson correlation, Kendall, and correlation based on Raymaekers and Rousseeuw (2021) study to the proposed robust correlation coefficient. Fourthly, we compare the RCS+LAD-Lasso, RCS+LAD-SCAD, WCS+LAD-Lasso, and WCS+LAD-SCAD with our proposed MRFCHCS+LAD-Lasso, and MRFCHCS+LAD-SCAD. These estimators are selected because they all use multivariate correlation estimates or location or dispersion matrix, or Mahalanobis distance function within their computations. Due to inadequate higher performing computer and time constrain, we only repeated our experiment for 100 and 200 iterations (Wang et al., 2015b). In addition, the same sets of data used repeatedly by previous researchers were adopted in this study to show consistent results with other existing works.

1.6 Outline of the Thesis

Following the objective and scopes of study, the contents of this thesis are designed into seven chapters. The thesis chapters are arranged so that each objective in the thesis is superficial in the sequence outline.

Chapter Two discusses the literature reviews on penalized and unpenalized regression estimators. The location and scatter matrix estimators for low dimensional (MCD, MVE, FMCD, DetMCD, and RFCH) and high dimensional (OGK, MDP, MRCD) models are presented. High leverage detection procedures based on the MDP and MRCD, including the outlier detection based distance correlation learning, are also reviewed. The concept of correlation estimators in the presence of outlying point was reviewed. Furthermore, the SIS, RSIS, DCSIS, variable selection based on the canonical correlation was deliberated. Finally, influential diagnostic methods are discussed.

Chapter Three discussed the proposed modified RFCH estimators, which utilize the original RFCH estimators. The MRFCH algorithm is presented. Simulation and real data examples were used to demonstrate the performance of the proposed MRFCH estimator. We also present simple examples using an artificial dataset for simplicity and a better understanding of the existing and proposed algorithm.

Chapter Four discusses the new outlier detection and classification procedure based on the modified RFCH (MRFCH) estimator of location and scatter matrix. The detection and classification power of the existing MRCD and MDP based on the robust distance (RD) are evaluated using simulations and three real life data (octane, NCI60, and Brain datasets).

Chapter Five discusses the new robust correlation algorithm developed based on the modified RFCH location, scatter matrix, and robust distance estimates. The existing Pearson correlation, Kendall rank correlation, and the wrapped correlation algorithm are compared with the new robust correlation learning algorithm based on simulation and real data (glass vessel, cardiomyopathy microarray, and octane datasets).

Chapter Six develops a penalized LAD-SCAD regression estimator based on a robust sure independence screening procedure for the sparse high dimensional regression model. The proposed MRFCHCS+LAD-SCAD and MRFCHCS+LAD-Lasso are compared with the RCS+LAD-Lasso, RCS+LAD-SCAD, WCS+LAD-Lasso, WCS+LAD-SCAD and using simulation and real data examples.

Chapter seven includes the summary, conclusions, recommendations, and possible future research areas.

REFERENCES

Abdullah, M. B. (1991). On a robust correlation coefficient. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 40(2):245–245.

Ahmed, T. and Bajwa, W. U. (2019). Exsis: Extended sure independence screening for ultrahigh-dimensional linear models. *Signal processing*, 159:33–48.

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*, pages 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.

Amato, U., Antoniadis, A., De Feis, I., and Gijbels, I. (2021). Penalised robust estimators for sparse and high-dimensional linear models. *Statistical Methods & Applications*, 30(1):1–48.

Arslan, O. (2012). Weighted lad-lasso method for robust parameter estimation and variable selection in regression. *Computational Statistics & Data Analysis*, 56(6):1952–1965.

Arslan, O. and Billor, N. (2000). Robust liu estimator for regression based on an m-estimator. *Journal of applied statistics*, 27(1):39–47.

Bai, Z. and Wu, Y. (1997). Generalm-estimation. *Journal of multivariate analysis*, 63(1):119–135.

Beckman, R. J. and Cook, R. D. (1983). Outlier.......... s. *Technometrics*, 25(2):119–149.

Belsley, D. A., Kuh, E., and Welsch, R. E. (2005). *Regression diagnostics: Identifying influential data and sources of collinearity*. John Wiley & Sons.

Boudt, K., Rousseeuw, P. J., Vanduffel, S., and Verdonck, T. (2020). The minimum regularized covariance determinant estimator. *Statistics and Computing*, 30(1):113–128.

Bravais, A. (1844). *Analyse mathématique sur les probabilités des erreurs de situation d'un point*. Impr. Royale.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *The annals of statistics*, 24(6):2350–2383.

Brown, P. J., Fearn, T., and Vannucci, M. (2001). Bayesian wavelet regression on curves with application to a spectroscopic calibration problem. *Journal of the American Statistical Association*, 96(454):398–408.

Bulut, H. (2020). Mahalanobis distance based on minimum regularized covariance determinant estimators for high dimensional data. *Communications in Statistics-Theory and Methods*, 49(24):5897–5907.

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

Chang, L., Roberts, S., and Welsh, A. (2018). Robust lasso regression using tukey's biweight criterion. *Technometrics*, 60(1):36–47.

Chen, S.-H., Kuo, Y., and Lin, J.-K. (2020). Using mahalanobis distance and decision tree to analyze abnormal patterns of behavior in a maintenance outsourcing process-a case study. *Journal of Quality in Maintenance Engineering*.

Chen, S. X., Qin, Y.-L., et al. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *The Annals of Statistics*, 38(2):808–835.

Chernick, M. R. (1982). The influence function and its application to data validation. *American Journal of Mathematical and Management Sciences*, 2(4):263–288.

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

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

Croux, C. and Haesbroeck, G. (2000). Principal component analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies. *Biometrika*, 87(3):603–618.

Davies, P. L. (1987). Asymptotic behaviour of s-estimates of multivariate location parameters and dispersion matrices. *The Annals of Statistics*, pages 1269–1292.

Dhhan, W., Rana, S., and Midi, H. (2017). A high breakdown, high efficiency and bounded influence modified gm estimator based on support vector regression. *Journal of Applied Statistics*, 44(4):700–714.

Donoho, D. L. (1982). Breakdown properties of multivariate location estimators. Technical report, Technical report, Harvard University, Boston. URL http://www-stat. stanford ....

Ellenberg, J. H. (1976). Testing for a single outlier from a general linear regression. *Biometrics*, pages 637–645.

Erickson, J., Har-Peled, S., and Mount, D. M. (2006). On the least median square problem. *Discrete & Computational Geometry*, 36(4):593–607.

Fan, J., Feng, Y., and Song, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*, 106(494):544–557.

Fan, J., Feng, Y., and Wu, Y. (2010). High-dimensional variable selection for cox's proportional hazards model. In *Borrowing strength: Theory powering applications–a Festschrift for Lawrence D. Brown*, pages 70–86. Institute of Mathematical Statistics.

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

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911.

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The annals of statistics*, 32(3):928–961.

Filzmoser, P., Maronna, R., and Werner, M. (2008). Outlier identification in high dimensions. *Computational statistics & data analysis*, 52(3):1694–1711.

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

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1.

Fritsch, V., Varoquaux, G., Thyreau, B., Poline, J.-B., and Thirion, B. (2011). Detecting outlying subjects in high-dimensional neuroimaging datasets with regularized minimum covariance determinant. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 264–271. Springer.

Gao, X. and Huang, J. (2010). Asymptotic analysis of high-dimensional lad regression with lasso. *Statistica Sinica*, pages 1485–1506.

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

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

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

Hall, P. and Miller, H. (2009). Using generalized correlation to effect variable selection in very high dimensional problems. *Journal of Computational and Graphical Statistics*, 18(3):533–550.

Hampel, F. R., Rousseeuw, P. J., and Ronchetti, E. (1981). The change-of-variance curve and optimal redescending m-estimators. *Journal of the American Statistical Association*, 76(375):643–648.

Hoaglin, D. C. and Welsch, R. E. (1978). The hat matrix in regression and anova. *The American Statistician*, 32(1):17–22.

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

Huang, J. and Xie, H. (2007). Asymptotic oracle properties of scad-penalized least squares estimators. In *Asymptotics: Particles, processes and inverse problems*, pages 149–166. Institute of Mathematical Statistics.

Huang, J. Z. and Shen, H. (2004). Functional Coefficient Regression Models for Non-linear Time Series: A Polynomial Spline Approach. *Scandinavian Journal of Statistics*, 31:515–534.

Huber, P. (1981). Robust statistics. new york: John wiley and sons. *HuberRobust statistics1981*.

Huber, P. J. (1973). Robust regression: asymptotics, conjectures and monte carlo. *The annals of statistics*, pages 799–821.

Hubert, M., Rousseeuw, P. J., and Vanden Branden, K. (2005). Robpca: a new approach to robust principal component analysis. *Technometrics*, 47(1):64–79.

Hubert, M., Rousseeuw, P. J., and Verdonck, T. (2012). A deterministic algorithm for robust location and scatter. *Journal of Computational and Graphical Statistics*, 21(3):618–637.

Hubert, M. and Van der Veeken, S. (2008). Outlier detection for skewed data. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 22(3-4):235–246.

Imon, A. R. (1996). *Sub-sample Methods in Regression Residual Prediction and Diagnostics*. PhD thesis, University of Birmingham.

Kong, X.-B., Liu, Z., Yao, Y., and Zhou, W. (2017). Sure screening by ranking the canonical correlations. *Test*, 26(1):46–70.

Lee, Y.-C. and Teng, H.-L. (2009). Predicting the financial crisis by mahalanobis–taguchi system–examples of taiwan's electronic sector. *Expert Systems with Applications*, 36(4):7469–7478.

Lemberge, P., De Raedt, I., Janssens, K. H., Wei, F., and Van Espen, P. J. (2000). Quantitative analysis of 16–17th century archaeological glass vessels using pls regression of epxma and μ-xrf data. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 14(5-6):751–763.

Li, G., Peng, H., Zhang, J., and Zhu, L. (2012). Robust rank correlation based screening. *The Annals of Statistics*, 40(3):1846–1877.

Li, G., Peng, H., and Zhu, L. (2011). Nonconcave penalized m-estimation with a diverging number of parameters. *Statistica Sinica*, pages 391–419.

Liebmann, B., Friedl, A., and Varmuza, K. (2009). Determination of glucose and ethanol in bioethanol production by near infrared spectroscopy and chemometrics. *Analytica Chimica Acta*, 642(1-2):171–178.

Lim, H. A. and Midi, H. (2016). Diagnostic robust generalized potential based on index set equality (drgp (ise)) for the identification of high leverage points in linear model. *Computational statistics*, 31(3):859–877.

Lu, S., Chen, X., and Wang, H. (2021). Conditional distance correlation sure independence screening for ultra-high dimensional survival data. *Communications in Statistics-Theory and Methods*, 50(8):1936–1953.

Ma, X. and Zhang, J. (2016). Robust model-free feature screening via quantile correlation. *Journal of Multivariate Analysis*, 143:472–480.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. National Institute of Science of India.

Maronna, R. A., Martin, R. D., Yohai, V. J., and Salibián-Barrera, M. (2019). *Robust statistics: theory and methods (with R)*. John Wiley & Sons.

Maronna, R. A. and Yohai, V. J. (1995). The behavior of the stahel-donoho robust multivariate estimator. *Journal of the American Statistical Association*, 90(429):330–341.

Maronna, R. A. and Zamar, R. H. (2002). Robust estimates of location and dispersion for high-dimensional datasets. *Technometrics*, 44(4):307–317.

Martinez, W. G., Weese, M. L., and Jones-Farmer, L. A. (2020). A one-class peeling method for multivariate outlier detection with applications in phase i spc. *Quality and Reliability Engineering International*, 36(4):1272–1295.

Midi, H., Hendi, H. T., Arasan, J., and Uraibi, H. (2020). Fast and robust diagnostic technique for the detection of high leverage points. *Pertanika Journal of Science & Technology*, 28(4).

Midi, H., Rana, M. S., and Imon, A. (2009). The performance of robust weighted least squares in the presence of outliers and heteroscedastic errors. *WSEAS Transactions on Mathematics*, 8(7):351–361.

Militino, A., Palacios, M., and Ugarte, M. (2006). Outliers detection in multivariate spatial linear models. *Journal of statistical planning and inference*, 136(1):125–146.

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

Oosterhoff, J. (1994). Trimmed mean or sample median? *Statistics & Probability Letters*, 20(5):401–409.

Pearson, K. (1895). Correlation coefficient. In *Royal Society Proceedings*, volume 58, page 214.

Qian, W. and Yang, Y. (2013). Model selection via standard error adjusted adaptive lasso. *Annals of the Institute of Statistical Mathematics*, 65(2):295–318.

Rajaratnam, B., Roberts, S., Sparks, D., and Yu, H. (2019). Influence diagnostics for high-dimensional lasso regression. *Journal of Computational and Graphical Statistics*, 28(4):877–890.

Rashid, A. M., Midi, H., Slwabi, W. D., and Arasan, J. (2021). An efficient estimation and classification methods for high dimensional data using robust iteratively reweighted simpls algorithm based on nu-support vector regression. *IEEE Access*, 9:45955–45967.

Raymaekers, J. and Rousseeuw, P. J. (2021). Fast robust correlation for high-dimensional data. *Technometrics*, 63(2):184–198.

Restrepo, A. and Bovik, A. C. (1988). Adaptive trimmed mean filters for image restoration. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(8):1326–1337.

Ro, K., Zou, C., Wang, Z., and Yin, G. (2015). Outlier detection for high-dimensional data. *Biometrika*, 102(3):589–599.

Rousseeuw, P. and Leroy, A. (1987). Robust regression and outlier detection. new york: John wiley& sons.

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):37.

Rousseeuw, P. J. and Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical association*, 88(424):1273–1283.

Rousseeuw, P. J., Debruyne, M., Engelen, S., and Hubert, M. (2006). Robustness and outlier detection in chemometrics. *Critical reviews in analytical chemistry*, 36(3-4):221–242.

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

Ryan, T. (1997). Modern regression method, john willey & sons. *Inc. New York*.

Segal, M. R., Dahlquist, K. D., and Conklin, B. R. (2003). Regression approaches for microarray data analysis. *Journal of Computational Biology*, 10(6):961–980.

Shevlyakov, G. and Khvatova, T. Y. (1998). On robust estimation of a correlation coefficient and correlation matrix. In *MODA 5—Advances in Model-Oriented Data Analysis and Experimental Design*, pages 153–162. Springer.

Shevlyakov, G. and Smirnov, P. (2011). Robust estimation of the correlation coefficient: An attempt of survey. *Austrian Journal of Statistics*, 40(1&2):147–156.

Smucler, E. and Yohai, V. J. (2017). Robust and sparse estimators for linear regression models. *Computational Statistics & Data Analysis*, 111:116–130.

Stahel, W. A. (1981). *Robuste schätzungen: infinitesimale optimalität und schätzungen von kovarianzmatrizen*. PhD thesis, ETH Zurich.

Stöckl, S. and Hanke, M. (2014). Financial applications of the mahalanobis distance. *Applied Economics and Finance*, 1(2):78–84.

Stuart, C. (2011). Robust regression. *Department of Mathematical Sciences, Durham University*, 169.

Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *The annals of statistics*, 35(6):2769–2794.

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

Todorov, V., Filzmoser, P., et al. (2009). An object-oriented framework for robust multivariate analysis.

Uraibi, H. S. and Midi, H. (2019). On robust bivariate and multivariate correlation coefficient. *Economic Computation & Economic Cybernetics Studies & Research*, 53(2).

Uraibi, H. S., Midi, H., and Rana, S. (2015). Robust stability best subset selection for autocorrelated data based on robust location and dispersion estimator. *Journal of Probability and Statistics*, 2015.

Uraibi, H. S., Midi, H., and Rana, S. (2017). Selective overview of forward selection in terms of robust correlations. *Communications in Statistics-Simulation and Computation*, 46(7):5479–5503.

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

Visuri, S., Koivunen, V., and Oja, H. (2000). Sign and rank covariance matrices. *Journal of Statistical Planning and Inference*, 91(2):557–575.

Wan, W. and Jamaludin, S. (2015). Smoothing Wind and Rainfall Data through Functional Data Analysis Technique. *Jurnal Teknologi*, 74:105–112.

Wang, H. and Leng, C. (2007). Unified lasso estimation by least squares approximation. *Journal of the American Statistical Association*, 102(479):1039–1048.

Wang, H., Li, G., and Jiang, G. (2007). Robust regression shrinkage and consistent variable selection through the lad-lasso. *Journal of Business & Economic Statistics*, 25(3):347–355.

Wang, J. L., Chiou, J. M., and Muller, H. G. (2015a). Review of Functional Data Analysis. *Annual Reviews Statistics*, 1:1–41.

Wang, M., Song, L., and Tian, G.-l. (2015b). Scad-penalized least absolute deviation regression in high-dimensional models. *Communications in Statistics-Theory and Methods*, 44(12):2452–2472.

Wang, T., Zheng, L., Li, Z., and Liu, H. (2017). A robust variable screening method for high-dimensional data. *Journal of Applied Statistics*, 44(10):1839–1855.

Wang, T. and Zhu, L. (2011). Consistent tuning parameter selection in high dimensional sparse linear regression. *Journal of Multivariate Analysis*, 102(7):1141–1151.

Wright, S. (1921). Correlation and causation.

Xiao, X., Fu, D., Shi, Y., and Wen, J. (2020). Optimized mahalanobis–taguchi system for high-dimensional small sample data classification. *Computational intelligence and neuroscience*, 2020.

Xie, H. and Huang, J. (2009). Scad-penalized regression in high-dimensional partially linear models. *The Annals of Statistics*, 37(2):673–696.

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):119.

Zhang, Y., Huang, D., Ji, M., and Xie, F. (2011). Image segmentation using pso and pcm with mahalanobis distance. *Expert Systems with Applications*, 38(7):9036–9040.

Zhang, Y., Li, Z., Cai, J., and Wang, J. (2010). Image segmentation based on fcm with mahalanobis distance. In *International Conference on Information Computing and Applications*, pages 205–212. Springer.

Zhao, J., Leng, C., Li, L., and Wang, H. (2013). High-dimensional influence measure. *The Annals of Statistics*, 41(5):2639–2667.

Zhong, W. and Zhu, L. (2015). An iterative approach to distance correlation-based sure independence screening. *Journal of Statistical Computation and Simulation*, 85(11):2331–2345.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429.