ROBUST DIAGNOSTICS AND PARAMETER ESTIMATION IN LINEAR REGRESSION FOR HIGH DIMENSIONAL DATA

By

SITI ZAHARIAH BINTI ABDUL WAHAB

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

November 2023

IPM 2023 12

DEDICATION

To my parents,
Abdul Wahab bin Mohd Nor & Faridah binti Ayub

And

To my husband,
Muhammad Adli bin Yahaya

And

To my sons,
Luqman Alhakim bin Muhammad Adli & Muhammad Alfateh bin Muhammad Adli

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

ROBUST DIAGNOSTICS AND PARAMETER ESTIMATION IN LINEAR REGRESSION FOR HIGH DIMENSIONAL DATA

By

SITI ZAHARIAH BINTI ABDUL WAHAB

November 2023

Chairman : Professor Habshah binti Midi, PhD
Institute : Mathematical Research

Several methods of identification of HLPs in HDD have been put forth, including the methods of Robust Mahalanobis Distance (RMD) based on Minimum Regularized Covariance Determinant (MRCD) and Robust Principal Component Analysis (ROBPCA). However, they suffer from masking and swamping effects when the predictor variables are at least 700. In addressing this problem, a modification of HLPs detection method called Robust Mahalanobis Distance based on the combination of the Minimum Regularized Covariance Determinant and Principal Component Analysis (RMD-MRCD-PCA) is proposed. Empirical evidence from simulation studies and real data show that the RMD-MRCD-PCA method is very successful in the detection HLPs with negligible masking and swamping effects.

Numerous classical methods, such as leave-one-out cross-validation (LOOCV) and *K*-fold cross-validation (*K*-FoldCV) are developed to determine the optimal number of PLS components. Nonetheless, they are easily affected by HLPs. Thus, robust cross validation techniques, denoted as RMD-MRCD-PCA-LOOCV and RMD-MRCD-PCA-*K*-FoldCV are proposed to remedy this problem. The results of the simulation study and real data sets indicate that the proposed methods successfully select the appropriate number of PLS components.

The statistically inspired modification of partial least squares (SIMPLS) is the popular method to deal with multicollinearity in high dimensional data. Nonetheless, the SIMPLS method is vulnerable to the existence of HLPs. Hence, the robust weight based on RMD-MRCD-PCA of SIMPLS (RMD-MRCD-PCA-RWSIMPLS) is established to overcome this issue. Simulation experiments and real examples have demonstrated that the RMD-MRCD-PCA-RWSIMPLS is more efficient than the SIMPLS and the RWSIMPLS methods.

Partial least squares discriminant analysis (PLSDA) is the popular classifier for HDD. Nevertheless, the PLSDA is easily affected by the presence of HLPs. Hence, a robust weighted partial least squares discriminant analysis based on the weighting function of RMD-MRCD-PCA (RMD-MRCD-PCA-RWPLSDA) is proposed to close the gap in the literature. The results of the simulation study and real datasets show that the RMD-MRCD-PCA-RWPLSDA method successfully and efficiently classifies the data into binary and multiple groups.

Hotelling T^2 based on PLS (T^2-PLS) method has been proposed for variable selection technique in HDD. However, the T^2-PLS is not resistant to the HLPs. To rectify this issue, the robust Hotelling T^2 variable selection method, which is based on the RMD-MRCD-PCA-RWSIMPLS, is proposed. The results of simulation study and real datasets indicate that the T^2-RMD-MRCD-PCA-RWSIMPLS method successfully selects appropriate number of important variables to be included in the model with the least value of mean square error.

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

DIAGNOSTIK TEGUH DAN ANGGARAN PARAMETER DALAM REGRESI LINEAR UNTUK DATA DIMENSI TINGGI

Oleh

SITI ZAHARIAH BINTI ABDUL WAHAB

November 2023

Pengerusi : Profesor Habshah binti Midi, PhD
Institut : Penyelidikan Matematik

Beberapa kaedah pengecaman HLP dalam HDD telah dikemukakan, termasuk kaedah Jarak Mahalanobis Teguh (RMD) berdasarkan Penentu Kovarian Teratur Minimum (MRCD) dan Analisis Komponen Utama Teguh (ROBPCA). Walau bagaimanapun, mereka mengalami masalah kesan penyorokan dan limpahan apabila pembolehubah peramal sekurang-kurangnya 700. Dalam menangani masalah ini, kaedah pengesanan HLP baharu yang dipanggil Jarak Mahalanobis Teguh berdasarkan gabungan Penentu Kovarian Teratur Minimum dan Analisis Komponen Utama (RMD-MRCD-PCA) dicadangkan. Bukti empirikal daripada kajian simulasi dan data sebenar menunjukkan bahawa kaedah RMD-MRCD-PCA sangat berjaya dalam pengesanan HLP dengan kesan penyorokan dan kesan limpahan yang boleh diabaikan.

Banyak kaedah klasik, seperti pengesahan silang *leave-one out* (LOOCV) dan pengesahan silang lipatan *K* (*K*-FoldCV) dibangunkan untuk menentukan bilangan komponen PLS yang optimum. Namun begitu, mereka mudah dipengaruhi oleh HLP. Oleh itu, teknik pengesahan silang yang teguh, yang ditandakan sebagai MRCD-PCA-LOOCV dan MRCD-PCA-*K*-FoldCV dibangunkan untuk menyelesaikan masalah ini. Keputusan kajian simulasi dan set data sebenar menunjukkan kaedah yang dicadangkan berjaya memilih bilangan komponen PLS yang sesuai

Pengubahsuaian statistik kuasa dua terkecil separa (SIMPLS) ialah kaedah popular untuk menangani multikolineariti dalam data dimensi tinggi. Namun begitu, kaedah SIMPLS terdedah kepada kewujudan HLP. Oleh itu, pemberat teguh berdasarkan RMD-MRCD-PCA SIMPLS (RMD-MRCD-PCA-RWSIMPLS) dibangunkan untuk mengatasi isu ini. Eksperimen simulasi dan contoh sebenar telah menunjukkan bahawa RMD-MRCD-PCA-RWSIMPLS adalah lebih cekap daripada kaedah SIMPLS dan RWSIMPLS.

Analisis diskriminasi kuasa dua terkecil separa (PLSDA) ialah pengelas popular untuk HDD. Namun begitu, PLSDA mudah terjejas dengan kehadiran HLP. Oleh itu, analisis diskriminasi kuasa dua terkecil separa berpemberat teguh berdasarkan fungsi pemberat RMD-MRCD-PCA (MRCD-PCA-RWPLSDA) dicadangkan untuk merapatkan jurang dalam literatur. Hasil kajian simulasi dan set data sebenar menunjukkan kaedah RMD-MRCD-PCA-RWPLSDA berjaya mengelaskan data kepada kumpulan binari dan kumpulan berbilang dengan cekap.

Hotelling T^2 berdasarkan kaedah PLS (T^2-PLS) telah dicadangkan untuk teknik pemilihan pembolehubah dalam HDD. Walau bagaimanapun, T^2-PLS tidak tahan terhadap HLP. Bagi mengatasi masalah ini, kaedah pemilihan pembolehubah Hotelling teguh T^2, yang berdasarkan RMD-MRCD-PCA-RWSIMPLS, dibangunkan. Hasil kajian simulasi dan set data sebenar menunjukkan bahawa kaedah T^2-RMD-MRCD-PCA-RWSIMPLS berjaya memilih bilangan pembolehubah penting yang sesuai untuk dimasukkan ke dalam model dengan nilai min kuasadua ralat yang paling kecil.

ACKNOWLEDGEMENTS

All praises and glorification are due to Allah (SWT). I thank Him for the ability, wisdom and strength bestowed on me to conduct this study. I really thank Him.

I would like to express my sincere gratitude to my supervisor, Prof. Dr. Habshah Midi, who has taught and put me through over the years. I have benefited enormously from her continuous support and confidence throughout my research. Without her help and support, this dissertation would have been impossible, I feel truly privileged to have been her student. I am also very grateful to my committee members Associate Dr. Norhaslinda binti Ali, Dr. Mohd Shafie Bin Mustafa for their guidance, encouragement, and support. I thank all the staff of Institute for Mathematical Research for their help and support in one way or the other.

I would also like to express my gratitude to my mother Faridah Ayub and my father Abdul Wahab Mohd Nor for their prayers, support and encouragement, to also my brothers, sisters and relatives for their moral support and prayers. My special thanks go to my beloved husband Muhammad Adli Yahaya for his prayers, endurance, support, and encouragement. I would not forget the patient and endurance of our beloved children Luqman Alhakim and Muhammad Alfateh. They have really sacrificed much to see this dream has become a reality; may Allah bless them. Similarly, I would like to appreciate my friends, colleagues in research, officemate, and well-wishers for their prayers.

Lastly, I am very grateful to UniKL for their financial support during my studies, without their support this study wouldn’t have been possible. I also sincerely appreciate Institute for Mathematical Research and UPM in general for providing support and assistance in one way or the other.

This thesis was submitted to the Senate of the 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
(Chairman)

Mohd Shafie bin Mustafa, **PhD**
Senior Lecturer
Faculty of Science
Universiti Putra Malaysia
(Member)

Norhaslinda binti Ali, **PhD**
Senior Lecturer
Faculty of Science
Universiti Putra Malaysia
(Member)

Zalizah binti Awang Long, PhD
Professor
Malaysian Institute of Information Technology
Universiti Kuala Lumpur
(Member)

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

Date: 14 March 2024

TABLE OF CONTENTS

Page

ABSTRACT i
ABSTRAK iii
ACKNOWLEDGEMENTS v
APPROVAL vi
DECLARATION viii
LIST OF TABLES xiii
LIST OF FIGURES xv
LIST OF ABBREVIATIONS xvii

CHAPTER

1 INTRODUCTION 1
1.1 Background and Purpose 1
1.2 Mahalanobis Distance (MD) 2
1.3 Minimum Covariance Determinant (MCD) 2
1.4 Minimum Volume Ellipsoid (MVE) 3
1.5 Minimum Regularized Covariance Determinant (MRCD) 3
1.6 Principal Component Analysis 5
1.6.1 Data Matrix 6
1.6.2 Variance-Covariance Matrix 6
1.6.3 Linear Combination 6
1.6.4 Linear Independence 6
1.7 Partial Least Squares (PLS) 7
1.8 Motivation of the Study 7
1.9 Objective of Thesis 10
1.10 Scope and Limitation of Study 11
1.11 Outline of the Thesis 11

2 LITERATURE REVIEW 13
2.1 Introduction 13
2.2 Outliers Identification Methods in High Dimensional Data 13
2.3 Selecting an Optimal Number of PLS Components 17
2.4 Partial Least Squares Regression and Classification in Linear Regression 20
2.5 Classification of Outliers in High Dimensional Data 25
2.6 Classification of Observations in High Dimensional Data 26
2.7 Variable Selection Methods 30

3 IDENTIFICATION OF HIGH LEVERAGE POINTS IN HIGH DIMENSIONAL DATA BASED ON THE MINIMUM REGULARIZED COVARIANCE DETERMINANT AND PRINCIPAL COMPONENT ANALYSIS 34
3.1 Introduction 34

3.2 The Algorithm of Minimum Regularized Covariance Determinant (MRCD) 34
3.3 Robust Principal Component Analysis (ROBPCA) 37
3.4 The Improvised Robust Mahalanobis Distance based on the Combination of Minimum Regularized Covariance Determinant and Principal Component Analysis 38
3.5 Monte Carlo Simulation 40
3.6 Numerical Example 46
3.6.1 Octane Dataset 47
3.6.2 Biscuit Dough Dataset 48
3.6.3 Glass Spectra Dataset 50
3.7 Conclusion 51

4 THE SELECTION OF OPTIMAL NUMBER OF PLS COMPONENTS BASED ON THE COMBINATION OF CROSS-VALIDATION AND RMD-MRCD-PCA WEIGHTING FUNCTION 53
4.1 Introduction 53
4.2 Cross Validation (CV) 54
4.2.1 Leave-One-Out Cross Validation (LOOCV) 54
4.2.2 *K*-Fold Cross Validation (*K*-Fold CV) 56
4.3 The Modification of Weighted Robust Cross-Validation based on RMD-MRCD-PCA 57
4.4 Monte Carlo Simulation 61
4.5 Numerical Example 66
4.5.1 Octane Dataset 66
4.5.2 Biscuit Dough Dataset 69
4.6 Conclusion 73

5 AN IMPROVISED SIMPLS ESTIMATOR BASED ON THE RMD-MRCD-PCA WEIGHTING FUNCTION 74
5.1 Introduction 74
5.2 lassical SIMPLS 74
5.3 Robust Iteratively Reweighted SIMPLS (RWSIMPLS) 77
5.4 The Proposed Robust Weighted SIMPLS based on RMD-MRCD-PCA 80
5.5 The Proposed RMD-MRCD-PCA-RWSIMPLS Diagnostic Plot 82
5.6 Monte Carlo Simulation 85
5.7 Numerical Study 89
5.7.1 Gasoline Dataset 89
5.7.1.1 Influence Function 91
5.7.2 Octane Dataset 92
5.7.2.1 Influence Function 95
5.8 Conclusion 95

6 ROBUST CLASSIFICATION METHOD BASED ON THE INTERGRATION OF PLSDA AND ROBUST WEIGHTING FUNCTION OF RMD-MRCD-PCA 97
6.1 Introduction 97

6.2 Linear Discriminant Analysis (LDA) 97
6.3 Partial least Squares Discriminant Analysis 100
6.4 The Proposed Robust Weighted Partial Least Squares Discriminant Analysis based on MRCD-PCA 102
6.5 Monte Carlo Simulation Study 103
6.6 Numerical Study 114
6.6.1 Covid and Healthy Dataset 114
6.6.2 Colon Dataset 117
6.6.3 Coffee Dataset 119
6.7 Conclusion 122

7 ROBUST VARIABLE SELECTION METHOD BASED ON HOTELING T^2 AND RMD-MRCD-PCA-RWSIMPLS 123
7.1 Introduction 123
7.2 Hotelling T^2 for PLS 123
7.3 Chi-squared for PLS 125
7.4 The Proposed Hoteling T^2 Variable Selection based on RMD-MRCD-PCA-RWSIMPLS 127
7.5 Monte Carlo Simulation 131
7.6 Numerical Study 136
7.6.1 Gasoline Dataset 136
7.6.2 Fish Oil Dataset 138
7.7 Conclusion 140

8 SUMMARY, CONCLUSIONS AND RECOMMENDATIONS FOR FURTHER STUDIES 141
8.1 Introduction 141
8.2 Research Contribution 141
8.2.1 Identification of High Leverage Points in High Dimensional Data based on the Minimum Regularized Covariance Determinant and Principal Component Analysis 141
8.2.2 The Selection of Optimal Number of PLS Components based on the Combination of Cross-Validation and RMD-MRCD-PCA Weighting Function 142
8.2.3 An Improvised SIMPLS Estimator based on the RMD-MRCD-PCA Weighting Function 142
8.2.4 Robust Classification Method based on the Integration of PLSDA and Robust Weighting Function of RMD-MRCD-PCA 143
8.2.5 Robust Variable Selection Method based on Hoteling T^2 and RMD-MRCD-PCA- RWSIMPLS 143
8.3 Areas of Further Studies 143

REFERENCES 145
APPENDICES 160
BIODATA OF STUDENT 168
LIST OF PUBLICATIONS 169

LIST OF TABLES

Table		Page
2.1	Overview of the Literature Review for HLPs Detection in HDD	17
2.2	Overview of the Literature Review for Selecting the Optimal Number of PLS Components in HDD	20
2.3	Overview of the Literature Review for Partial Least Squares Regression in Linear Regression	24
2.4	Overview of the Literature Review for Classification of Outliers in High Dimensional Data	26
2.5	Overview of the Literature Review for Classification of Observations in High Dimensional Data	30
2.6	Overview of the Literature Review for Variable Selection Methods in HDD	33
3.1	Percentage of Correct Detection of HLPs, Masking and Swamping by RMD-MRCD, ROBPCA and RMD-MRCD-PCA, *n* = 50	42
3.2	Percentage of Correct Detection of HLPs, Masking and Swamping by RMD-MRCD, ROBPCA and RMD-MRCD-PCA, *n* = 100	43
4.1	The Number of Selected Optimal Number of PLS Components for 5% of HLPs Contamination	63
4.2	The Number of Selected Optimal Number of PLS Components for 10% of HLPs Contamination	64
4.3	The Number of Selected Optimal Number of PLS Components for 30% of HLPs Contamination	65
4.4	The RMSEP for Octane Data Set	67
4.5	The RMSEP Biscuit Dough Data Set	71
5.1	Mean Square Errors for Regression Parameter Estimates. Contamination in X Direction	86
5.2	Mean Square Errors for Regression Parameter Estimates. Contamination in X and Y Directions	87
6.1	Confusion Matrix between the Real Data Partition and the Predicted Partition	106
6.2	The Average of APER for Two Equal Groups	108

6.3 The Average of APER for Two Unequal Groups 109

6.4 The Average of APER for Three Equal Groups 112

6.5 The Average of APER for Three Unequal Groups 113

6.6 Confusion Matrix of RMD-MRCD-PCA-RWPLSDA for Covid-19 115

6.7 Confusion Matrix of SPLSDA for Covid-19 115

6.8 Confusion Matrix of PLSDA for Covid-19 115

6.9 Confusion Matrix of RMD-MRCD-PCA-RWPLSDA for Colon 118

6.10 Confusion Matrix of SPLSDA for Colon 118

6.11 Confusion Matrix of PLSDA for Colon 118

6.12 Confusion Matrix of RMD-MRCD-PCA-RWPLSDA for Coffee 120

6.13 Confusion Matrix of SPSLDA for Coffee 120

6.14 Confusion matrix of PLSDA for Coffee 120

7.1 The Selected Number of Variables and the Values of RMSEP 133

7.2 The Selected Number of Variables and the Values of RMSEP 134

7.3 The Selected Number of Variables and the Values of RMSEP 135

7.4 The Selected Number of Variables and the Values of RMSEP 136

7.5 The Selected Number of Variables and the Values of RMSEP 137

7.6 The Selected Number of Variables and the Values of RMSEP 139

LIST OF FIGURES

Figure **Page**

3.1 Percentage of Detection of HLPs, n = 50 for Graph (a) – (d), and n = 100 for Graph (e) – (h) 46

3.2 Index Plot of RMD-MRCD for Octane 47

3.3 Index Plot of ROBPCA for Octane 48

3.4 Index Plot of RMD-MRCD-PCA for Octane 48

3.5 Index Plot of RMD-MRCD for Biscuit Dough 49

3.6 Index Plot of ROBPCA for Biscuit Dough 50

3.7 Index Plot of RMD-MRCD-PCA for Biscuit Dough 50

3.8 Index Plot of the ROBPCA for Glass Spectra 51

3.9 Index Plot of the RMD-MRCD-PCA for Glass Spectra 51

4.1 Scree Plot of RMSEP curves for the Octane Data Set 69

4.2 Scree Plot of RMSEP Curves for the Biscuit Dough Data Set 73

5.1 Robust Standardized Residuals Versus Robust Score Distance of RMD-MRCD-PCA-RWSIMPLS Plot 83

5.2 Diagnostic Plot of RMD-MRCD-PCA-RWSIMPLS for Gasoline 90

5.3 Diagnostic Plot of RWSIMPLS for Gasoline 90

5.4 Diagnostic Plot of SIMPLS for Gasoline 91

5.5 Empirical Influence Functions of SIMPLS (red line), RWSIMPLS (blue line), and RMD-MRCD-PCA-RWSIMPLS (green line) for Gasoline 92

5.6 Diagnostic Plot of RMD-MRCD-PCA-RWSIMPLS for Octane 93

5.7 Diagnostic Plot of RWSIMPLS for Octane 94

5.8 Diagnostic Plot of SIMPLS for Octane 94

5.9 Empirical Influence Functions of SIMPLS (red line), RWSIMPLS (blue line), and RMD-MRCD-PCA-RWSIMPLS (green line) for Octane 95

6.1 Plots of Individual of RMD-MRCD-PCA-RWPLSDA on Covid-19 116

6.2 Plots of Individual of SPLSDA on Covid-19 116

6.3 Plots of Individual of PLSDA on Covid-19 117

6.4 Plots of Individual of RMD-MRCD-PCA-RWPLSDA on Colon 118

6.5 Plots of Individual of SPLSDA on Colon 119

6.6 Plots of Individual of PLSDA on Colon 119

6.7 Plots of Individual of RMD-MRCD-PCA-RWPLSDA on Coffee 121

6.8 Plots of Individual of SPLSDA on Coffee 121

6.9 Plots of Individual of PLSDA on Coffee 122

7.1 Variable Selection Plot for T^2-RMD-MRCD-PCA-RWSIMPLS Method 137

7.2 Variable Selection Plot for T^2-PLS Method 137

7.3 Variable Selection Plot for $\chi^2 - \mathsf{PLS}$ Method 138

7.4 Variable Selection Plot for T^2-RMD-MRCD-PCA-RWSIMPLS 139

7.5 Variable Selection Plot for T^2-PLS 139

7.6 Variable Selection Plot for $\chi^2 - \mathsf{PLS}$ 140

LIST OF ABBREVIATIONS

CCPLS	Canonical Correlation Partial Least Squares
CV	Cross - Validation
DA	Discriminant analysis
DLDA	Deterministic Minimum Covariance Determinant
DNA	Deoxyribonucleic Acid
GA	Genetic Algorithm
HDD	High Dimensional Data
HDRDA	High Dimensional Regularized Discriminant Analysis
HLPs	High Leverage Points
IPW	Iterative Predictor Weighting
IRPLS	Iteratively Reweighted Partial Least Squares
IRWPLS	Iteratively Weighted Partial Least Squares
LDA	Linear Discriminant Analysis
LOOCV	Leave-One-Out Cross Validation
LW	Loading Weight
MCCV	Monte Carlo Cross Validation
MCD	Minimum Covariance Determinant
MDP	Minimum Diagonal Product
MRCD	Minimum Regularized Covariance Determinant
MRI	Magnetic Resonance Imaging
MVE	Minimum Volume Ellipsoid
MWCD	Minimum Weighted Covariance Determinant
NIPALS	Nonlinear Estimation by Iterative Partial Least Squares
OD	Orthogonal Distance
OLS	Ordinary Least Square

PCA	Principal Component Analysis
PLS	Partial Least Squares
PLD-DA	Partial Least Squares Discriminant Analysis
PRM	Partial Robust M-Regression
QDA	Quadratic Discriminant analysis
RDA	Regularized Discriminant Analysis
rdCV	Repeated Double Cross Validation
R-MCD	Regularized Minimum Covariance Determinant
RMD	Robust Mahalanobis Distance
RMT	Random Matrix Theory
SCRDA	Shrunken Centroids Regularized Discriminant analysis
SD	Score Distance
SDE	Stahel Donoho Estimator
SIMPLS	Statistically Inspired Modification of the Partial Least Squares
sMC	Significance Multivariate Correlation
SPLS	Sparse Partial Least Squares
SPLSDA	Sparse Partial Least Squares Discriminant Analysis
SR	Selectivity Ratio
SS-PLS	Spatial Sign Partial Least Squares
SwPA-PLS	Sub-Window Permutation Analysis Partial Least Squares
UVE-PLS	Uninformative Variable elimination Partial Least Squares

CHAPTER 1

INTRODUCTION

1.1 Background and Purpose

The rapid development of computer technology and statistics has contributed to the generation of high-dimensional data (HDD). Some examples of HDD are images, curves, or movies such that a single observation has a dimension in hundreds, thousands, or even millions of variables to consider. A wide range of recent research has focused on high-dimensional data sets, for instance, in gene analyses, millions of genes are measured for a single individual (Boulesteix & Strimmer, 2007) while an image analysis contains thousands of resolution images in pixels with a small number of samples (Cao, 2006). Other varied applications in high dimensional data are in chemometrics, fraud detection, climate studies, geography, and satellite processing. In other words, high-dimensional data set signifies data with $p >> n$. This characteristic leads to a sparsity problem known as the curse of dimensionality phenomena (Mehmood & Ahmed (2016) and Bolon-Canedo et al. (2015)). In the curse of high dimensionality, conventional statistical methods do not work well, and most of them fail to perform especially when dealing with contaminated data or outliers.

Several versions of outliers exist in regression problems, such as vertical outliers, residual outliers, and high leverage points. Residual outliers refer to any observations that have a large residual while vertical outliers refer to observations that are extreme or are outlying in the *Y*-space. On the other hand, high leverage points (HLPs) refer to observations which fall far from the majority of the explanatory variables or are outlying in the *X*-space. HLPs show an abnormal behavior in the system and often bring or contain meaningful information to data (Yu & Aggarwal, 2001). The detection of high leverage points is very crucial, for example in a microarray data analysis to spot a malignant tumor in an MRI scan (Foss, 2010), in a business analysis to identify unforeseen models in a network traffic to indicate a hack (Agarwal & Mittal, 2012), and in classifying fraud detection in credit card transactions (Porwal & Mukund, 2019). These problems frequently arise in very high dimensional data sets. Therefore, high leverage points detection in HDD has become an issue of great importance to study although it has not received significant attention from statisticians. Accurate identification of high leverage points plays an important role in statistical analysis as incorrect detection of high leverage points will substantially affect the standard error of estimates and cause multicollinearity problems, masking and/or swamping of outliers, overfitting and/or underfitting of a model which will lead to a false prediction. Moreover, HLPs are responsible for detrimental effects on various statistical analysis such as in parameter estimation, classification and variable selection methods. This will lead to misleading conclusions and inaccurate predictions. Therefore, in this thesis, several alternatives robust methods have been developed to deal with HLPs. Thus, a few basic concepts and some commonly used methods need to be introduced in the following sections.

1.2 Mahalanobis Distance (MD)

Mahalanobis Distance (Mahalanobis, 1936) is broadly used in a multivariate analysis to measure the gap between two individual points with several variables (Varmuza and Filzmoser, 2009). Let $i = 1, 2, \ldots, n$, the i^{th} vector of explanatory variables can be written as $x_i = (1, x_{i1}, x_{i2}, \ldots, x_{ip}) = (1, v_{ip})$, where v_{ip} is a p-dimensional row vector. The mean vector $\bar{v}$ is calculated as, $\bar{v} = \frac{1}{n}\sum_{i=1}^{n} v_i$ and the variance covariance matrix $\hat{\Sigma}$ is computed as: $\hat{\Sigma} = \left(\frac{1}{n-1}\right)\sum_{i=1}^{n}(v_i - \bar{v})^T (v_i - \bar{v})$.

Then, the Mahalanobis distance for each observation is defined as

$$MD_i = \sqrt{(v_i - \bar{v})\hat{\Sigma}^{-1}(v_i - \bar{v})^T}\ , i = 1, 2, 3, \ldots, n \qquad (1.1)$$

And Mahalanobis (squared) distance is given by,

$$MD_i^2 = (v_i - \bar{v})\hat{\Sigma}^{-1}(v_i - \bar{v})^T\ , i = 1, 2, 3, \ldots, n$$

1.3 Minimum Covariance Determinant (MCD)

Rousseeuw (1985) introduced the Minimum Covariance Determinant (MCD), which is a robust multivariate estimator. These location and scatter estimates are very robust and affine equivariant.

The MCD method is widely used by statistical practitioners, but it is not computationally efficient. To overcome this drawback, Rousseeuw and Driessen (1999) developed the Fast-Minimum Covariance Determinant algorithm.

The main purpose of MCD is to find a subset of $h = 1,2,\ldots, h$ data points with the smallest determinant of the covariance matrix, where $\frac{n}{2} \le h \le n$. A robust location estimator $\hat{\mu}_{mcd}$ is an average of these h data points, while a robust scatter estimator, $\hat{\Sigma}_{mcd}^{-1}$, is a covariance matrix of h data points multiplied by consistency factor, c_α. Croux and Haesbroeck (1999) noted that c_α equals $\frac{\alpha}{F_{\chi^2_{p\,2}}(x^2_{p,2})}$ where $\alpha = \frac{n-h}{n}$; p is a dimension of data set; and $h = 0.75n$ or $h = (n + p + 1)/2$. The asymptotic efficiency of MCD is almost 50% when $h = 0.75n$ or $h = (n + p + 1)/2$. Hubert et al. (2018) stated that the MCD location and scatter estimator have the highest breakdown values and thus become more robust. MCD estimators can be determined when $h > p$; otherwise, the covariance matrix will be singular, and the determinant of the covariance matrix will be zero. The principal rule of thumb suggested by Rousseeuw et al. (1990) and Hubert et al. (2012) is to avoid the curse of dimensionality, and it requires $n > 5p$

1.4 Minimum Volume Ellipsoid (MVE)

The minimum volume ellipsoid (MVE) is another robust multivariate estimator. Rousseeuw (1985) pointed out that the center of the minimal volume ellipsoid covers at least half of the *h* points of *X*, where *h* can be taken equal to $\left[\frac{n}{2}\right] + 1$. The ellipsoid can be used as the corresponding covariance predictor. In most applications it is not feasible to consider all "halves" of data. Habshah, et al. (2009) revealed that the calculation of MVE can be started by drawing a sub sample of (*p* + 1) different observations, indexed by $J = (i_1, i_2, \dots i_p)$. Then the mean and covariance matrix are determined, respectively, given by:

$$\bar{x}_j = \frac{1}{p+1}\sum_{i \in J} x_i \quad \text{and} \quad C_j = \frac{1}{p}\sum_{i \in J}(x_i - \bar{x})^T (x_i - \bar{x})$$

where C_j is non-singular. The corresponding ellipsoid should then be inflated or deflated to contain exactly *h* points, which corresponds to compute $m_j^2 = Med(x_i - \bar{x}_J)C_J^{-1}(x_i - \bar{x}_J)^T$. The volume of the resulting ellipsoid, corresponding to $m_J^2 C_J$ is proportional to $\sqrt{\det(m_J^2 C_J)} = \sqrt{\det(C_J)}(m_J)^P$. It is repeated for many *J* so that the above determinant becomes the minimum and its corresponding values yield:

$T(X) = \bar{x}_J$ and $C(X) = (\chi^2_{p,0.5})^{-1} m_J^2 C_J$, where $\chi^2_{p,0.5}$ is the median of the Chi-squared distribution with *p* degrees of freedom. This correction factor is required to attain the consistency for multivariate normal data.

1.5 Minimum Regularized Covariance Determinant (MRCD)

There is a primary constraint in the MCD system to be applied to high dimensional data. For the MCD, the criterion of dimension *p* must satisfy *p* < *h* for any *h*-subset. This requirement must be fulfilled in order to obtain a non-singular covariance matrix. An improvement to the MCD is of high importance to make it work for high-dimension data.

Boudt et al. (2018) formulated a new modification of the MCD, the so-called minimum regularized covariance determinant (MRCD). The fundamental objective of the MRCD is to substitute a regularized covariance estimate for the MCD subset-based covariance. *H*-subset of MRCD that minimizes the determinant of regularized covariance of MRCD, $K(H)$ is as shown below,

$$H_{mrcd} = \arg\min_{H \in \mathrm{H}_h} \left(\det K(H)\right)^{1/p}$$

where H_h is set of all H subsets. *K*(*H*) represents a regularized covariance matrix in MRCD and can be written as

$$K(H) = \rho T + (1-\rho) c_\alpha S(H)$$

where $S(H) = h^{-1}(X_H - m(H))^T (X_H - m(H))$ is a sample covariance estimate based on subset H with $h \times p$ submatrix of X and can be denoted as X_H. The mean estimate of subset H is $m(H) = h^{-1} X_H{}^T$ with $[(n+p+1)/2] \le h \le n$ where *h* is the number of observations in subset H .

T is a predetermined, symmetric, and positive definite target matrix. In other words, assume $T = I_p$ where I_p is an identity matrix with dimension $p \times p$. The identity matrix is used as a target matrix because it has very good statistical properties and is well-conditioned in high dimensional data sets**.** $\rho = (0,1]$ is a regularization intensity parameter and the consistency factor, c_α is to obtain the consistency at the normal multivariate distribution and unbiased at small sample. The consistency factor used in the MRCD method is the same as the one used in the MCD procedure. The value of ρ is set such that *K(H)* is well-conditioned such that $\frac{\lambda_{\max}}{\lambda_{\min}} \le 1000$, where $\lambda_{\max}$ and $\lambda_{\min}$ are the largest and the smallest eigenvalues of $c_\alpha S(H)$, respectively. The eigenvalue of MRCD covariance is equals to $\rho + (1-\rho)\lambda$, and the regularization is used when needed. Then C-steps is applied until the estimated MRCD covariance converges. The C-step (Boudt et al., 2018) of MCD has been generalized to regularized covariance matrices in MRCD method. In MCD, The C-step theorem computes the mean and the covariance matrix of *h*-subset H_1 , and then puts the observations with smallest Mahalanobis distance in a new subset H_2 . The C-step theorem proves that the covariance determinant of H_2 is less than or equal to H_1 . The generalized C-step theorem is discussed at length in this section.

The generalized C-steps (Boudt et al., 2018) theorem is summarized as follows:

Starting from an h-subset H_1 *compute* $m_1=\frac{1}{h}\sum_{i\in H_1} x_i$ *and* $S_1=\frac{1}{h}\sum_{i\in H_i}(x_i-m_i)^T(x_i-m_i)$*. The matrix* $K_1=\rho T+(1-\rho)\,c_\alpha S_1$ *is positive definite, hence invertible. Then compute,*

$$d_1(i)=(x_i-m_1)K_1^{-1}(x_i-m_1)^T \qquad i=1,\dots,n\,.$$

Let H_2 *be an h-subset for which,*

$$\sum_{i\in H_2} d_1(i)\le\sum_{i\in H_1} d_1(i)$$

And compute $m_2=\frac{1}{h}\sum_{i\in H_2} x_i$, $S_2=\frac{1}{h}\sum_{i\in H_2}(x_i-m_2)^T(x_i-m_2)$ *and* $K_2=\rho T+(1-\rho)S_2$*. The,*

$$\det(K_2)\le\det(K_1),$$

With equality if and only if $m_1=m_2$ *and* $K_1=K_2$. The iteration process in the C-steps theorem stop if $\det(K_2)=0$ or $\det(K_2)=\det(K_1)$; otherwise the iteration process continues until it converges.

By using this generalized C-steps, the MRCD regularized covariance matrix is constructed to find the MRCD subset.

1.6 Principal Component Analysis

Principal Component Analysis (PCA) is an unsupervised dimension reduction procedure, and it is commonly used in multiple linear regression analyses. It was first established by Pearson (1901) before it becomes a prevalent method in many fields such as chemometrics, engineering, computer vision, face recognition Li et al. (2016) and other domains such as in gene selection and tumor classification as illustrated by Feng et al. (2019), and feature selection as shown by Hu et al. (2019). PCA analysis aims to find a small number of linear combinations of the predictors that can be used to summarize data without losing too much information. This statistical method transforms a broad set of correlated variables into a smaller number of uncorrelated factors called principal components. These orthogonal principal components solve the multicollinearity problem.

1.6.1 Data Matrix

Let X be $n \times p$ matrix where n is a number of observations, $i = 1,2,3,...,n$, and p is a set of predictive variables $x_{i1}, x_{i2}, ..., x_{ip}$. The set of X represent a random observation of $x_{i1}, x_{i2}, \ldots, x_{ip}$. The issue is selecting a subset of the predictive variables which contains most information of the data set.

1.6.2 Variance-Covariance Matrix

The PCA procedure involves the covariance structure of data to obtain the values of eigenvalues and eigenvectors. Let σ_{ij} denotes the covariance between X_i and X_j in the data matrix. The matrix σ_{ij} is written as $\sum$ and known as a variance-covariance matrix. The values of the diagonal element in matrix $\sum$ are the variances of X_i. This covariance matrix $\sum$ must be a symmetric and a square matrix.

1.6.3 Linear Combination

The principal components (PCs) of PCA are the linear combination of the original variables, $x_{i1}, x_{i2}, ..., x_{ip}$, that can be expressed as $\sum a_i x_i$, where a_i is the scalar. A principal component is a normalized linear combination of the original predictors in a data set if $\sum |a_i| = 1$.

1.6.4 Linear Independence

The PCA method successfully solves the problem of correlated variables by choosing a set of orthogonal principal components (PCs). This is a potent mathematical property. All the principal components are independent of each other. The first principal component, PC1, is uncorrelated to PC2, and PC2 and PC3 are uncorrelated and so on. It can be proven by computing the product of $b_i^T b_j = 0$, where *i*, *j* = 1, 2, 3, …, *p*. The linear combination of variables for PCs is formulated as $t_i = x_{i1}b_1 + x_{i2}b_2 + ... + x_{ip}b_p$, where *i* = 1, 2, 3, …, *n*, and $b = b_1, b_2, ..., b_p$ are the loading vectors, and p is the number of predictor variables.

1.7 Partial Least Squares (PLS)

Partial least squares (PLS) regression is a supervised multivariate technique to address the problem of multicollinearity in HDD. The general idea of PLS is to extract a set of orthogonal latent variables that account for most of the variation measured by the original variables to represent the data. These orthogonal latent variables are also known as PLS components and are very crucial for developing the best predicted PLS model. More formally, the PLS method is to relate a matrix *X* to a vector *y* or matrix *Y*, through a linear relationship, $Y = X\beta + \varepsilon$, which maximizes the covariance between the predictor and response variables. The predictor and response variable in PLS can be modelled as $X = TP^T + \varepsilon_X$ and $Y = UQ^T + \varepsilon_Y$, where ε_X and ε_Y are error matrices, T and U are score matrices of *X* and *Y*, respectively. P and Q are the loading matrices of *X* and *Y*, respectively. PLS has been shown to be an extremely versatile method for the analysis of multivariate data, and the number of applications is rapidly growing. Numerous extensions of PLS have been designed to address multivariate issues, such as regression, classification and variable selection. However, it is now evident that PLS is not robust in the sense that its estimate is easily affected by outliers and high leverage points.

1.8 Motivation of the Study

Mahalanobis distance is a popular technique to detect outliers in multivariate dataset by measuring the distance of an observation from a location with respect to a shape in the sample. However, the classical location and shape of Mahalanobis distance are not robust to outliers. Rousseeuw (1985) proposed a new robust method named minimum volume ellipsoid (MVE) to replace the classical location and scatter of Mahalanobis distance, but it is a highly biased-robust method. Minimum covariance determinant (MCD) is the first affine equivariant and highly robust estimators of multivariate location and scatter (Rousseeuw, 1985). It is very resistant to outlying observations that makes the MCD highly effective for outlier detection. The MCD was first introduced in 1984, but its main application started since the development of FastMCD by Rousseeuw & Driessen, (1999). The FastMCD involves the calculation of covariance matrix and its determinant based on *h*-subset of observations. Then the method was improvised by giving a zero-one weighting function to improve the efficiency of the FastMCD (Lopuhaä & Rousseeuw (1991) and (Lopuhaä, 1999).

The robust FastMCD location and scatter estimates replace the classical estimates of Mahalanobis distance (Mahalanobis, 1936) thus it is known as robust Mahalanobis distance. There are many further research has been done in improving the MCD procedure for outlier detection such as minimum weighted covariance determinant (FAST-MWCD) by Roelant et al. (2009) and deterministic MCD (DetMCD) by Hubert et al. (2012). However, none of the approaches discussed are applicable to p >> n scenario, as the covariance matrix of the MCD is not invertible in high dimension cases. Therefore, Boudt et al. (2018) proposed a minimum regularized covariance determinant (MRCD) to overcome the curse of dimensionality issue. Afterwards, the robust Mahalanobis distance which is based on the MRCD (RMD-MRCD) is introduced.

However, the RMD-MRCD method indicates a decrease in its performance as the number of independent variables (p) increases. This shortcoming has motivated us to propose a robust Mahalanobis distance (RMD) based on the combined methods of the minimum regularized covariance determinant and the principal component analysis (RMD-MRCD-PCA). The RMD-MRCD-PCA is developed by incorporating the PCA method in the MRCD algorithm whereby this robust approach shrinks the covariance matrix to make it invertible and thus, can be employed to compute the RMD for high dimensional data.

Multivariate partial least squares (PLS) regression is a very useful tool for the analysis of high-dimensional data. Choosing the number of PLS components (also known as latent variables or factors) is a vital step in developing the best model. Preserving too few components means that the calibration data are not well-fitted, and there is still information left that can be modelled and over-fitting results from selecting too many components. This means that although the calibration data are accurately defined, the model's ability to predict future samples will be subpar. Cross-validation (CV) is the most widely used method to find the PLS components. The CV is basically a leave-one-out cross validation (LOOCV). It evaluates the prediction power of the predicted models according to the number of components included in the model. Nevertheless, when dealing with huge datasets, LOOCV can be computationally intensive and time consuming. This is due to the fact that LOOCV fits the model iteratively on the entire training set. The other problem with LOOCV is that it can be prone to high variance or overfitting, which means that it requires almost all of the training data to learn and only a single observation to evaluate. K-Fold cross validation (K-FoldCV) is designed to address the drawbacks of LOOCV. In K-FoldCV, the data are split into k randomly equal folds or groups. Then, in k different iterations, each of these folds is treated as a validation set. Both CVs stated above are evaluated using the classical mean square error (MSE), which is easily influenced by outliers and high leverage points. Thus, these weaknesses have inspired us to establish a robust weighted RMD-MRCD-PCA-LOOCV and robust weighted RMD-MRCD-PCA-K-FoldCV to determine the optimal number of PLS components.

Multicollinearity often occurs when two or more predictor variables are correlated, especially for high dimensional data (HDD) where p>>n. The statistically inspired modification of the partial least squares (SIMPLS) is a very popular technique for solving a partial least squares regression problem due to its efficiency, speed, and ease of understanding. The execution of SIMPLS is based on the empirical covariance matrix of explanatory variables and response variables. Nevertheless, SIMPLS is very easily affected by outliers. In order to rectify this problem, a robust iteratively reweighted SIMPLS (RWSIMPLS) is introduced. Nonetheless, it is still not very efficient as the algorithm of RWSIMPLS is based on a weighting function that does not specify any method of identification of high leverage points (HLPs), i.e., outlying observations in the X-direction. HLPs have the most detrimental effect on the computed values of various estimates, which results in misleading conclusions about the fitted regression model. Hence, their effects need to be reduced by assigning smaller weights to them. As a solution to this problem, we propose an improvised SIMPLS based on a new weight function obtained from the RMD-MRCD-PCA diagnostic method of the identification of HLPs for HDD and name this method RMD-MRCD-PCA-RWSIMPLS.

This thesis also concerns constructing diagnostic plots to group observations into four types of data points, such as regular observations, good leverage points, vertical outliers, and bad leverage points. It is very important to classify the outliers into its correct category. For instance, good leverage points are not removed because removing the good leverage points decreases the efficiency of the estimator. The ROBPCA-diagnostic plot was the earliest graphical plot developed for HDD. It plots the outliers based on the score and the orthogonal distances. However, there are some drawbacks in ROBPCA method as the method suffers from swamping effect. Moreover, the classical Chi-squared which is easily influenced by the existence of outliers or HLPs was used as a cut-off point. Another diagnostic plot for HDD is the RWSIMPLS-plot to classify observations into four types of data points by plotting RWSIMPLS standardized residuals versus leverage values, i.e., the diagonal elements of the hat matrix that is used to identify HLPs. Any observation that corresponds to a leverage value larger than the cut-off point, i.e., 2p/n, is considered as HLPs. Using leverage values as diagnostic measure of detecting HLPs will produce unsatisfactory results. These limitations have inspired us to construct a new RMD-MRCD-PCA-RWSIMPLS diagnostic plot to classify observations into four data points, i.e., regular observations, vertical outliers, and good and bad leverage points.

This thesis also addresses the issue of classification of observations in the HDD. Linear discriminant analysis (LDA) is the most often applied classification rule under normality. In LDA, a separate covariance matrix is estimated for each group. However, if there are more variables than observations in a group, the typical covariance estimates are singular and can no longer be employed. Numerous improvements and regularizations of the LDA approach have been done to address the dimensionality issue. The earliest is the regularized discriminant analysis developed by Friedman in 1989 by adding a regularization parameter to minimize the classification rate. Kemsley (1996) suggested a combination of PLS and PCA with the LDA. Yu & Yang (2001) constructed a direct LDA (DLDA) by diagonalizing the two variance matrices (between-class variance and within-class variance) simultaneously. However, the LDA method possesses less classification accuracy (Ding & Gentleman, 2005). PLSDA was developed by Barker & Rayens (2003) and extended by Nocairi et al. (2005), then, the method was improved in 2007 by Indahl et al. The discriminant score is determined based on PLSDA scores. Nevertheless, Aminu & Ahmad (2020) claimed that the PLSDA method has no significant advantage over classical method for low dimension cases. It provides similar results as the classical approach of LDA. Furthermore, PLSDA is highly sensitive to the presence of outliers. Cao et al. (2011) developed a sparse version of the PLS which is an extension of the sparse partial least squares (SPLS) proposed by Cao et al. (2008). The method employs the sparse element (Lasso parameter) to reduce the dataset's dimension and only uses nonzero variables in the classification process. However, the Lasso parameter fails to locate HLPs and diminishes their effect. As a result, they cause the SPLSDA to compute a high misclassification rate. Moreover, the main purpose of the development of the SPLSDA is for variable selection procedures. Furthermore, the SPLSDA is not resistant to HLPs. Their work has inspired us to develop a robust weighted partial least squares discriminant analysis as a novel robust weighted classifier (RMD-MRCD-PCA-RWPLSDA) to remedy the problem of classification for high dimensional contaminated dataset. To the best of our knowledge, no one has attempted to develop such a method.

This thesis also focusses on the issue of variable selection in HDD. In high-dimensional data, the number of predictor variables is typically more extensive than the number of observations and is frequently correlated. However, it is assumed that only a small subset of these variables is relevant to the model. Therefore, choosing informative variables to be included in model building is critical, as an appropriate variable selection improves prediction and model understanding. Partial least squares (PLS) based variable selection is the most common approach to deal with this issue. Frank et al. (1987) proposed a backward variable elimination in PLS but there is no clear explanation how to group the explanatory variables before the PLS is fitted. Centner et al. (1996) suggested an uninformative variable elimination in PLS by adding artificial outlying points to the dataset whereby variables that are less influence than noisy variables are eliminated. However, adding artificial outlying variables could influence the model if outlying variables are not properly selected. Several other methods are developed to solve the issue of variable selection in the model such as genetic algorithm in PLS by Hasegawa et al. (1997) and iterative predictor weighting PLS by Forina et al. in (1999). However, when there are many predictor variables in a dataset, both approaches require a lot of time and effort. To remedy this problem, Mehmood (2016) proposed a Hotelling T^2 variable selection based on PLS (T^2-PLS) which is less complicated and easier to compute. Nevertheless, the introduced method is NIPALS-based PLS, and the estimators rely on the classical mean and classical covariance, which are highly affected by the outlying observations. These weaknesses have motivated us to develop a new variable selection method by integrating a new weight function from RMD-MRCD-PCA in its establishment and call this method a robust Hotelling T^2 denoted as T^2-RMD-MRCD-PCA-RWSIMPLS.

1.9 Objective of Thesis

The objectives of this research are as follows:

1. To extend a method for the identification of high leverage points in high dimensional data by incorporating the PCA approach and the RMD-MRCD method.
2. To formulate an efficient selection method of determining an optimal number of PLS latent variables by integrating the cross-validations approaches with a robust weighting function obtained from the RMD-MRCD-PCA.
3. To modify the existing robust partial least squares estimation method via the incorporation of the RMD-MRCD-PCA weighting function and to extend a diagnostic method of outliers classification in high dimensional data by employing the robust weight of the RMD-MRCD-PCA and the SIMPLS technique.
4. To improve a robust classification method based on the integration of PLSDA and robust weighting function of RMD-MRCD-PCA in the presence of high leverage points for binary and multiple classes.
5. To extend a robust Hotelling T^2 variables selection method by incorporating a new weight function constructed from a newly developed method of identification of outliers in HDD (RMD-MRCD-PCA) and robust SIMPLS (RWSIMPLS).

1.10 Scope and Limitation of Study

The analysis of high dimensional data has become increasingly important in many disciplines such as chemometrics, climate studies, fraud detection, geography and satellite processing. It forms major statistical challenges as such data leads to sparsity problem known as curse of dimensionality phenomena. With the increase of dimensionality and in the presence of outliers, the analysis becomes very complicated, time consuming and computationally intense. Hence the scope of this thesis concentrates on the establishment of several robust methods in HDD.

The detection of high leverage points in high-dimensional data is crucial, particularly because the high dimensions of HDD lead to a sparsity problem known as the curse of dimensionality. The occurrence of outliers in the *X* direction, as well as multicollinearity among the predictors, are the two key concerns with this sparsity problem. However, no research focusing on the identification of HLPs has been conducted. Therefore, there are not many real data sets available in high dimensions that are appropriate for our research.

The simulation experiments for HDD require a significant amount of computer time and can be expensive. Thus, the Monte Carlo simulations of the proposed methods are conducted using limited dimensions of the dataset and due to an inadequately high performing computer, some of our experiments were only repeated for 500 iterations.

1.11 Outline of the Thesis

Following the objectives and scopes of study, the contents of this thesis are organized into eight chapters. The arrangement of thesis chapters is designed to align with each objective, ensuring a coherent sequence in the outline.

Chapter Two discusses the literature reviews on outliers and high leverage points detection techniques for low and high dimension cases, the weighting functions and the optimum number of latent variables selection methods in PLS and the types of PLS methods are also reviewed. The classification methods for high dimension situations are deliberated. Finally, the variable selection methods are discussed.

Chapter Three discusses the high leverage points detection based on the MRCD-PCA estimators of location and scatter matrix. The proposed method is the integration of reduction method PCA and the robust Mahalanobis distance based on MRCD (RMD-MRCD-PCA). The RMD-MRCD-PCA algorithm is presented. Monte Carlo simulation and real data examples were used to demonstrate the performance of the proposed robust Mahalanobis distance based on the MRCD-PCA estimator. The existing methods of RMD-MRCD and the ROBPCA are used for comparison. The three real data examples are Octane data, Biscuit dough and Glass spectra dataset.

Chapter Four discusses the proposed optimal number of latent variable selection in PLS based on the weight of RMD-MRCD-PCA and the cross validations (LOOCV and *K*-FoldCV) procedures. The MRCD-PCA-LOOCV and MRCD-PCA-*K*-FoldCV algorithms are presented. This proposed method is evaluated using Monte Carlo simulations and 2 real life data (Fish and Biscuit dough dataset).

Chapter Five discusses the new improvised robust SIMPLS method based on the weighting function of RMD-MRCD-PCA (RMD-MRCD-PCA-RWSIMPLS). Monte Carlo simulation and two real datasets are used to evaluate the performance of the diagnostic algorithm (RMD-MRCD-PCA-RWSIMPLS) and the existing methods, i.e., classical SIMPLS and RWSIMPLS. The real datasets are Octane and Gasoline.

Chapter Six discusses the new robust classification procedure based on the weight of RMD-MRCD-PCA and partial least squares discriminant analysis (PLSDA) under HLPs contamination dataset. The RMD-MRCD-PCA-RWPLSDA is evaluated on the Monte Carlo simulation and three real datasets (Covid 19, Colon and Coffee)

Chapter Seven discusses the new robust Hotelling T^2 based on the RMD-MRCD-PCA-RWSIMPLS variable selection approach. The proposed T^2-RMD-MRCD-PCA-RWSIMPLS is compared with other three existing methods T^2-PLS, Chi-PLS and UVE-PLS by using Monte Carlo simulation methods and two real datasets (Gasoline and Fish)

Chapter Eight provides the summary, conclusions, recommendations, and possible future research areas.

REFERENCES

Agarwal, B., & Mittal, N. (2012). *Hybrid Approach for Detection of Anomaly Network Traffic using Data Mining Techniques*. Procedia Technology, 6, 996–1003. https://doi.org/10.1016/j.protcy.2012.10.121

Agostinelli, C., Leung, A., Yohai, V. J., & Zamar, R. H. (2015). *Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination*. Test, 24(3), 441–461. https://doi.org/10.1007/s11749-015-0450-6

Akaike, H. (1969). *Statistical predictor identification*. ann. inst. statist. math., (1), 203–217.

Aylin, A., & Agostinelli, C. (2017). *Robust iteratively reweighted SIMPLS*. Journal of Chemometrics, 31(3), 1–9. https://doi.org/10.1002/cem.2881

Alin, A., & Ali, M. M. (2012). *Improve straightforward implementation of a statistically inspired modification of the partial least squares algorithm*. Pakistan Journal of Statistics, 28(2), 217–229. Retrieved from https://www.rics.org/south-asia/upholding-professional-standards/standards-of-conduct/ethics/

Allen, D. M. (1971). *Mean square error of prediction as a criterion for selecting variables*. Technometrics, 13(3), 469–475. https://doi.org/10.1080/00401706.1971.10488811

Alma, Ö. G. (2013). *Performance comparisons of model selection criteria : aic , bic , icomp and wold ' s for plsr*, (December), 15–34.

Almeida, M. R., Correa, D. N., Rocha, W. F. C., Scafi, F. J. O., & Poppi, R. J. (2013). *Discrimination between authentic and counterfeit banknotes using raman spectroscopy and PLS-DA with uncertainty estimation*. Microchemical Journal, 109, 170–177. https://doi.org/10.1016/j.microc.2012.03.006

Alon, U., Barka, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., & Levine, A. J. (1999). *Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays*. Proceedings of the National Academy of Sciences of the United States of America. https://doi.org/10.1073/pnas.96.12.6745

Aminu, M., & Ahmad, N. A. (2020). *Complex Chemical Data Classification and Discrimination Using Locality Preserving Partial Least Squares Discriminant Analysis*. ACS Omega, 5(41), 26601–26610. https://doi.org/10.1021/acsomega.0c03362

Andries, J. P. M., Vander Heyden, Y., & Buydens, L. M. C. (2017). *Improved variable reduction in partial least squares modelling by Global-Minimum Error Uninformative-Variable Elimination*. Analytica Chimica Acta, 982(2017), 37–47. https://doi.org/10.1016/j.aca.2017.06.001

Asuman Seda Turkmen. (2008). *Robust Partial Least Squares For Regression and Classification*. Auburn University.

Baba, A. M., & Midi, H. (2022). *Spatial Outlier Accommodation Using a Spatial Variance Shift Outlier Model*, 1–19.

Bagheri, A., Habshah, M., & Imon, R. H. M. R. (2012). *A novel collinearity-influential observation diagnostic measure based on a group deletion approach*. Communications in Statistics: Simulation and Computation, 41(8), 1379–1396. https://doi.org/10.1080/03610918.2011.600497

Barker, M., & Rayens, W. (2003). *Partial least squares for discrimination*. Journal of Chemometrics, 17(3), 166–173. https://doi.org/10.1002/cem.785

Bin, J., Ai, F., Fan, W., Zhou, J., Li, X., Tang, W., & Liang, Y. (2016). *An efficient variable selection method based on variable permutation and model population analysis for multivariate calibration of NIR spectra*. Chemometrics and Intelligent Laboratory Systems (Vol. 158). https://doi.org/10.1016/j.chemolab.2016.08.006

Bolon-Canedo, V., Sanchez-Marono, N., & Alonso-Betanzos, A. (2015). *Feature Selection for High-Dimensional Data*. Springer. https://doi.org/10.1007/978-3-319-21858-8

Boudt, K., Rousseeuw, P. J., Vanduffel, S., & Verdonck, T. (2018). *The minimum regularized covariance determinant estimator*. Statistics and Computing. https://doi.org/10.1007/s11222-019-09869-x

Boulesteix, A.-L. (2004). *PLS dimension reduction for classification of microarray data.*

Boulesteix, A. L., & Strimmer, K. (2007). *Partial least squares: A versatile tool for the analysis of high-dimensional genomic data*. Briefings in Bioinformatics, 8(1), 32–44. https://doi.org/10.1093/bib/bbl016

Bouwmans, T., Javed, S., Zhang, H., Lin, Z., & Otazo, R. (2018). *On the Applications of Robust PCA in Image and Video Processing*. *Proceedings of the IEEE, 106*(8), 1427–1457. https://doi.org/10.1109/JPROC.2018.2853589

Bozdogan, H. (1987). *Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions*. Psychometrika, 52(3), 345–370. https://doi.org/10.1007/BF02294361

Branden, K., & Hubert, M. (2004). *Robustness properties of a robust partial least squares regression method*. Analytica Chimica Acta, 515(1), 229–241. https://doi.org/10.1016/j.aca.2004.01.004

Brodinová, Š., Filzmoser, P., Ortner, T., Breiteneder, C., & Rohm, M. (2017). *Robust and sparse k-means clustering for high-dimensional data*. Advances in Data Analysis and Classification, 13(4), 905–932. https://doi.org/10.1007/s11634-019-00356-9

Brown, P. J., Fearn, T., & 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. https://doi.org/10.1198/016214501753168118

Bulut, E., & Alma, Ö. (2012). *A Performance Assessment of Model Selection Criteria When the Number of Objects Is Much Larger than the Number of Variables in PLSR*. European Journal of Applied Sciences, 4(6), 257–264. https://doi.org/10.5829/idosi.ejas.2012.4.6.1111

Burian, A., Vehviläinen, M., & Kangas, J. (2008). *Camera Barcode Reader with Automatic Localization , Detection of Orientation and Type Classification*. COMPUTERS and SIMULATION in MODERN SCIENCE, I, 214–219.

Calvini, R., Ulrici, A., & Amigo, J. M. (2015). *Practical comparison of sparse methods for classification of Arabica and Robusta coffee species using near infrared hyperspectral imaging*. Chemometrics and Intelligent Laboratory Systems, 146, 503–511. https://doi.org/10.1016/j.chemolab.2015.07.010

Cao, L. (2006). *Singular Value Decomposition applied to digital image processing*. Division of Computing Studies, Arizona State University, 1–15. Retrieved from http://www.lokminglui.com/CaoSVDintro.pdf

Cao, K. A., Boitard, S., & Besse, P. (2011). *Sparse PLS discriminant analysis: Biologically relevant feature selection and graphical displays for multiclass problems*. BMC Bioinformatics, 12(June 2011), 0–16. https://doi.org/10.1186/1471-2105-12-253

Cao, K. A., Rossouw, D., Robert-Granié, C., & Besse, P. (2008). *A sparse PLS for variable selection when integrating omics data*. Statistical Applications in Genetics and Molecular Biology, 7(1). https://doi.org/10.2202/1544-6115.1390

Centner, V., Massart, D. L., De Noord, O. E., De Jong, S., Vandeginste, B. M., & Sterna, C. (1996). *Elimination of Uninformative Variables for Multivariate Calibration*. Analytical Chemistry, 68(21), 3851–3858. https://doi.org/10.1021/ac960321m

Chen, D., Hu, B., Shao, X., & Su, Q. (2004). *Variable selection by modified IPW (iterative predictor weighting)-PLS (partial least squares) in continuous wavelet regression models*. Analyst, 129(7), 664–669. https://doi.org/10.1039/b400410h

Chiang, J. (2008). *The Algorithm for Multiple Outliers Detection The Formulation of the Methods for Multiple Outliers. Business, 3*(17), 839–859.

Coakley, C. W., & Hettmansperger, T. P. (1993). *A bounded influence, high breakdown, efficient regression estimator*. Journal of the American Statistical Association, 88(423), 872–880. https://doi.org/10.1080/01621459.1993.10476352

Corman, V. M., Landt, O., Kaiser, M., Molenkamp, R., Meijer, A., Chu, D. K. W., … Drosten, C. (2020). *Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR*. Eurosurveillance, 25(3), 1–8. https://doi.org/10.2807/1560-7917.ES.2020.25.3.2000045

Croux, C., & Haesbroeck, G. (1999). *Influence Function and Efficiency of the Minimum Covariance Determinant Scatter Matrix Estimator*. Journal of Multivariate Analysis, 190(2), 161–190. https://doi.org/https://doi.org/10.1006/jmva.1999.1839

Cummins, D. J., & Andrews, C. W. (1995). *Iteratively reweighted partial least squares: A performance analysis by monte carlo simulation*. Journal of Chemometrics, 9(6), 489–507. https://doi.org/10.1002/cem.1180090607

Dhhan, W., Rana, S., & Midi, H. (2016). *A high breakdown, high efficiency and bounded influence modified GM estimator based on support vector regression*. Journal of Applied Statistics, 44(4), 700–714. https://doi.org/10.1080/02664763.2016.1182133

Ding, B., & Gentleman, R. (2005). *Classification using generalized partial least squares. Journal of Computational and Graphical Statistics*, 14(2), 280–298. https://doi.org/10.1198/106186005X47697

Downey, G., Briandet, R., Wilson, R. H., & Kemsley, E. K. (1997). *Near- and Mid-Infrared Spectroscopies in Food Authentication: Coffee Varietal Identification*. Journal of Agricultural and Food Chemistry, 45(11), 4357–4361. https://doi.org/10.1021/jf970337t

Engelen, Hubert, Branden, V., & Verboven. (2004). *Robust PCR and Robust PLSR: a Comparative Study*. Theory and Applications of Recent Robust Methods, 105–117. https://doi.org/10.1007/978-3-0348-7958-3_10

Fan, J., & Lv, J. (2008). *Sure Independence Screening for Ultra-High Dimensional Feature Space* ∗, (April 2013).

Feng, C. M., Xu, Y., Liu, J. X., Gao, Y. L., & Zheng, C. H. (2019). *Supervised discriminative sparse PCA for com-characteristic gene selection and tumor classification on multiview biological data*. IEEE transactions on neural networks and learning systems, 1–12.

Filzmoser, P., Liebmann, B., & Varmuza, K. (2009). *Repeated double cross validation*. Journal of Chemometrics, 23(4), 160–171. https://doi.org/10.1002/cem.1225

Fonseca, J. B. (2006). *From Fisher's Linear Discriminant Analysis to NLDA or the Story of the Solution of a Very Difficult Nonlinear Classification Problem*. Proceedings of the 6th WSEAS International Conference on Simulation, Modelling and Optimization, (January 2006), 378–384.

Fordellone, M. (2019). *Dimensionality reduction and simultaneous classification approaches for complex data: methods and applications*. universita di roma.

Forina, M., Casolino, C., & Pizarro Millan, C. (1999). *Iterative predictor weighting (IPW) PLS: A technique for the elimination of useless predictors in regression problems.* Journal of Chemometrics, 13(2), 165–184. https://doi.org/10.1002/(SICI)1099-128X(199903/04)13:2<165::AID-CEM535>3.0.CO;2-Y

Foss, A. P. O. (2010). *high-dimensional data mining: subspace clustering, outlier detection and applications to classification*.

Frank, I. E. (1987). *Intermediate least squares regression method*. Chemometrics and Intelligent Laboratory Systems, 1(3), 233–242. https://doi.org/10.1016/0169-7439(87)80067-9

Friedman, J. H. (1989). *Regularized discriminant analysis.* Journal of the American Statistical Association, 84(405), 165–175. https://doi.org/10.1080/01621459.1989.10478752

Fritsch, V., Varoquaux, G., Thyreau, B., Poline, J. B., & Thirion, B. (2011). *Detecting outlying subjects in high-dimensional neuroimaging datasets with regularized minimum covariance determinant*. MICCAI 2011, Part III, LNCS 6893, 6893 LNCS(PART 3), 264–271. https://doi.org/10.1007/978-3-642-23626-6_33

Geisser, S. (1974). *The predictive sample reuse method with applications*. Journal of the American Statistical Association, 70(350), 320–328. https://doi.org/10.1080/01621459.1975.10479865

Geladi, P., & Kowalski, B. R. (1986). *Partial least-squares regression: a tutorial.* Elsevier Science Publishers, 186.

Gil, J. A., & Romera, R. (1998). *On robust partial least squares (PLS) methods*. Journal of Chemometrics, 12(6), 365–378. https://doi.org/10.1002/(SICI)1099-128X(199811/12)12:6<365::AID-CEM519>3.0.CO;2-G

Giraud, C. (2015). *Introduction to Statistics*. Research and Biostatistics for Nurses. https://doi.org/10.5005/jp/books/13016_12

González, J., Peña, D., & Romera, R. (2008). *A robust partial least squares regression method with applications*. Journal of Chemometrics, 23(2), 78–90. https://doi.org/10.1002/cem.1195

Gschwandtner, M., & Filzmoser, P. (2013). *Outlier detection in high dimension using regularization.* Advances in Intelligent Systems and Computing, 190 AISC, 237–244. https://doi.org/10.1007/978-3-642-33042-1_26

Guo, Y., Hastie, T., & Tibshirani, R. (2007). *Regularized linear discriminant analysis and its application in microarrays.* Biostatistics, 8(1), 86–100. https://doi.org/10.1093/biostatistics/kxj035

Habshah, M., Norazan, M. R., & Imon, A. H. M. R. (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. https://doi.org/10.1080/02664760802553463

Hasegawa, K., Miyashita, Y., & Funatsu, K. (1997). *GA strategy for variable selection in QSAR studies: GA-based PLS analysis of calcium channel antagonists.* Journal of Chemical Information and Computer Sciences, 37(2), 306–310. https://doi.org/10.1021/ci960047x

Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The elements of statistical learning*. https://doi.org/10.1007/b94608_4

He, X., & Fung, W. K. (2000). *High Breakdown Estimation for Multiple Populations with Applications to Discriminant Analysis.* Journal of Multivariate Analysis, 72(2), 151–162. https://doi.org/10.1006/jmva.1999.1857

Hoaglin, D. C., & Welsch, R. E. (1978). *The hat matrix in regression and anova.* American Statistician, 32(1), 17–22. https://doi.org/10.1080/00031305.1978.10479237

Hoffmann, I., Serneels, S., Filzmoser, P., & Croux, C. (2015). *Sparse partial robust M regression.* Chemometrics and Intelligent Laboratory Systems, 149, 50–59. https://doi.org/10.1016/j.chemolab.2015.09.019

Hu, Y., Liu, J.-X., Gao, Y.-L., & Shang, J. (2019). *DSTPCA: Double-Sparse Constrained Tensor Principal Component Analysis Method for Feature Selection.* IEEE/ACM Transactions on Computational Biology and Bioinformatics, 5963(c), 1–1. https://doi.org/10.1109/tcbb.2019.2943459

Huang, C., Du, J., Nie, B., Yu, R., Xiong, W., & Zeng, Q. (2019). *Feature selection method based on partial least squares and analysis of traditional Chinese medicine data.* Computational and Mathematical Methods in Medicine, 2019. https://doi.org/10.1155/2019/9580126

Hubert, M., & Branden, K. Vanden. (2003). *Robust methods for partial least squares regression.* Journal of Chemometrics, 17(10), 537–549. https://doi.org/10.1002/cem.822

Hubert, M., Debruyne, M., & Rousseeuw, P. J. (2018). *Minimum covariance determinant and extensions*. Wiley Interdisciplinary Reviews: Computational Statistics, 10(3), 1–11. https://doi.org/10.1002/wics.1421

Hubert, Mia, & Driessen, K. Van. (2004). *Fast and Robust Discriminant Analysis,* 9473(March). https://doi.org/10.1016/S0167-9473(02)00299-2

Hubert, Mia, Reynkens, T., Schmitt, E., & Verdonck, T. (2015). *Sparse PCA for High-Dimensional Data With Outliers*. Technometrics, 58(4), 424–434. https://doi.org/10.1080/00401706.2015.1093962

Hubert, Mia, Rousseeuw, P. J., & Vanden Branden, K. (2005). *ROBPCA: A new approach to robust principal component analysis*. Technometrics (Vol. 47). https://doi.org/10.1198/004017004000000563

Hubert, Mia, Rousseeuw, P. J., & Verdonck, T. (2012). *A deterministic algorithm for robust location and scatter.* Journal of Computational and Graphical Statistics, 21(3), 618–637. https://doi.org/10.1080/10618600.2012.672100

Hubert, Mia, & Verboven, S. (2003). *A robust PCR method for high-dimensional regressors*. Journal of Chemometrics, 17(8–9), 438–452. https://doi.org/10.1002/cem.783

Indahl, U. G., Martens, H., & Næs, T. (2007). *From dummy regression to prior probabilities in PLS-DA*. Journal of Chemometrics, 21(12), 529–536. https://doi.org/10.1002/cem.1061

Jiménez-Carvelo, A. M., Martín-Torres, S., Ortega-Gavilán, F., & Camacho, J. (2021). *PLS-DA vs sparse PLS-DA in food traceability. A case study: Authentication of avocado samples.* Talanta, 224(August). https://doi.org/10.1016/j.talanta.2020.121904

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

Joswiak, M., Peng, Y., Castillo, I., & Chiang, L. H. (2019). *Dimensionality reduction for visualizing industrial chemical process data.* Control Engineering Practice, 93(June). https://doi.org/10.1016/j.conengprac.2019.104189

Kaewumpai, P., & Chongcharoen, S. (2017). *Discriminant methods for high dimensional data*. Songklanakarin Journal of Science and Technology, 41(2), 319–331. https://doi.org/10.14456/sjst-psu.2019.41

Kalivas, J. H. (1997). *Two data sets of near infrared spectra*. Chemometrics and Intelligent Laboratory Systems, 37(2), 255–259. https://doi.org/10.1016/S0169-7439(97)00038-5

Kemsley, E. K. (1996). *Discriminant analysis of high-dimensional data : a comparison of principal components analysis and partial least squares data reduction methods*, 33, 47–61.

Killeen, D. P., Card, A., Gordon, K. C., & Perry, N. B. (2019). *First Use of Handheld Raman Spectroscopy to Analyze Omega-3 Fatty Acids in Intact Fish Oil Capsules.* Applied Spectroscopy, 74(3), 365–371. https://doi.org/10.1177/0003702819877415

Krämer, N., & Braun, M. L. (2007). *Kernelizing PLS, degrees of freedom, and efficient model selection.* ACM International Conference Proceeding Series, 227, 441–448. https://doi.org/10.1145/1273496.1273552

Kruger, U., & Joe Qin, S. (2003). *Canonical Correlation Partial Least Squares.* IFAC Proceedings Volumes. https://doi.org/10.1016/S1474-6670(17)34989-3

Kucirka, L. M., Lauer, S. A., Laeyendecker, O., Boon, D., & Lessler, J. (2020). *Variation in false-negative rate of reverse transcriptase polymerase chain reaction–based SARS-CoV-2 tests by time since exposure.* Annals of Internal Medicine, 173(4), 262–268. https://doi.org/10.7326/M20-1495

Kvalheim, O. M., & Karstang, T. V. (1989). *Interpretation of latent-variable regression models.* Chemometrics and Intelligent Laboratory Systems, 7(1–2), 39–51. https://doi.org/10.1016/0169-7439(89)80110-8

Leardi, R., & Lupiáñez González, A. (1998). *Genetic algorithms applied to feature selection in PLS regression: How and when to use them*. Chemometrics and Intelligent Laboratory Systems, 41(2), 195–207. https://doi.org/10.1016/ S0169-7439(98)00051-3

Leardi, R., Seasholtz, M. B., & Pell, R. J. (2002). *Variable selection for multivariate calibration using a genetic algorithm: Prediction of additive concentrations in polymer films from Fourier transform-infrared spectral data*. Analytica Chimica Acta, 461(2), 189–200. https://doi.org/10.1016/S0003-2670(02)00272-6

Lee, L. C., Liong, C. Y., & Jemain, A. A. (2018). *Partial least squares-discriminant analysis (PLS-DA) for classification of high-dimensional (HD) data: A review of contemporary practice strategies and knowledge gaps*. Analyst, 143(15), 3526–3539. https://doi.org/10.1039/c8an00599k

Lemberge, P., De Raedt, I., Janssens, K. H., Wei, F., & 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, 14(5–6), 751–763. https://doi.org/10.1002/1099-128X(200009/12)14:5/6<751::AID-CEM622>3.0.CO;2-D

Li, B. B., Morris, J., & Martin, E. B. (2002). *Model selection for partial least squares regression.* Chemometrics Intell. Lab. Syst., 64(1), 79–89. https://doi.org/10.1016/S0169-7439(02)00051-5

Li, H. D., Zeng, M. M., Tan, B. Bin, Liang, Y. Z., Xu, Q. S., & Cao, D. S. (2010). *Recipe for revealing informative metabolites based on model population analysis*. Metabolomics, 6(3), 353–361. https://doi.org/10.1007/s11306-010-0213-z

Li, L., Liu, S., Peng, Y., & Sun, Z. (2016). *Overview of principal component analysis algorithm*. Optik, 127(9), 3935–3944. https://doi.org/10.1016/j.ijleo. 2016.01.033

Lim, H. A., & 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. https://doi.org/10.1007/s00180-016-0662-6

Lindgren, F., Geladi, P., & Wold, S. (1993). *The Kernel Algorithm for PLS,* 7(April 1992), 45–59.

Liu, H., & Setiono, R. (1995). *Chi2: Feature Selection and Discretization of Numeric Attributes*. IEEE, 388–391.

Lopuhaä, H. P. (1999). *Asymptotics of reweighted estimators of multivariate location and scatter*. Annals of Statistics, 27(5), 1638–1665.

Lopuhaa, H. P., & Rousseeuw, P. J. (1991). *Breakdown Points of Affine Equivariant Estimators of Multivariate Location and Covariance Matrices*. The Annals of Statistics. https://doi.org/10.1214/aos/1176347978

Lottering, R. T., Govender, M., Peerbhay, K., & Lottering, S. (2020). *Comparing partial least squares (PLS) discriminant analysis and sparse PLS discriminant analysis in detecting and mapping Solanum mauritianum in commercial forest plantations using image texture.* ISPRS Journal of Photogrammetry and Remote Sensing, 159(April 2019), 271–280. https://doi.org/10.1016/j.isprsjprs. 2019.11.019

Mahalanobis, P. C. (1936). *On the generalized distance in statistics*.

Mallows, C. L. (1973). *Some comments on Cp.* Technometrics, 15(4), 661–675. https://doi.org/10.1080/00401706.1973.10489103

Markatou, M., Basu, A., & Lindsay, B. G. (1998). *Weighted likelihood estimating equations with a bootstrap search.* Journal of the American Statistical Association, 93(442), 740–750.

Maronna, R. A., Martin, R. D., Yohai, V. J., & Salibian-Barrera, M. (2006). *Robust Statistics*. John Wiley & Sons.

Maronna, R. A., & Zamar, R. H. (2002). *Robust estimates of location and dispersion for high-dimensional datasets*. Technometrics, 44(4), 307–317. https://doi.org/10.1198/004017002188618509

Martens, H. A., & Dardenne, P. (1998). *Validation and verification of regression in small data sets*. Chemometrics and Intelligent Laboratory Systems, 44(1–2), 99–121. https://doi.org/10.1016/S0169-7439(98)00167-1

Mehmood, T. (2016). *Hotelling T2 based variable selection in partial least squares regression.* Chemometrics and Intelligent Laboratory Systems, 154, 23–28. https://doi.org/10.1016/j.chemolab.2016.03.001

Mehmood, T. (2021). *Regularized Feature Selection in Categorical PLS for Multicollinear Data.* Mathematical Problems in Engineering, 2021. https://doi.org/10.1155/2021/5561752

Mehmood, T., & Ahmed, B. (2016). *The diversity in the applications of partial least squares: An overview.* Journal of Chemometrics, 30(1), 4–17. https://doi.org/10.1002/cem.2762

Mehmood, T., Sæbø, S., & Liland, K. H. (2020). *Comparison of variable selection methods in partial least squares regression.* Journal of Chemometrics, 34(6), 1–14. https://doi.org/10.1002/cem.3226

Mejia, A. F., Nebel, M. B., Eloyan, A., Caffo, B., & Lindquist, M. A. (2017). *PCA leverage: Outlier detection for high-dimensional functional magnetic resonance imaging data*. Biostatistics, 18(3), 521–536. https://doi.org/10.1093/ biostatistics/kxw050

Mevik, B.-H., & Wehrens, R. (2015). *Introduction to the pls Package*. Help Section of the "Pls" Package of RStudio Software, (Section 7), 1–23.

Midi, Habshah, Hendi, H. T., Arasan, J., & Uraibi, H. (2020). SCIENCE & TECHNOLOGY *Fast and Robust Diagnostic Technique for the Detection of High Leverage Points*, *28*(4), 1203–1220.

Midi, Habshah, Rana, S., & Imon, R. (2014). *Two-step robust estimator in heteroscedastic regression model in the presence of outliers. Economics Computation and Economics Cybernatics Studies and Research*, *48(3).*

Midi, H., Ramli, N. M., & Imon, A. H. M. R. (2009). *The performance of diagnostic-robust generaliozed potential approach for the identification of multiple high leverage points in linear regression*. Journal of Applied Statistics, 36(5), 1–15.

Mosteller, F., & L.Wallace, D. (1963). *Inference in an authorship problem. Encyclopedia of Research Design*. https://doi.org/10.4135/9781412961288.n9

Ndaoud, M., & Tsybakov, A. B. (2020). *Optimal variable selection and adaptive noisy compressed sensing*. IEEE Transactions on Information Theory, 66(4), 2517–2532. https://doi.org/10.1109/TIT.2020.2965738

Nengsih, T. A., Bertrand, F., Maumy-Bertrand, M., & Meyer, N. (2019). *Determining the number of components in PLS regression on incomplete data set.* Statistical Applications in Genetics and Molecular Biology, (November). https://doi.org/10.1515/sagmb-2018-0059

Nocairi, H., Qannari, E. M., Vigneau, E., & Bertrand, D. (2005). *Discrimination on latent components with respect to patterns.* Application to multicollinear data. Computational Statistics and Data Analysis, 48(1), 139–147. https://doi.org/10.1016/j.csda.2003.09.008

Nunes, A., Martins, J., Barros, A. S., Galvis-Sánchez, A. C., & Delgadillo, I. (2009). *Estimation of olive oil acidity using FT-IR and partial least squares regression*. Sensing and Instrumentation for Food Quality and Safety, 3(3), 187–191. https://doi.org/10.1007/s11694-009-9084-2

Peña and Victor J. Yohai. (1991). *The detection of influential subsets in linear regression using an influence matrix*. Journal of Chemical Information and Modeling, 53(9), 1689–1699. https://doi.org/10.1017/CBO9781107415324.004

Polat, & Gunay, S. (2019). *A new robust partial least squares regression method based on a robust and an efficient adaptive reweighted estimator of covariance.* AMS, (Mcd).

Pires, A. M. (2003). *Robust Linear Discriminant Analysis and the Projection Pursuit Approach.* Developments in Robust Statistics, (2001), 317–329. https://doi.org/10.1007/978-3-642-57338-5_28

Polat, E. (2019). *The effects of different weight functions on partial robust M-regression performance: A simulation study. Communications in Statistics: Simulation and Computation*, 0918. https://doi.org/10.1080/03610918.2019.1586926

Polat, E., & Gunay, S. (2014). *A New Approach to Robust Partial Least Squares Regression Analysis.* International Journal of Mathematics Trends and Technology, 9(3), 197–205. https://doi.org/10.14445/22315373/ijmtt-v9p524

Polat, E., & Gunay, S. (2019). *A new robust partial least squares regression method based on a robust and an efficient adaptive reweighted estimator of covariance.* Revstat Statistical Journal, 17(4), 449–474.

Porwal, U., & Mukund, S. (2019). *Credit Card Fraud Detection in e-Commerce: An Outlier Detection Approach.* Retrieved from http://arxiv.org/abs/1811.02196

Quenouille, M. H. (1949). *Approximate Tests of Correlation in Time-Series. Journal of the Royal Statistical Society: Series B (Methodological)*, *11*(1), 68–84. https://doi.org/10.1111/j.2517-6161.1949.tb00023.x

Quinn, T. P., Nguyen, T., Lee, S. C., & Venkatesh, S. (2019). *Cancer as a tissue anomaly: Classifying tumor transcriptomes based only on healthy data.* Frontiers in Genetics, 10(JUL), 1–6. https://doi.org/10.3389/fgene.2019.00599

Rahmatullah Imon, A. H. M. (2005). *Identifying multiple influential observations in linear regression.* Journal of Applied Statistics, 32(9), 929–946. https://doi.org/10.1080/02664760500163599

Ramey, J. A., Stein, C. K., Young, P. D., & Young, D. M. (2017). *High-Dimensional Regularized Discriminant Analysis*. Retrieved from http://arxiv.org/abs/1602.01182

Rashid, A., Midi, H., Dhhan, W., & Arasan, J. (2021). *Detection of outliers in high-dimensional data using nu-support vector regression*. Journal of Applied Statistics. https://doi.org/10.1080/02664763.2021.1911965

Riazoshams, H., & Midi, H. (2014). *The Performance of a Robust Multistage Estimator in Nonlinear Regression with Heteroscedastic Errors*, *0918*(June 2016). https://doi.org/10.1080/03610918.2014.944657

Ro, K., Zou, C., Wang, Z., & Yin, G. (2015). *Outlier detection for high-dimensional data*. Biometrika, 102(3), 589–599. https://doi.org/10.1093/biomet/asv021

Roelant, E., Van Aelst, S., Willems, G., Aelst, S. Van, Willems, G., Van Aelst, S., & Willems, G. (2009). *The minimum weighted covariance determinant estimator*. Metrika, 70(2), 177–204. https://doi.org/10.1007/s00184-008-0186-3

Rousseeuw, P. (1985). *Multivariate Estimation With High Breakdown Point*. Mathematical Statistics and Applications, (June). https://doi.org/10.1007/978-94-009-5438-0

Rousseeuw, P., & Driessen, K. (1999). *A Fast Algorithm for the Minimum Covariance*. Technometrics, 41(3), 212–223.

Rousseeuw, P. J., & Croux, C. (1993). *Alternatives to the median absolute deviation*. Journal of the American Statistical Association, 88(424), 1273–1283. https://doi.org/10.1080/01621459.1993.10476408

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. (Vol. 152). Bristol-Myers-Squibb B-1170 Brussels, Belgium: John wiley & sons,. https://doi.org/10.2307/2982847

Rousseeuw, P. J., & Zomeren, B. C. van. (1990). *Unmasking multivariate outliers and leverage points*. Journal of the American Statistical Association. https://doi.org/10.1080/01621459.1990.10474920

Salibian-Barrera, M., & Yohai, V. J. (2006). *A fast algorithm for S-regression estimates*. Journal of Computational and Graphical Statistics, 15(2), 414–427. https://doi.org/10.1198/106186006X113629

Santos-Rufo, A., Mesas-Carrascosa, F. J., García-Ferrer, A., & Meroño-Larriva, J. E. (2020). *Wavelength selection method based on partial least square from hyperspectral unmanned aerial vehicle orthomosaic of irrigated olive orchards*. Remote Sensing, 12(20), 1–20. https://doi.org/10.3390/rs12203426

Schwarz, G. (1978). *Estimating the dimension of a model*. Annals of Statistics, 4(6), 1280–1289. Retrieved from http://projecteuclid.org/euclid.aop/1176996548

Serfling, R., & Wang, S. (2012). *General foundations for studying masking and swamping robustness of outlier identifiers*. Statistical Methodology, 20(May), 79–90. https://doi.org/10.1016/j.stamet.2013.08.004

Serneels, S., Croux, C., Filzmoser, P., & Van Espen, P. J. (2005). *Partial robust M-regression*. Chemometrics and Intelligent Laboratory Systems, 79(1–2), 55–64. https://doi.org/10.1016/j.chemolab.2005.04.007

Serneels, S., De Nolf, E., & Van Espen, P. J. (2006). *Spatial sign preprocessing: A simple way to impart moderate robustness to multivariate estimators*. Journal of Chemical Information and Modeling, 46(3), 1402–1409. https://doi.org/10.1021/ci050498u

Shieh, A. D., & Hung, Y. S. (2009). *Detecting Outlier Samples in Microarray Data Detecting Outlier Samples in Microarray Data*, 8(1). https://doi.org/10.2202/1544-6115.1426

Singhabahu, D. M. (2013). *Robust partial least squares regression and outlier detection using repeated minimum covariance determinant method and a resampling method*. university of pittsburgh.

Stefansson, P., Liland, K. H., Thiis, T., & Burud, I. (2020). *Fast method for GA-PLS with simultaneous feature selection and identification of optimal preprocessing technique for datasets with many observations*. Journal of Chemometrics, 34(3), 1–15. https://doi.org/10.1002/cem.3195

Stone, M. (1973). *Cross-Validatory Choice and Assessment of Statistical Predictions*. Journal of the Royal Statistical Society: Series B (Methodological), 38(1), 102–102. https://doi.org/10.1111/j.2517-6161.1976.tb01573.x

Su, P., Tarr, G., & Muller, S. (2022). *Robust Variable Selection under Cellwise Contamination*, 1–16. Retrieved from http://arxiv.org/abs/2110.12406

Sun, F., Chen, Y., Wang, K. Y., Wang, S. M., & Liang, S. W. (2019). *Identification of Genuine and Adulterated Pinellia ternata by Mid-Infrared (MIR) and Near-Infrared (NIR) Spectroscopy with Partial Least Squares - Discriminant Analysis (PLS-DA).* Analytical Letters, 53(6), 937–959. https://doi.org/10.1080/00032719.2019.1687507

T.Cacoullos. (1972). *Discriminant analysis and applications.* Academic Press.

Thakkar, S., Perkins, R., Hong, H., & Tong, W. (2018). *Computational Toxicology*. Comprehensive Toxicology: Third Edition (Third Edit, Vol. 5–15). Elsevier Ltd. https://doi.org/10.1016/B978-0-12-801238-3.64317-9

Thomaz, C. E., Kitani, E. C., & Gillies, D. F. (2005). *A maximum uncertainty LDA-based approach for limited sample size problems — with application to face recognition.* Journal of the Brazilian Computer Society, 12(2), 7–18. https://doi.org/10.1007/BF03192391

Tibshirani, R., Hastie, T., Narasimhan, B., & Chu, G. (2002). *Diagnosis of multiple cancer types by shrunken centroids of gene expression*. Proceedings of the National Academy of Sciences of the United States of America, 99(10), 6567–6572. https://doi.org/10.1073/pnas.082099299

Tibshirani, R. (1996). *Regression Shrinkage and Selection via the Lasso.* Journal of the Royal Statistical Society. Series B (Methodological), 58(1), 267–288. Retrieved from jstor.org/stable/2346178

Tran, T. N., Afanador, N. L., Buydens, L. M. C., & Blanchet, L. (2014). *Interpretation of variable importance in Partial Least Squares with Significance Multivariate Correlation (sMC).* Chemometrics and Intelligent Laboratory Systems, 138, 153–160. https://doi.org/10.1016/j.chemolab.2014.08.005

Uraibi, H. S., Midi, H., & Rana, S. (2017). *Selective overview of forward selection in terms of robust correlations, 0918*(May). https://doi.org/10.1080/03610918.2016.1164862

Varmuza, K., & Filzmoser, P. (2008). *Introduction to Multivariate Statistical Analysis in Chemometrics*. Applied Spectroscopy (Vol. 64). CRC Press. https://doi.org/10.1201/9781420059496

Velleman, P. F., & Welsch, R. O. Y. E. (1981). *Efficient Computing of Regression Diagnostics,* Author (s): Paul F . Velleman and Roy E . Welsch Source : The American Statistician , Vol . 35 , No . 4 (Nov ., 1981), pp . 234-242 Published by : Taylor & Francis , Ltd . on behalf of the American Statist, 35(4), 234–242.

Vijaya kumar, V., Srikrishna, A., Babu, B. R., & Mani, M. R. (2010). *Classification and recognition of handwritten digits by using mathematical morphology*. Indian Academy of Sciences, 35(4), 419–426. https://doi.org/10.1007/s12046-010-0031-z

Wold, H. (1975). *Soft Modelling by Latent Variables: The Non-Linear Iterative Partial Least Squares (NIPALS) Approach*. Journal of Applied Probability, 12(S1), 117–142. https://doi.org/10.1017/s0021900200047604

Wold, S. (1978). *Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models*. Technometrics, 20(4), 397–405. https://doi.org/10.1080/00401706.1978.10489693

Xu, Q. S., & Liang, Y. Z. (2000). *Monte Carlo cross validation*. Chemometrics and Intelligent Laboratory Systems, 56(1), 1–11. https://doi.org/10.1016/S0169-7439(00)00122-2

Yahaya, S. S. S., Lim, Y. F., Ali, H., & Omar, Z. (2016). *Robust linear discriminant analysis*. Journal of Mathematics and Statistics, 12(4), 312–316. https://doi.org/10.3844/jmssp.2016.312.316

Yazgan, N. N., Genis, H. E., Bulat, T., Topcu, A., Durna, S., Yetisemiyen, A., & Boyaci, I. H. (2020). *Discrimination of milk species using Raman spectroscopy coupled with partial least squares discriminant analysis in raw and pasteurized milk*. Journal of the Science of Food and Agriculture, 100(13), 4756–4765. https://doi.org/10.1002/jsfa.10534

Ye, B., & Liu, P. (2019). *Classification of high-dimensional data: A random-matrix regularized discriminant analysis approach*. International Journal of Innovative Computing, Information and Control, 15(3), 955–967. https://doi.org/10.24507/ijicic.15.03.955

Yin, G., Li, L., Lu, S., Yin, Y., Su, Y., Zeng, Y., Lang, J. (2021). *An efficient primary screening of COVID-19 by serum Raman spectroscopy.* Journal of Raman Spectroscopy, 52(5), 949–958. https://doi.org/10.1002/jrs.6080

Yu, H., & Yang, J. (2001). *A direct LDA algorithm for high-dimensional data — with application to face recognition*. Pattern Recognition, 34(10), 2067–2070. https://doi.org/10.1016/s0031-3203(00)00162-x

Yu, P. S., & Aggarwal, C. C. (2001). *Outlier detection for high dimensional data*. Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data, 37–46. https://doi.org/10.1109/ICUMT.2009.5345621

Yuan, Z., Zhang, L., Wang, D., Jiang, J., Harrington, P. de B., Mao, J., Li, P. (2020). *Detection of flaxseed oil multiple adulteration by near-infrared spectroscopy and nonlinear one class partial least squares discriminant analysis.* Lwt, 125(March). https://doi.org/10.1016/j.lwt.2020.109247

Zheng, W., Shu, H., Tang, H., & Zhang, H. (2019). *Spectra data classification with kernel extreme learning machine.* Chemometrics and Intelligent Laboratory Systems, 192(January). https://doi.org/10.1016/j.chemolab.2019.103815

Zhu, M., & Ghodsi, A. (2006). *Automatic dimensionality selection from the scree plot via the use of profile likelihood.* Computational Statistics and Data Analysis, 51(2), 918–930. https://doi.org/10.1016/j.csda.2005.09.010

Zontov, Y. V., Rodionova, O. Y., Kucheryavskiy, S. V., & Pomerantsev, A. L. (2020). *PLS-DA – A MATLAB GUI tool for hard and soft approaches to partial least squares discriminant analysis*. Chemometrics and Intelligent Laboratory Systems, 203(June). https://doi.org/10.1016/j.chemolab.2020.104064

Zou, H., Hastie, T., & Tibshirani, R. (2006). *Sparse Principal Component Analysis*, 15(2), 265–286. https://doi.org/10.1198/106186006X113430