

ROBUST PARTIAL LEAST SQUARES REGRESSION AND OUTLIER DETECTION
USING REPEATED MINIMUM COVARIANCE DETERMINANT METHOD AND A
RESAMPLING METHOD

by

Dilrukshika Manori Singhabahu

BS, Information Technology, Slippery Rock University, 2008

Submitted to the Graduate Faculty of
Graduate School of Public Health in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

University of Pittsburgh

2013

UNIVERSITY OF PITTSBURGH
GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Dilrukshika Manori Singhabahu

It was defended on

June 5th, 2013

and approved by

Howard Aizenstein, MD, PhD, Associate Professor of Psychiatry, Bioengineering,
and Clinical and Translational Science, Department of Psychiatry, University of
Pittsburgh Medical Center

Chung-Chou Ho Chang, PhD, Associate Professor of Medicine, Biostatistics, and Clinical
and Translational Science, Department of Medicine, University of Pittsburgh

Yan Lin, PhD, Research Assistant Professor, Department of Biostatistics, Graduate School of
Public Health, University of Pittsburgh

Dissertation Advisor: Lisa A. Weissfeld, Professor, PhD, Department of Biostatistics,
Graduate School of Public Health, University of Pittsburgh

Copyright © by Dilrukshika Manori Singhabahu

2013

Lisa A. Weissfeld, PhD

**ROBUST PARTIAL LEAST SQUARES REGRESSION AND OUTLIER DETECTION
USING REPEATED MINIMUM COVARIANCE DETERMINANT METHOD AND A
RESAMPLING METHOD**

Dilrukshika Manori Singhabahu, PhD

University of Pittsburgh, 2013

ABSTRACT

Partial Least Squares Regression (PLSR) is often used for high dimensional data analysis where the sample size is limited, the number of variables is large, and the variables are collinear. Like other types of regression, PLSR is influenced by outliers and/or influential observations. Since PLSR is based on the covariance matrix of the outcome and the predictor variables, this is a natural starting point for the development of techniques that can be used to identify outliers and to provide stable estimates in the presence of outliers. We focus on the use of the minimum covariance determinant (MCD) method for robust estimation of the covariance matrix when $n \gg p$ and modify this method for application to a magnetic resonance imaging (MRI) data set. We extend this approach by applying the MCD to generate robust Mahalanobis squared distances (RMSD) in the Y vector and the X matrix separately and then identify the outliers based on the RMSD. We then remove these observations from the data set and apply PLSR to the remaining data. This approach is applied iteratively until no new outliers are detected. Simulation studies demonstrate that the PLSR results are improved when using this approach.

Another approach to outlier detection is explored for the setting where $n < p$. This approach, resampling by half-means (RHM), was introduced in 1998 by William Egan and

Stephen Morgan. We adapt this method for use in MRI data to detect outliers and then to develop a robust PLSR model. This method can be used for small or large datasets overcoming the limitation of the leading multivariate outlier detection methods such as the MSD method that cannot be used for small sample sizes ($n < p$).

The two methods proposed improve the accuracy of predictions on brain imaging data (MRI in our example). Thus the public health significance is increasing the accuracy in brain imaging diagnosis and predictions.

TABLE OF CONTENTS

PREFACE.....	X
1.0 INTRODUCTION.....	1
2.0 LITERATURE REVIEW	3
2.1 PARTIAL LEAST SQUARES REGRESSION (PLSR)	3
2.2 ROBUST PARTIAL LEAST SQUARES REGRESSION AND OUTLIERS	4
2.3 MINIMUM COVARIANCE DETERMINANT (MCD).....	6
2.4 TWO ALTERNATE METHODS FOR MULTIVARIATE OUTLIER DETECTION	8
3.0 SCOPE	10
3.1 METHOD 1 – INTRODUCTION	10
3.2 METHOD 2 – INTRODUCTION	10
4.0 METHOD 1 – PAPER.....	11
ROBUST PARTIAL LEAST SQUARES REGRESSION AND OUTLIER DETECTION USING MINIMUM COVARIANCE DETERMIANAT MRTHOD....	11
4.1 ABSTRACT.....	11
4.2 INTRODUCTION	12
4.2.1 Data	13
4.2.2 Partial Least Squares Regression (PLSR)	13
4.2.3 Mahalanobis Squared Distance (MSD):	15
4.2.4 Minimum Covariance Determinant (MCD).....	15

4.2.5	Detecting Outliers and Leverage points using Robust Mahalanobis Squared Distance (RMSD)	17
4.3	METHOD	17
4.3.1	Simulation	19
4.4	RESULTS	21
4.5	DISCUSSION	26
5.0	METHOD 2 – PAPER	27
	ROBUST PARTIAL LEAST SQUARES REGRESSION AND OUTLIER DETECTION USING RESAMPLING BY HALF-MEANS METHOD.....	27
5.1	ABSTRACT.....	27
5.2	INTRODUCTION	28
5.2.1	Data:	28
5.2.2	Partial Least Squares Regression (PLSR)	29
5.3	METHOD	30
5.4	RESULTS	33
5.5	DISCUSSION.....	40
6.0	SUMMARY	42
	BIBLIOGRAPHY.....	44

LIST OF TABLES

Table 4-1: MRI data - Chi-squared distributed -cutoff >0.975	21
Table 4-2: MRI data - Chi-squared distributed - cutoff > 0.999	22
Table 4-3: MRI Data – cutoff largest 5% of the robust MSD	22
Table 4-4: Simulation1 – Chi-squared distributed – cutoff > 0.975.....	23
Table 4-5: Simulation1 – Chi-squared distributed – cutoff > 0.999.....	23
Table 4-6: Simulation1 – cutoff largest 5% of the robust MSD	24
Table 5-1: After 1000 rounds the subjects with the largest 5% of the vector lengths sorted by ID for (X)	33
Table 5-2: Comparison of PLSR simulation results for large datasets.....	34
Table 5-3: Small sample simulation results for 2 trials.....	34
Table 5-4: Comparison of PLSR results for smaller datasets.....	35
Table 5-5: MRI data (305 observations) for 5 iterations (samples) for the predictors (X)..	35
Table 5-6: MRI example data after 500 iterations the subjects with the largest 5% of the lengths sorted by ID for the X.....	37
Table 5-7: MRI data (305 observations) for 5 iterations (samples) for the response (Y).....	38
Table 5-8: After 500 iterations the subjects with the largest 5% of the lengths sorted by ID for the Y vector.....	39
Table 5-9: Comparison of PLSR results for the MRI example data.....	40

LIST OF FIGURES

Figure 4-1: MRI Data - Robust Mahalanobis Squared Distances for the three repeats.....	21
Figure 4-2: Simulation1 data - Robust Mahalanobis Squared Distances for the three repeats	23
Figure 4-3: MRI data - Error differences between the repeats, with 2 components chosen for PLSR	25
Figure 4-4: Simulation-1000 data points generated - Error differences between the repeats, with 2 components for PLSR	25
Figure 5-1: Flow chart showing the steps used in the RHM method of outlier detection....	31
Figure 5-2: Vector length distributions for the predictors of MRI example data	36
Figure 5-3: Vector lengths for all the 305 subjects for the Y vector	38

PREFACE

To my parents, Walter and Indrani Singhabahu, my husband Athula Herat, and my daughter Hiruni, thank you for all your support and love!

In loving memory of my sister Varuni!

1.0 INTRODUCTION

Although there is no formal definition describing what an outlier is, Hawkins(1980) describes an outlier as the following, “an outlier is an observation that deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism” [28]. In high-dimensional or p -variate data, when p is greater than two, it is difficult to identify outliers visually. The greatest challenge when analyzing data that has more than two dimensions is that an observation that appears to be an outlier in one-dimension may not unduly influence analysis of the data when using multivariate techniques. Another issue is that of masking, when outliers are grouped with non-outlier data due to the presence of extreme outliers, and therefore are not detectable as outliers. Although there are several multivariate outlier detection techniques available, they are not widely used as they may be difficult to program, require significant processing time and are prone to masking. Techniques such as the Mahalanobis distance, which are based on covariance matrix, are highly influenced by the presence of outliers. Robust multivariate outlier detection techniques such as Minimum Volume Ellipsoid (MVE), Multivariate Trimming (MVT), M-Estimators, and Minimum Covariance Determinant (MCD) are difficult to program and require significant processing time. Rousseeuw and Driessen introduced a fast algorithm for the minimum covariance determinant estimator [35], currently considered as the best performing multivariate outlier detection method even though it is still subject to masking. To address these issues, we propose our method 1, a technique based on a robust Mahalanobis Squared Distance (MSD) that uses the MCD to generate robust location and scale estimates. This proposed technique will identify all possible outliers including the observations that are masked. Then it is possible to use outlier free observations to generate the location vector and the covariance matrix that can be used to develop a robust PLSR.

We have two objectives in this study. The first objective is to identify possible outliers including observations that are masked. Once these observations are identified and examined it can be decided, either to use, modify, or discard the observations. The second objective is to develop a robust Partial Least Squares Regression that is not influenced by possible outliers including points that might be masked. The limitation of method 1 is, it only works for large ($n > p$) datasets.

Our method 2, overcomes the limitation of method 1. Method 2 can be used on large ($n > p$) or small ($n < p$) datasets. To identify outliers we focus on a method called resampling by half-means (RHM), introduced in 1998 by William Egan and Stephen Morgan in Analytical Chemistry [27]. We adapt this method for Magnetic Resonance Imaging (MRI) data to detect outliers and then to develop a robust PLSR model. The method uses sampling without replacement to generate i samples of size $n/2$, and calculate the vector lengths for all observations by auto-scaling the original dataset with the medians and the median absolute deviation of the i samples. Outliers are identified empirically or using the chi-squared cutoff. The outlier free dataset is then used to generate the scatter and the location values to be used in the robust PLSR.

The significance of method 1 is that the method was successful in identifying masked outliers in multivariate datasets which then lead to robust PLSR. Addressing the identification of masked outliers in multivariate data was the most significant outcome of method 1, which then improves the accuracy of the PLSR. The significance of method 2 is that the method is successful in identifying outliers specifically, in small datasets ($n < p$). Identifying outliers in small datasets becomes important in the analysis of brain voxel level data. This is one example where method 2 will improve the accuracy of the statistical analysis of the data.

2.0 LITERATURE REVIEW

2.1 PARTIAL LEAST SQUARES REGRESSION (PLSR)

The method of PLSR is used for predicting a response or set of responses from a group of predictors. While there are many other methods to achieve this objective, PLSR is ideal when there is a large number of predictors and when these predictors are multi-collinear. To implement PLSR, the cross-covariance matrix of the response and the predictor variables is used to generate latent variables based on least squares regression in the prediction model. Partial Least Squares (PLS) is used to investigate the association between the dependent and the independent variables whereas PLSR is used for prediction.

PLSR was first introduced in the literature in the field of social sciences by Herman Wold in 1966 [1]. It was then used in the chemometrics literature, for example, by Geladi and Kowalski in 1986[2]. In their article the authors go through PLSR in detail, including a brief discussion of multiple linear regression, Principal Components Analysis (PCA), and principal components regression as these topics are necessary to understand PLSR. Herman Wold's son, Svante Wold, also contributed to the literature on PLSR during this period [3][4]. One of the initial algorithms used to perform PLSR was the nonlinear iterative partial least squares (NIPALS) algorithm [2][35]. The orthogonal X block factors and the Y block factors are found using the original centered X and the Y matrices. The factors are computed sequentially, using an iterative method that uses the direction and the projection of sample points onto the factors. A detailed explanation of the NIPALS algorithm is found in [3] and [35]. The factors identified by the NIPLAS method are the same as those obtained from a Singular Value Decomposition (SVD) of the covariance matrix of X and Y. In 1993, another algorithm was introduced for

PLSR called SIMPLS [3] [8]. In this method the covariance matrix formed from the covariance of the X and Y matrices is used to compute the singular factors via the iterative power method or the singular value decomposition (SVD). Abdi (2003) [5] provides a detailed description of the PLSR algorithm and the use of the singular value decomposition for the computation of the latent variables. Baker (2005) wrote a tutorial [6] on SVD that gives a straightforward explanation starting from an introduction to matrices, eigenvectors and eigenvalues.

The methodology of PLS was later introduced to other fields such as imaging and genomics. In 1996 A.R. McIntosh [7] published a paper where he used PLS on Positron Emission Tomography (PET) data to explain the relationship between image pixels and a given task or behavior. The example presented was that of a face encoding and recognition PET rCBF study using 10 young subjects. Here McIntosh goes through the PLS algorithm in detail and explains how the SVD of the covariance matrix is used to generate the singular images. A conceptual illustration of the PLS steps is given using functional brain imaging as an example with interpretation of the output from PLS as a focus. The working example uses randomly generated data to represent measures from four pixel images obtained from three conditions with five observations for each condition. The design matrix X has two contrasts. After the PLS analysis the resulting two singular images for the two contrasts were explained in relation to the three conditions given.

2.2 ROBUST PARTIAL LEAST SQUARES REGRESSION AND OUTLIERS

Partial least squares regression uses the empirical cross covariance matrix to develop the latent variables and then uses least squares regression for prediction. The combination of these

two methods makes PLSR highly sensitive to outliers. The recognition of this fact, and the complexity of the methods, has generated interest in the development of methodology to detect outliers for these methods.

A paper published by M. Huber and K. Vanden Branden in 2003[8] gives an overall introduction to the robust methods available for PLSR. The paper begins with a brief introduction to the popular PLSR algorithms, NIPALS [9] and SIMPLS [35]. In the NIPALS algorithm, the orthogonal X block factors and the Y block factors are computed using the original centered X and the Y matrices. The factors are computed sequentially, using an iterative method that uses the direction and the projection of the sample points onto the factors. A detailed explanation of the NIPALS algorithm is found in [30] and [35]. The factors identified by the NIPLAS method are the same as those obtained from a singular value decomposition (SVD) of the covariance matrix of X and Y. In 1993, SIMPLS algorithm was introduced [30] [8]. In this method the covariance matrix formed from the covariance of the X and Y matrices is used to develop the singular factors through the use of the iterative power method or the singular value decomposition (SVD). Both of these algorithms are sensitive to outliers. Let PLSR with a single response be denoted as PLS1, and be denoted as PLS2 when there is more than one response. One of the first robust algorithms mentioned in the paper was developed by Wakelinc and Macfie in 1992[11]. Two iteratively reweighted algorithms[12][13] are mentioned by the authors and the disadvantages of these algorithms are given as that they are only applicable for problems where there is a single response and are not resistant to leverage points. Another algorithm mentioned is the PLS1 method proposed by Gil and Romera in 1998[13]. This method robustifies the x-matrix sample covariance and the cross-covariance matrix of x and y variables. The method used to develop the robust covariance matrices is the Stahel-Donoho estimator

[14][15]. The disadvantages of this method are that it cannot be applied to high-dimensional regressors ($n \ll p$) or when there is more than one response variable. Next, the paper introduces a more recent method developed in 2003 by the authors Hubert and Verboven [17] for Principal Components Regression (PCR) which are applicable to high dimensional x-variables and multiple y-variables. The authors Huber and Branden then present several robust methods for the SIMPLS algorithm. The two methods introduced in the paper are denoted as the RSIMCD and RSIMPLS methods and can be applied when there is more than one response variable. Both of these methods are variants of the SIMPLS algorithm discussed earlier. The estimators developed in these two algorithms are based on robust covariance methods for high dimensional data using the ROBPCA method [18]. The ROBPCA method is based on the Minimum Covariance Determinant (MCD) [19][20] for the development of a robust covariance matrix when the dimension of the data is small ($n > p$). For high dimensional data ($n < p$) projection pursuit [21][22] methods are used.

2.3 MINIMUM COVARIANCE DETERMINANT (MCD)

A paper published by authors Rousseeuw and Driessen in 1999[20] describes a fast algorithm for the Minimum Covariance Determinant Estimator. The MCD provides a robust scatter and location estimate for a given data set where $n > p+q$ where p is the number of variables in the X matrix and q is the number of response variables. The paper discusses the use of the MCD covariance and scatter estimates in the Mahalanobis distance calculation for outlier detection, as the Mahalanobis distances are affected by the presence of outliers. The authors mention that there are several methods for estimating the location and scatter matrices and

compare the Minimum Volume Ellipsoid (MVE) method developed in 1984[23] to MCD. The MVE looks for the ellipsoid with the smallest volume that covers h data points, where $n/2 \leq h < n$. In 1997 Rousseeuw and Leroy proposed a resampling algorithm called the Minimum Volume (MINVOL) algorithm for approximating the MVE. For the implementation of the MINVOL algorithm, the mean and the covariance matrix of a trial subset of $p + 1$ observations are calculated. Then the corresponding ellipsoid from the trial dataset is deflated or inflated until there are exactly h data points. This method is repeated and the smallest volume used in estimation. The paper also discusses other methods used to approximate the MVE [24]. Next, the paper states reasons as to why the MCD is an improvement over the MVE. The idea behind the MCD is to find h observations out of n that have the lowest determinant of its covariance matrix. Location estimates are then the average of these h data points. Compared to the MVE, the MCD is statistically efficient because it is asymptotically normal [25], and MVE has a lower convergence rate [26]. It is stated that robust distances based on the MCD are more accurate and stable than the MVE-based robust distances. The fast MCD algorithm given in the paper [20] is faster when compared to the MVE method. The fast MCD algorithm starts off by randomly choosing a subset of data H_1 from n . To construct the dataset H_1 , first randomly pick $p+1$ data points and calculate the average and the covariance of that data subset, where p is the number of variables in the matrix. If the determinant of the covariance matrix of the subset of data is not greater than 0, randomly add one more data point. Continue this process until the determinant of the covariance matrix is greater than 0. The determinant is the product of the eigenvalues of a matrix and measure the p -dimensional volume of the data. The eigenvalues represent the variance or the scales of each of the eigenvectors. The use of the robust location and scatter

matrix developed using the MCD in the Mahalanobis distance calculation is effective in outlier detection from a multivariate dataset.

2.4 TWO ALTERNATE METHODS FOR MULTIVARIATE OUTLIER DETECTION

A paper published in *Anal. In Chem.* in 1999, titled *Outlier Detection in Multivariate Analytical Chemical Data* [27], by William J. Egan and Stephen L. Morgan, introduces two multivariate outlier detection methods called “resampling by the half-means method” and “the smallest half-volume method”. The authors claim that these two methods are simple to use, are conceptually clear, and are superior to the current best performing method Minimum Covariance Determinant. Method 1, resampling by the half-means method, uses resampling without replacement. A sample of size $n/2$ is obtained by sampling without replacement and stored in $X(i)$. Then the mean, $m(i)$, and the standard deviation, $s(i)$, of $X(i)$ is calculated. Next, the column lengths, $l(i)$, are calculated using the autoscaled original dataset X as,

$$L(i) = \sqrt{\sum_{k=1:p} (X_{k(i)} - m_k(i))^2 / s_k(i)^2}.$$

Then all of the $l(i)$ are stored in the $n \times i$ matrix L . These lengths do not have an apparent distribution that can be used to derive a statistic to define a cutoff. Thus, an empirical derived distribution is used. Plotting a histogram of the vector lengths provides a visualization of the possible outliers. The number of times each observation appears in the upper 5% of the distribution is tabulated and these are identified as possible outliers.

The second proposed method, the smallest half-volume method, uses the distances between each observation in the multivariate space expressed as vector lengths. A vector length between two observations I and j is given as,

$$l_{ij} = \sqrt{\sum(x_i - x_j)^2},$$

which is summed over all of the variables. The vector lengths are stored in a distance matrix and each column is sorted from the lowest to the highest distance. Then, for each column the first $n/2$ smallest distances are summed. Next, their mean vector and the covariance matrix are calculated. Using these location and dispersion measures, the Mahalanobis distances are calculated and the potential outliers are identified. This method is essentially the same as the Minimum Covariance Determinant method without the need to perform the eigen-decomposition and determinant calculations.

3.0 SCOPE

3.1 METHOD 1 – INTRODUCTION

The Minimum Covariance Determinant (MCD) method is used to detect possible outliers, and the outlier free data set is then used to generate the location and scatter matrices that can be used to develop robust Partial Least Squares Regression (PLSR). This procedure is repeated to overcome possible masking and to obtain a dataset that is outlier free to generate location and scatter matrices that will be used in robust PLSR.

3.2 METHOD 2 – INTRODUCTION

To identify outliers we focus on a method called resampling by half-means (RHM), introduced in 1998 by William Egan and Stephen Morgan (1998). We adapt this method for application to the detection of outliers and the development of a robust PLSR method that can be used for the analysis of Magnetic Resonance Imaging (MRI) data.

4.0 METHOD 1 – PAPER

ROBUST PARTIAL LEAST SQUARES REGRESSION AND OUTLIER DETECTION USING MINIMUM COVARIANCE DETERMINANT METHOD

4.1 ABSTRACT

Partial Least Squares Regression (PLSR) is often used for high dimensional data analysis where the sample size is limited, the number of variables is large, and the variables are collinear. One weakness of PLSR is that the method is influenced by outliers and/or influential observations. Since PLSR is based on the covariance matrix of the outcome and the predictor variables, this is a natural starting point for the development of techniques that can be used to identify outliers and to provide stable estimates in the presence of outliers. We focus on the use of the minimum covariance determinant (MCD) method for robust estimation of the covariance matrix when $n \gg p$ and modify this method for application to a magnetic resonance imaging (MRI) data set with 1 outcome and 18 predictors. We extend this approach by applying the MCD to generate robust Mahalanobis squared distances (MSD) in the Y vector and the X matrix separately and to detect outliers based on the robust MSD. We then remove these observations from the data set and compute the PLSR once more. This approach is applied iteratively until no new outliers and/or leverage points are detected. Simulation studies demonstrate that the PLSR results are improved when using this approach.

4.2 INTRODUCTION

Although there is no formal definition describing what an outlier is, the author Hawkins describes outliers as the following, “an outlier is an observation that deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism” [28]. In p -variate data, when p is greater than two, it is difficult to identify outliers visually leading to the need for methods that address this issue. Although there are several multivariate outlier detection techniques available, they are found to be unreliable or difficult to program, and require significant processing time. Techniques such as the Mahalanobis distance are based on the covariance matrix and are highly influenced by the presence of outliers. Robust multivariate outlier detection techniques such as the Minimum Volume Ellipsoid (MVE), Multivariate Trimming (MVT), M-Estimators, and Minimum Covariance Determinant (MCD) are difficult to program and require significant processing time. Rousseeuw and Driessen introduced a fast algorithm for the minimum covariance determinant estimator [35], currently considered as the best performing multivariate outlier detection method. One limitation is that the MCD method can be influenced by masking, a phenomenon that occurs when the outliers are grouped with non-outlier data due to the presence of extreme outliers. When this occurs, the points involved may not be identified as outliers providing for inaccuracies in the results obtained from the analysis. Therefore we propose a technique based on a robust Mahalanobis Squared Distance (MSD) that is computed using the MCD to generate robust location and scatter estimates. This proposed technique, which is a hybrid of MSD and MCD, will identify all possible outliers including the observations that are masked. Then it is possible to use outlier-free observations to generate the location vector and the covariance matrix that can be used to develop a robust PLSR.

We have two objectives in this study. The first objective is to identify possible outliers including observations that might be masked. Once these observations are identified and examined it can be decided, either to use, modify, or discard the observations. The second objective is to develop a robust Partial Least Squares Regression that is not influenced by possible outliers including points that might be masked.

4.2.1 Data

The data that motivated the study are behavioral response data and Magnetic Resonance Imaging (MRI) predictors. There are 305 subjects/observations. The response variable is the Digit Symbol Substitution Test (DSST). The predictor data consist of 18 MRI gray matter volumes (GMVs) in mm^3 , combined across left and right sides of the brain for 18 brain regions and the 18 brain regions are normalized. Therefore we have only one response vector Y (305×1) and our predictor matrix X is 305×18 . Multicollinearity is to be expected as the GMVs for different brain regions are taken from the same individual. We will use this data set to identify possible outliers using robust MSD and use robust MSD iteratively to develop a robust PLSR.

4.2.2 Partial Least Squares Regression (PLSR)

PLSR originated from PLS methods used by Herman World [1], [29], in econometric and sociometric fields. His son Svante World [30] used PLS in a regression framework for prediction. PLSR is a dimension reduction method best suited for multivariate datasets in the presence of multicollinearity. The response data will consist of an $i \times k$ matrix Y where k is the number of response vectors and an $i \times p$ predictor matrix, X , where p is the number of predictor

vectors and i denote the sample size. When the X predictor matrix is non-singular, ordinary multiple regression can be used. When the X matrix is singular, PLSR is one method that can be used to overcome this problem. Another method that can be used is principal components regression (PCR). In PCR the X matrix is used to generate the orthogonal components and the response variable is not included in the decomposition. Therefore, the components do not take into account the relationships between the predictor and the response variables. In PLSR the orthogonal components are generated using both the predictor data (X) and the response data (Y). Thus PLSR components explain, as much as possible, the covariance between X and Y leading to a better predictor model where you usually need fewer components. The PLSR projects the cross-covariance of the X and Y matrices into a new space, and uses the least squares regression for prediction. Both correlation and least squares regression are sensitive to outliers, thus the PLSR predictor model will be affected in the presence of outliers. Part of the computation of a PLSR uses the singular value decomposition (SVD) of the cross correlation matrix that overcomes collinearity, and works as a dimension reduction method. A brief description of the steps involved in SVD on the cross-covariance matrix is given next. The PDQ' is the decomposition result from the SVD on the cross-covariance matrix of X and Y . The P matrix is the x -loading and contains the eigenvectors from the X matrix, and XP becomes the x -score. Similarly, the Q matrix is the y -loading and contains the eigenvectors from the Y matrix, and YQ becomes the y -score. This decomposition is computed so that U has the maximum covariance with T , that is, u_1 has maximum covariance with t_1 , and u_2 has maximum covariance with t_2 , etc. Then Y is predicted using UQ' instead of X . This process is then described as follows:

$$\begin{aligned}
\text{svd}(X'Y') &= PDQ' & T &: x \text{ - score,} \\
\text{where, } T &= XP & P &: x \text{ - loading,} \\
U &= YQ & U &: y \text{ - score,} \\
X &= TP' + E & Q &: y \text{ - loading,} \\
Y &= UQ' + F
\end{aligned}$$

u_1 has maximum covariance with t_1 , etc.

$$u_1 = r_1 t_1$$

$u_2 = r_2 t_2$ etc, then use U to predict Y with y - loading Q .

4.2.3 Mahalanobis Squared Distance (MSD):

The MSD is a distance measure introduced by P. C. Mahalanobis [32] that is based on the correlations between variables. The MSD for a multivariate data vector, $x=(x_1, x_2, x_3, \dots, x_n)$, with mean $\mu=(\mu_1, \mu_2, \mu_3, \dots, \mu_n)$ and with a covariance matrix S is :

$$[6] D_M(x) = (x - \mu)^T S^{-1} (x - \mu).$$

An important fact that helps with outlier detection is that the MSD of multivariate normal data has a chi-square distribution with p degrees of freedom where p is the number of variables.

4.2.4 Minimum Covariance Determinant (MCD)

The MSD discussed previously is not robust in the presence of outliers. Rousseeuw 1984 [23] developed the Minimum Covariance Determinant (MCD), a robust shape and location estimate that can be used in the computation of the MSD to detect outliers and influential points. Given n data points, the MCD provides a sample of size h ($h \leq n$) that minimizes the determinant

of the covariance matrix. The determinant is the product of the eigenvalues of a matrix and measure the p-dimensional volume of the data. The eigenvalues represent the variance or the scales of each of the eigenvectors. The computation of the MCD proceeds as follows:

- Select a subset H_1 from n .
 - First randomly choose $p+1$ data points, subset J
 - Calculate the $T_0 = \text{avg}(J), S_0 = \text{cov}(J)$
 - If the $\det(S_0) < 0$, add one more data point.
 - Continue until $\det(S_0) > 0$
 - Compute average T_0 and average S_0
- Compute the MD using T_0 and S_0
 - $D_M(x) = \text{sqrt}((x - T_0)S_0^{-1}(x - T_0))$

Later Hardin and Rocke (2002) [34] developed a distributional fit to the Mahalanobis Squared Distance that used the Minimum Covariance Determinant for robust estimates. Since the Minimum Covariance Determinant is asymptotically normal, it has better statistical properties than other methods used for outlier detection in the multivariate setting such as the Minimum Volume Estimator (MVE) [25]. The limitation for Minimum Covariance Determinant is that it is only applicable when $n > p$, that is, the number of observations n , has to be greater than the number of variables p .

4.2.5 Detecting Outliers and Leverage points using Robust Mahalanobis Squared Distance (RMSD)

The robust MSD calculated using the MCD has a chi-square distribution with p degrees of freedom, where p is the number of variables. Generally, a chi-square value of 0.975 is used to find the cutoff for the robust MSD.

4.3 METHOD

The PLSR uses the location and scatter matrices to develop the regression model, thus the presence of outliers can influence the model. Therefore, we will use the robust location and scatter matrices generated by the robust Mahalanobis Squared Distances. Extending this approach, we will improve the predictive model by repeating the application of the robust Mahalanobis squared distances until all possible outliers are detected thereby identifying the masked outliers, and by using the robust location and scatter matrices generated by zooming into the data during each repeat in robust PLSR. The algorithm used is given below.

Step1:

The entire data set with potential outliers is used in PLSR to predict the one outcome variable using the 18 predictors. The effectiveness of the model is assessed by observing the Root Mean Squared Error Predicted (RMSEP) and R^2 using the cross validation leave-one-out-method as diagnostic measures.

Step2:

Run1: The robust Mahalanobis Squared Distances (MSD) will be calculated for both the Y vector and the X matrix. The Minimum Covariance Determinant will be used to generate the robust MSD. To achieve the first objective given in the introduction, we identify the possible outliers that are beyond the robust MSD cutoff. Next, to meet the second objective of obtaining robust location and scatter matrices, observations that are outside of the robust MSD cutoff, will be removed from the dataset. Then the outlier free data is used to compute the location and scatter matrices. We will then use the location and scatter values to compute the robust PLSR and to calculate the RMSEP and the R^2 . We are investigating a Chi-squared cutoff for 0.975, a Chi-squared cutoff for 0.999, and the largest 5% of the robust MSD as possible cutoffs for the identification of outliers.

Step3:

Run 2: We will use the reduced data set generated in step 2 by removing the outliers. Then, the robust MSD will be calculated for both the Y vector and the X matrix. The observations that are outside of the robust MSD cutoff will be identified as possible outliers that were masked by the outliers identified during the first run. Observations that are outside of the robust MSD cutoff will be removed from the dataset for the purpose of generating a more robust location and scatter values. We will use the reduced data set to run the PLSR and to calculate the RMSEP and the R^2 .

Step4:

Repeat this method until there are no more outliers detected.

The method used to generate the RMSEP and the R^2 in the PLSR is the leave-one-out cross validation method. In this cross validation method the PLSR is computed for $n - 1$ observations and the model will be used to predict the observation left out, and the RMSEP and the R^2 is recorded. This will be repeated n times, with a different observation left out each time and the PLSR will be run. At the end the average of the RMSEP and the average R^2 will be calculated. The RMSEP and the R^2 for each run is then compared.

4.3.1 Simulation

Simulation 1: The cross-correlation matrix from the above mentioned dataset with 305 observations was used to generate 1000 data points. The new dataset has one outcome variable and 18 predictor variables similar to the MRI data. From 1000 data points generated, 700 were generated from a multivariate normal distribution, $N_p(\mu_p, \epsilon_p)$, where μ_p is the mean vector from the 305 observations and ϵ_p is the covariance matrix from the MRI data. The other 300 observations were generated to be outliers by shifting the mean. The first one hundred outliers were shifted by four times the standard deviation of each of the variables. The second hundred outliers were shifted by five times the standard deviation of each of the variables. The third, hundred outliers were shifted by six times the standard deviation of each of the variables. The four steps described in the methods will be then performed on the new dataset generated. This process then will be repeated 500 times and the results will be averaged.

Assessment of minimal sample size for one response and 18 predictors:

For this purpose we ran the above 4 steps on the MRI data by reducing the sample size one observation at a time. We started with the entire data set of 305 observations, then we

removed 1 observation and used the sample of 304 observations and ran the 4 steps described above. This was repeated until the determinant of the covariance matrix is < 0 .

Simulation 2: We generated 1000 data points using the covariance matrix from the MRI data based on a single outcome and 18 predictor variables. We followed the same 4 steps given above on the 1000 data points. Then we repeated these 4 steps by reducing the sample size by 1 until the determinant of the covariance matrix is < 0 . We then repeated these steps 500 times and the results were generated by averaging the results.

4.4 RESULTS

Figure 4-1, Tables 4-1 and Table 4-2 results are from MRI data. Figure 4-2 and Table 4-3 results are from the simulation study1.

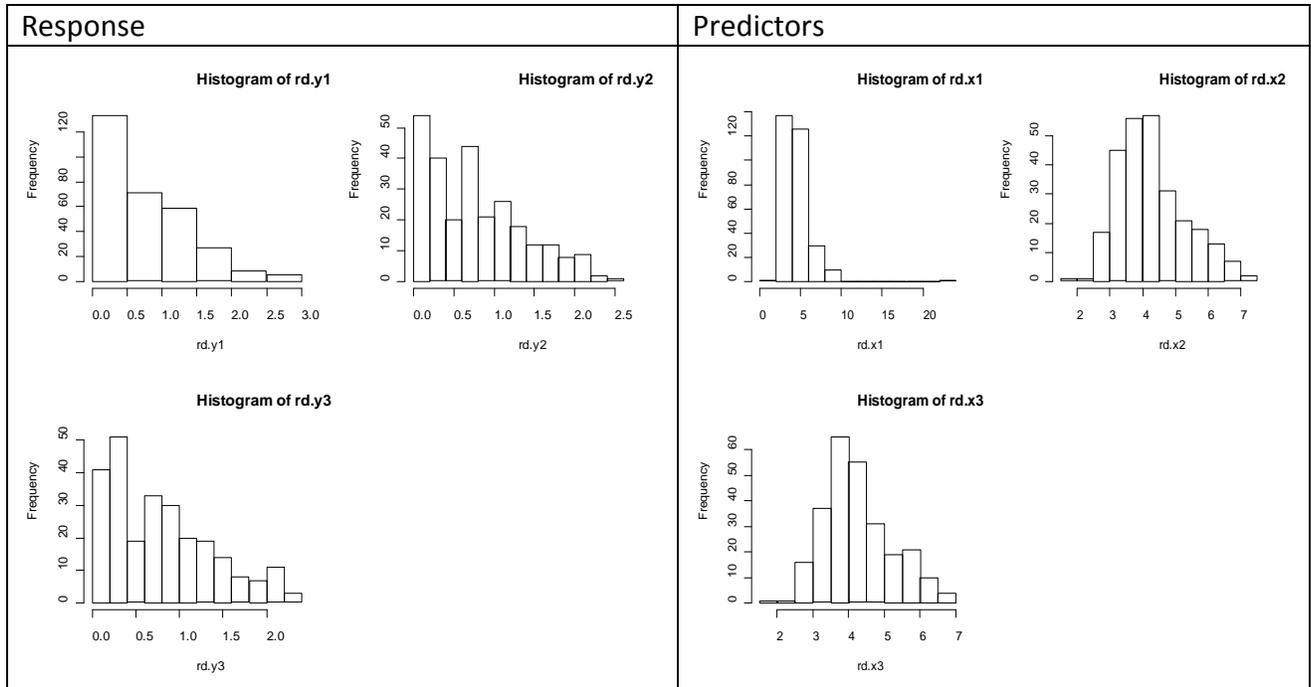


Figure 4-1: MRI Data - Robust Mahalanobis Squared Distances for the three repeats

Table 4-1: MRI data - Chi-squared distributed -cutoff >0.975

	N	RMSEP – 1 components	RMSEP – 2 components	R ² – 1 components	R ² – 2 components
Full dataset	305	13.29	13.3	81.9958	94.30
Repeat 1	244	11.8	11.81	67.4432	95.892
Repeat 2	223	11.74	11.39	23.191	96.019
Repeat 3	211	11.04	11.07	82.9921	96.186

Table 4-2: MRI data - Chi-squared distributed - cutoff > 0.999

	N	RMSEP – 1 components	RMSEP – 2 components	R ² – 1 components	R ² – 2 components
Full dataset	305	13.29	13.3	81.9958	94.30
Repeat 1	269	11.89	11.93	81.8858	95.4985
Repeat 2	260	11.63	11.6	52.5282	95.7627
Repeat 3	256	11.27	11.31	68.658	95.74073333

Table 4-3: MRI Data – cutoff largest 5% of the robust MSD

	N	RMSEP – 1 components	RMSEP – 2 components	R ² – 1 components	R ² – 2 components
Full dataset	305	13.29	13.3	81.9958	94.30
Repeat 1	276	11.46	11.5	83.338	94.586
Repeat 2	248	10.24	10.28	83.3816	93.23
Repeat 3	225	9.311	9.327	81.761	94.332

Figure 4-1 shows the distribution of the robust MSD for the response and the predictor variables for the MRI data. Table 4-1 shows the results of the PLSR with the chi-squared cutoff greater than 0.975 for the three runs, Table 4-2 shows the same for chi-squared cutoff greater than 0.999, and Table 1.3 shows the results of the PLSR for the three runs of PLSR with the outlier cutoff as the largest 5% of the robust MSD. The n value for runs 1, 2, and 3 is the sample size after removing the possible outliers from the previous run. For all three tables, the RMSEP decreases in value for either choice of one or two components while the R² for x increases.

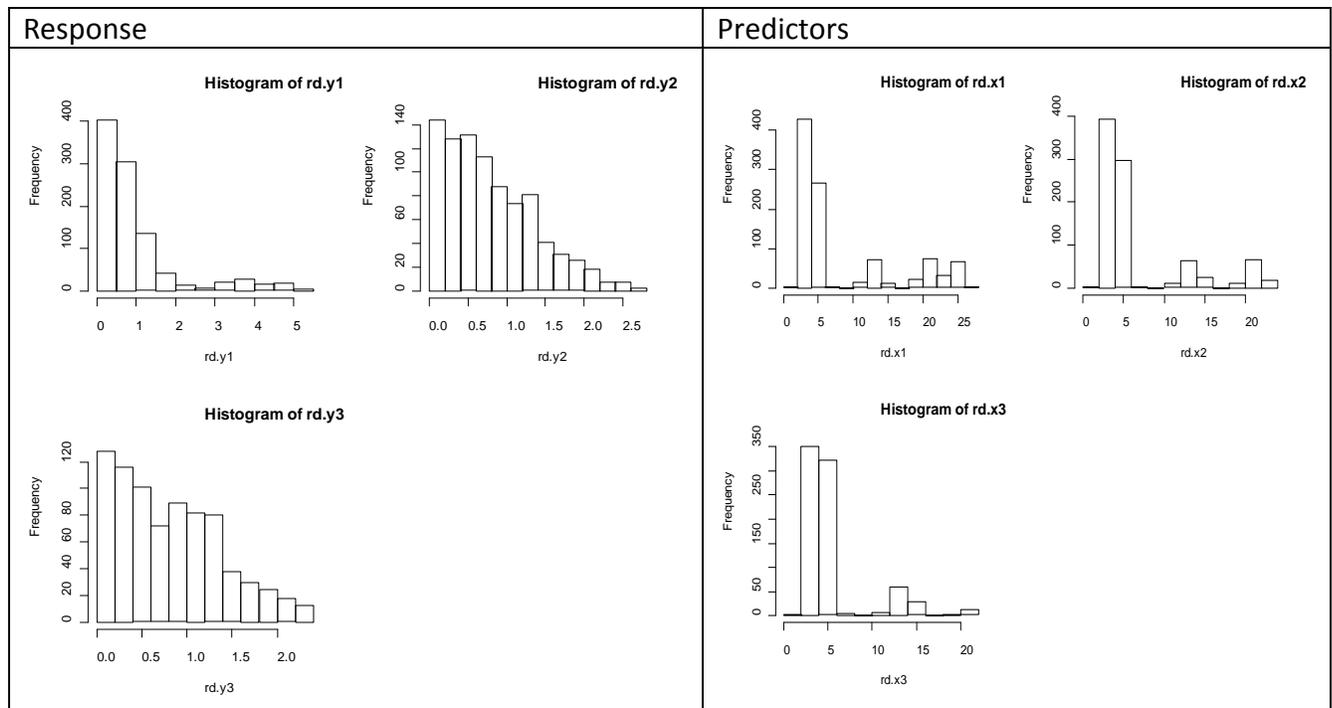


Figure 4-2: Simulation1 data - Robust Mahalanobis Squared Distances for the three repeats

Table 4-4: Simulation1 – Chi-squared distributed – cutoff > 0.975

	n	RMSEP – 1 components	RMSEP – 2 components	R ² – 1 components	R ² – 2 components
Full dataset	1000	17.33	17.01	96.46	97.77
Repeat 1	691	10.97	10.98	82.4425	94.6966
Repeat 2	658	10.24	10.25	82.5751	93.972
Repeat 3	647	9.908	9.936	81.84111	91.2016

Table 4-5: Simulation1 – Chi-squared distributed – cutoff > 0.999

	n	RMSEP – 1 components	RMSEP – 2 components	R ² – 1 components	R ² – 2 components
Full dataset	1000	17.37858	17.0682	96.46	97.75
Repeat 1	696	11.09299	11.10567	81.8107	93.3170
Repeat 2	680	10.42653	10.43953	81.7388	92.6241
Repeat 3	672	10.17238	10.18259	81.8436	92.042

Table 4-6: Simulation1 – cutoff largest 5% of the robust MSD

	n	RMSEP – 1 components	RMSEP – 2 components	R ² – 1 components	R ² – 2 components
Full dataset	1000	17.41	17.12	96.55	97.79
Repeat 1	920	13.60	13.56	95.558	97.095
Repeat 2	846	10.67	10.68	94.52	96.10
Repeat 3	764	9.54	9.55	92.7163	94.7228

Figure 4-2 and the Tables 4-4, 4-5, and 4-6, shows the results from the simulation 1.

Figure 4-2 shows the distribution of the robust MSD for the response and the predictor variables for the simulated data. Tables 4-4, 4-5, and 4-6 present the results from the PLSR, with a chi-squared cutoff greater than 0.975, chi-squared cutoff greater than 0.999 and the largest 5% of the robust MSD respectively. The n value for runs 1, 2, and 3 is the sample sizes after removing the possible outliers from the previous runs. The RMSEP goes down, either you choose one or two components and the R^2 for x stays above 90% for all three different cutoffs.

Next, to find the sample size requirement we ran repeated robust MSD reducing the sample size by 1 to obtain information related to the effect of sample size on the results. Figure 4-3 shows the results for the MRI data, and Figure 4-4 shows the results for the simulated data. The y-axis is the RMSEP difference from the previous repeat, while the x-axis the sample size. A positive value on the y axis indicates a reduction in error.

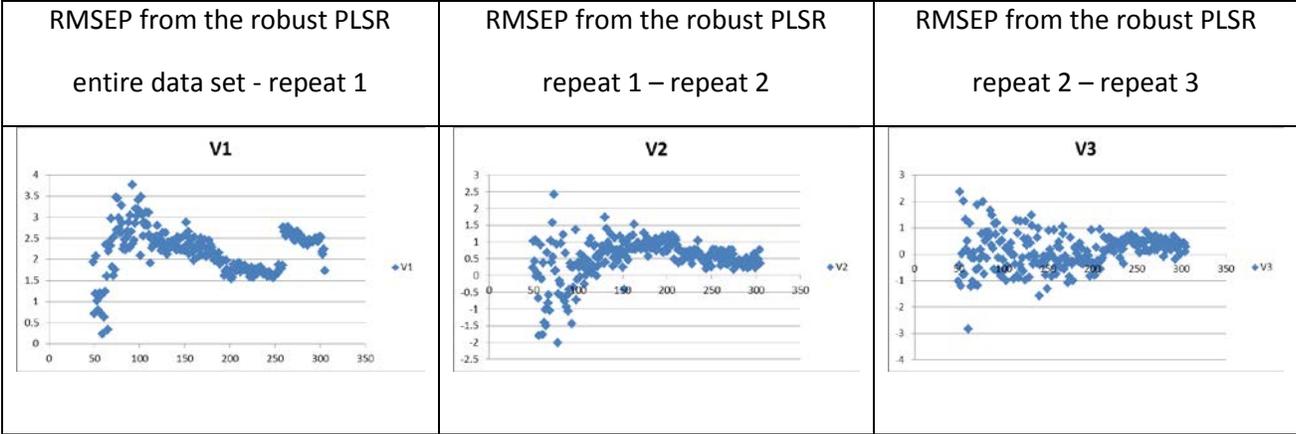


Figure 4-3: MRI data - Error differences between the repeats, with 2 components chosen for PLSR

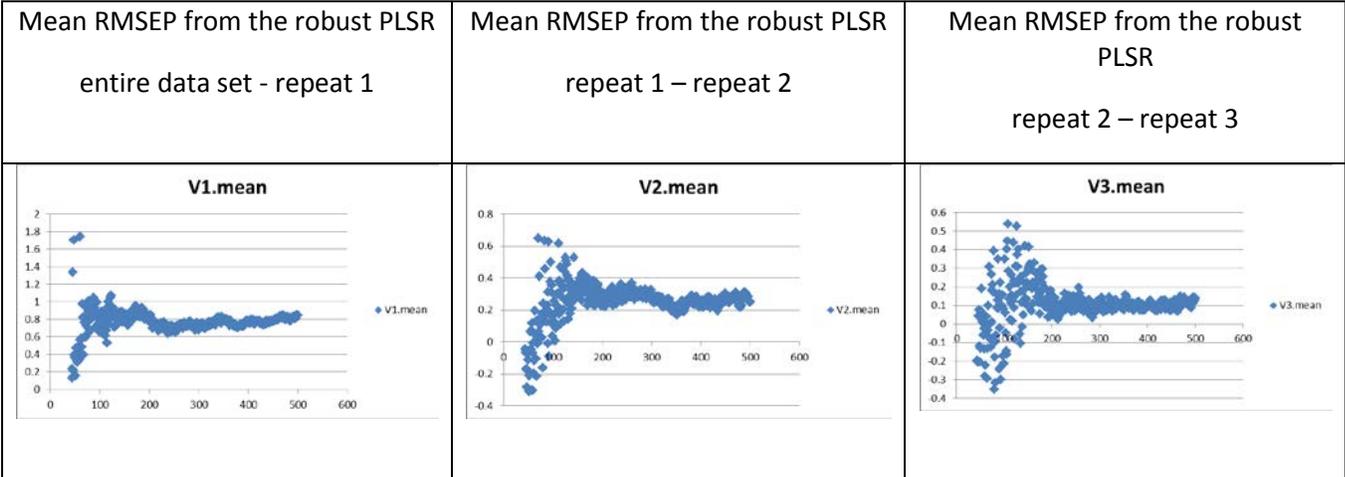


Figure 4-4: Simulation-1000 data points generated - Error differences between the repeats, with 2 components for PLSR

4.5 DISCUSSION

Tables 4-1, 4-2, and 4-3 demonstrate that the RMSEP is decreasing with each repeat for a chi-squared cutoff of 0.975, a chi-squared cutoff of 0.999, or the largest 5% of the robust MSD. This indicates that the location and the scatter matrices generated to be used in the robust PLSR after each layer of outliers removed are more robust than the previous layer. Whether to use the chi-squared cutoff of 0.975, the chi-squared cutoff of 0.999, or the largest 5% of the robust MSD is a call that the investigator will have to make by investigating the distribution of the robust MSD and the identified outliers. If the robust MSD values do not appear to have a chi-squared distribution you can use an empirical method where you compute the cutoff based on the largest 5% of the robust MSD as the possible outliers. The repeated layers identify possible outliers that were masked. Once these points are identified and examined it is possible to decide if the data will be used as is, or if there is a need to modify, correct, or to eliminate values.

Next, figure 4-3 presents the MRI dataset used in this paper. The method of repeated robust PLSR gives favorable results when the sample size is approximately greater than 150. This is confirmed by the simulation study results given in Figure 4-4. The method is stable and gives favorable results when the sample size is greater than about 150. Another important result from the simulation study is that the RMSEP values are stable for $n > 150$ for repeat 1 (after removing the first layer of outliers) and repeat 2 (after removing the 2nd layer of outliers). For repeat 3 (after removing the 3rd layer of outliers) the RMSEP values tend to be more unstable even when $n > 150$. Therefore it might not be advisable to use the location and scatter matrices generated from repeat 3, but to use the location and scatter matrices generated from repeat 2 in the robust PLSR for these datasets.

5.0 METHOD 2 – PAPER

ROBUST PARTIAL LEAST SQUARES REGRESSION AND OUTLIER DETECTION USING RESAMPLING BY HALF-MEANS METHOD

5.1 ABSTRACT

Partial Least Squares Regression (PLSR) is often used for high dimensional data analysis where the sample size is limited, the number of variables is large, and the variables are collinear. The PLSR results can be influenced by outliers and/or influential observations pointing to the need for methods to identify these observations. To identify outliers we focus on a method called resampling by half-means (RHM), introduced in 1998 by William Egan and Stephen Morgan published in Analytical Chemistry [27]. We adapt this method for Magnetic Resonance Imaging (MRI) data to detect outliers and then to develop a robust PLSR model. This method can be used for small or large datasets overcoming the limitation of the leading multivariate outlier detection methods such as Minimum Covariance Determinant method that cannot be used for small sample sizes ($n < p$). The method uses sampling without replacement to generate i samples of size $n/2$, and calculate the vector lengths for all observations by auto-scaling the original dataset with the medians and the absolute median deviations of the i samples. Outliers are identified empirically or using the chi-squared cutoff. The outlier free dataset is then used to generate the scatter and the location values to be used in the robust PLSR.

5.2 INTRODUCTION

The existing multivariate outlier detection techniques are mostly unreliable or difficult to program, and require significant processing time. Techniques based on the covariance matrix such as Mahalanobis distances are influenced by the presence of outliers. Other robust multivariate outlier detection techniques such as the Minimum Volume Ellipsoid (MVE), Multivariate Trimming (MVT), M-Estimators, and the Minimum Covariance Determinant (MCD) are difficult to program and require significant processing time. A fast algorithm for the minimum covariance determinant estimator [10], introduced by Rousseeuw and Driessen, is currently considered as the best performing multivariate outlier detection method. The Minimum Covariance Determinant method is valid when $n > p$, and thus not valid for smaller ($n < p$) datasets. The proposed resampling by half-means method is valid for larger ($n > p$) and smaller ($n < p$) datasets. Thus resampling by half-means leads to robust location and covariance estimates that can be used for robust PLSR over a wide range of applications.

The primary objective of the resampling by half-means method in our study is to explore a method that can be used regardless of the size of the dataset. Therefore PLSR which can be used for smaller datasets will be robust in its estimates.

5.2.1 Data:

The data used in the study comes from an MRI study with 305 subjects/observations. The response variable, Digit Symbol Substitution Test (DSST) is a neuropsychology test. The predictors are 18 normalized MRI gray matter volumes (GMVs) in mm^3 , combined across left

and right sides of the brain for 18 brain regions. Multicollinearity is to be expected as the gray matter volumes for different brain regions are taken from the same individual.

5.2.2 Partial Least Squares Regression (PLSR)

The PLSR originated from PLS methods used by Herman World [1], [2], in econometric and sociometric work. His son Svante World [3] used PLS in a regression framework for prediction. PLSR is a dimension reduction method best suited for multivariate datasets in the presence of multicollinearity. The response data will consist of an ixk matrix Y , where k is the number of response vectors and an ixp predictor matrix X , where p is the number of predictor vectors (usually $k < p$). When the X predictor matrix is non-singular ordinary multiple regression can be used. When the X matrix is singular, to overcome this problem you can use PLSR.

Another method used is Principal Components Regression (PCR). In PCR the X matrix is used to generate the orthogonal components and thus these components only explain the predictor data.

Therefore the components do not take into account the relationships between the predictor and the response variables. In PLSR the orthogonal components will be generated using both the predictor data (X) and the response data (Y). Thus the PLSR components explain as much of the covariance between X and Y as possible, leading to a better predictor model where you usually need fewer components in the model than in PCR. During PLSR, the singular value decomposition uses the cross-correlation matrix to project the X and Y variables in to a new space and uses least squares regression for prediction. Both correlation and least squares regression are sensitive to outliers, thus the PLSR predictor model will be affected by the presence of outliers. PLSR uses the singular value decomposition (SVD) on the cross correlation

matrix that overcomes collinearity, and works as a dimension reduction method. This leads to the following:

$$\begin{aligned}
 \text{svd}(X'Y') &= PDQ' & T &: \text{x - score,} \\
 \text{where, } T &= XP & P &: \text{x - loading,} \\
 U &= YQ & U &: \text{y - score,} \\
 X &= TP' + E & Q &: \text{y - loading,} \\
 Y &= UQ' + F
 \end{aligned}$$

u_1 has maximum covariance with t_1 , etc.

$$u_1 = r_1 t_1$$

$u_2 = r_2 t_2$ etc, then use U to predict Y with y - loading Q .

5.3 METHOD

The method can be applied to the $X_{n \times p}$ matrix (predictors) and the $Y_{n \times k}$ matrix (response) separately. If we consider the X matrix which has n observations and p predictor variables, first we would like to identify the possible outliers. First, using sampling without replacement we randomly pick a sample of size $n/2$. We generate i such samples of size $n/2$ and call each sample $X(i)$. Next, we calculate the medians $m(i)$ and the median absolute deviation $s(i)$ of the p predictors for the $X(i)$ samples. Then use the $m(i)$ and the $s(i)$ to auto-scale the original X matrix and calculate the column vector lengths $l(i)$, for each sample $X(i)$ for all the n observations as given below:

$$l(i) = \sqrt{\sum_{t=1}^p ((X_t - m_t(i)) / s_t(i))^2}.$$

Next, sort the lengths in each $l(i)$. The sorted vector lengths then can be examined empirically by observing the distribution by plotting the lengths. The squared vector lengths should follow a chi-squared distribution. If the data follow a chi-squared distribution it is possible to use a 0.999 or 0.975 cutoff to identify the possible outliers. If the lengths do not follow a chi-squared

distribution, which can be a result of the sampling method used, then identify the observations that have the largest lengths that appear in the extreme 5% of the data in all or most of the samples. This last cutoff method can be used even if the lengths have a chi-squared distribution. We will use the largest 5% as the cutoff for the results shown in this paper. Figure 5-1 flow chart illustrates the method further.

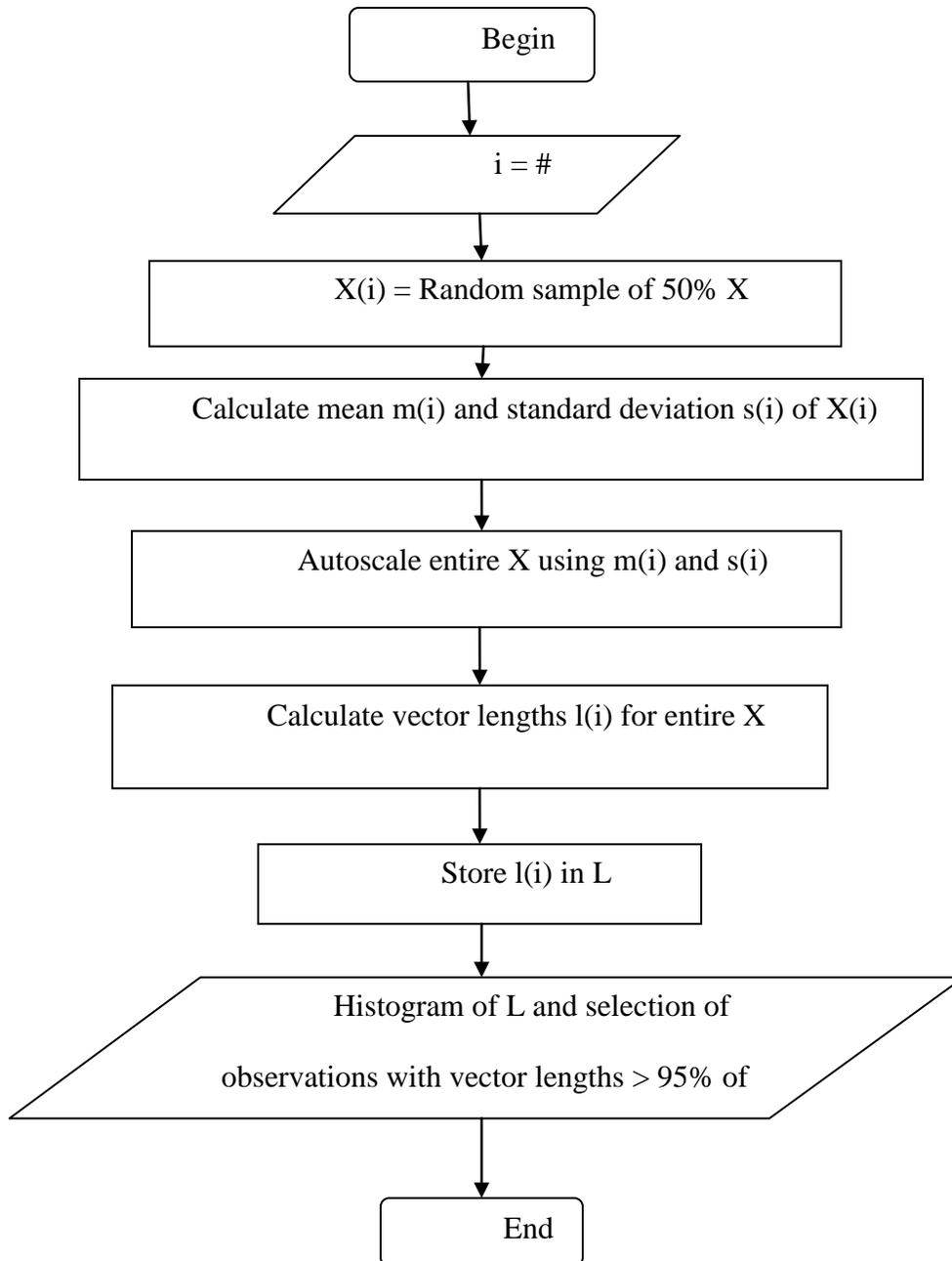


Figure 5-1: Flow chart showing the steps used in the RHM method of outlier detection

For the simulation study for the larger dataset ($n > p$), we generated 300 data points using the covariance matrix of the actual data set of 305 observations and for the simulation of smaller dataset, we used the outlier free data set identified by the method used on the larger dataset(305 subjects) to randomly select 15 subjects. Then we introduce 2 outliers by shifting the mean. Therefore the small dataset used has 17 subjects (n) and the 17 subjects are used with 18 predictors (p) thus $n < p$.

5.4 RESULTS

Results – Simulation

Table 5-1: After 1000 rounds the subjects with the largest 5% of the vector lengths sorted by ID for (X)

id	%
1	1.780645
6	100
7	100
10	6.051613
14	6.535484
25	0.019355
33	2.432258
39	3.23871
43	3.864516
47	100
50	100
63	10.25161
66	0.632258
88	12.42581
89	12.5871
96	100
144	0.36129
150	100
170	37.57419
180	40.56129
184	99.99355
195	100
203	100
205	64.60645
211	100
212	0.129032
222	75.97419
254	82.58065
256	0.141935
278	35.36774
287	99.74194
295	0.03871
296	100
301	3.109677

Table 5-1 provides a list of the subjects with the largest vector lengths sorted by ID. The % column gives the percentage of the times the subject had vector lengths in the largest 5%. The highlighted subjects have vector lengths in the largest 5%, 100% of the 1000 rounds or almost 100%.

Table 5-2: Comparison of PLSR simulation results for large datasets

	The entire dataset	Outlier free data,avg results
n	300	276
RMSEP	13.29972	10.7869
SD		0.05929556

Table 5-2 provides a comparison of the Root Mean Squared Error Predicted from the PLSR for the entire dataset and the dataset without the outliers for large datasets ($n > p$).

Table 5-3: Small sample simulation results for 2 trials

id	length	id	length
6	6.0578	13	6.3307
13	7.1798	14	6.5019
3	7.9936	6	6.5114
14	8.4699	3	9.1859
12	12.3805	12	10.6904
7	16.7109	11	12.5987
11	17.6238	7	16.6619
9	17.6439	9	18.8706
15	29.1166	15	27.7815
8	33.8549	4	28.3092
4	37.8842	8	32.0146
5	47.5412	5	36.3295
2	69.3513	2	64.3679
1	227.5194	10	207.0673
10	308.1954	1	225.9017
16	3269.573	17	2422.573
17	3305.677	16	3063.662

Table 5-3 presents the vector lengths for a small dataset ($n < p$) for two iterations sorted by the vector lengths. The highlighted two subjects are the two outliers introduced to the dataset. And they had the largest vector lengths for all the 1000 trials.

Table 5-4: Comparison of PLSR results for smaller datasets

	n	Average RMSEP-second component
The entire dataset	17	15.26175
Outlier free data	15	12.039

Table 5-4 also presents a comparison of the Root Mean Squared Error Predicted from the PLSR for the entire dataset and without the dataset outliers for small datasets ($n < p$).

Results – Example Data

Table 5-5: MRI data (305 observations) for 5 iterations (samples) for the predictors (X)

	id	length								
1	288	3.070653	288	2.656197	288	3.086783	288	2.786386	288	3.016295
2	200	4.697258	200	4.482265	200	4.984269	215	4.952709	215	4.799336
3	125	4.726309	215	4.93209	215	5.063546	200	5.101561	200	4.833593
4	215	5.448981	228	5.310261	228	5.832493	125	5.288661	234	5.271599
5	147	5.529308	147	5.383025	147	5.997001	147	5.633618	125	5.912405
.										
296	6	59.39116	195	51.69783	195	59.7656	195	56.54462	184	56.2545
297	203	59.90316	203	56.28099	203	65.04184	50	59.03495	203	57.83499
298	195	60.27253	50	57.43941	6	66.95201	203	59.5086	6	61.72536
299	50	62.12107	6	57.85139	50	67.95672	6	60.86533	50	66.00247
300	150	70.81358	211	59.50692	211	71.41161	211	61.97807	211	67.15726
301	7	71.62423	150	62.61466	7	79.30918	150	63.71189	150	67.63135
302	211	77.52548	7	62.79242	150	80.26347	7	65.34198	7	68.98318
303	96	81.74897	96	70.99351	96	86.6193	96	69.7778	96	76.06066
304	296	103.7884	296	95.82588	296	109.7778	296	100.016	296	111.3996
305	47	483.2647	47	507.1873	47	524.0656	47	510.1088	47	550.2659

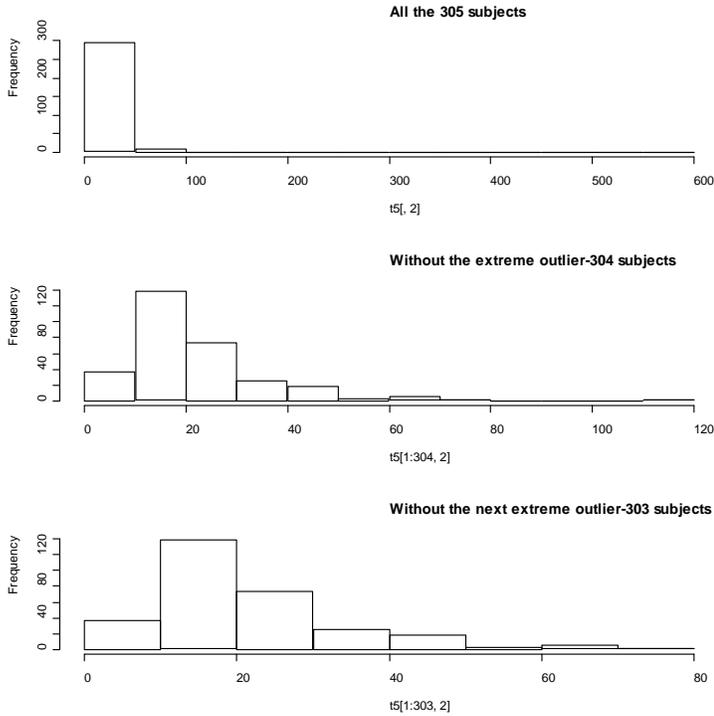


Figure 5-2: Vector length distributions for the predictors of MRI example data

In figure 5-2 the first histogram gives the vector lengths for all the subjects, the histogram in the center, is for the vector lengths without the subject with the largest vector length. The histogram at the end is for the vector length without the subjects with the two largest vector lengths.

Table 5-6: MRI example data after 500 iterations the subjects with the largest 5% of the lengths sorted by ID for the X

id	Total
1	5
6	500
7	500
10	35
14	25
33	9
39	17
43	25
47	500
50	500
63	47
66	3
88	51
89	63
96	500
144	1
150	500
170	202
180	197
184	500
195	500
203	500
205	331
211	500
212	1
222	383
254	401
278	190
287	498
295	1
296	500
301	15

Table 5-6 lists the subjects with the largest vector lengths sorted by ID for the example MRI data sorted by ID for the X (18 brain volume regions) matrix.. The total column gives the number of times the subject had vector lengths in the largest 5% for the 500 iterations.

Table 5-7: MRI data (305 observations) for 5 iterations (samples) for the response (Y)

	id	length								
1	20	0.00126	12	0.001574	25	0	20	0.00126	12	0
2	25	0.00126	25	0.001574	55	0	25	0.00126	76	0
3	55	0.00126	55	0.001574	105	0	55	0.00126	80	0
4	105	0.00126	76	0.001574	114	0	105	0.00126	155	0
5	114	0.00126	80	0.001574	156	0	114	0.00126	161	0
6	141	0.00126	105	0.001574	221	0	141	0.00126	177	0
7	156	0.00126	114	0.001574	234	0	156	0.00126	193	0
8	166	0.00126	155	0.001574	265	0	166	0.00126	203	0
9	179	0.00126	156	0.001574	274	0	179	0.00126	248	0
10	180	0.00126	161	0.001574	285	0	180	0.00126	262	0
.										
296	126	4.38681	194	5.114508	126	4.72349	126	4.38681	194	4.72349
297	102	4.689262	50	5.857521	50	5.397472	102	4.689262	50	5.054864
298	171	4.689262	102	6.247917	102	5.397472	171	4.689262	102	5.751312
299	50	5.001795	171	6.247917	171	5.397472	50	5.001795	171	5.751312
300	52	5.999886	52	7.935437	52	6.880232	52	5.999886	243	6.880232
301	243	6.715688	243	7.935437	243	7.279005	243	6.715688	52	7.279005
302	211	7.088712	211	8.388801	211	7.68901	211	7.088712	211	7.279005
303	238	7.088712	238	8.388801	238	7.68901	238	7.088712	238	7.279005
304	244	7.088712	244	8.388801	244	7.68901	244	7.088712	244	7.279005
305	255	7.088712	255	8.388801	255	7.68901	255	7.088712	255	7.279005

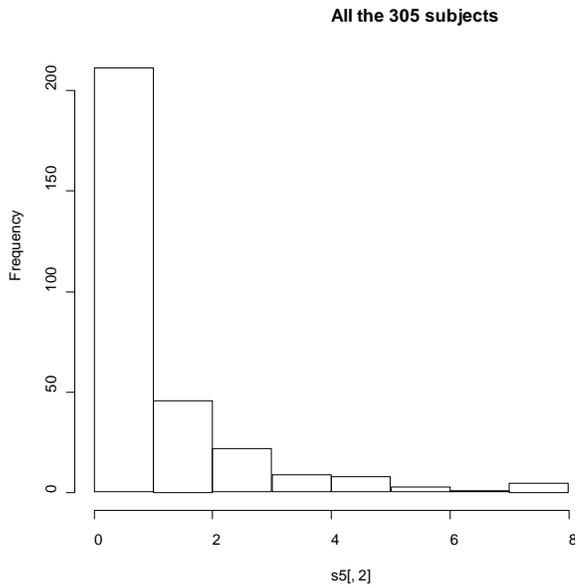


Figure 5-3: Vector lengths for all the 305 subjects for the Y vector

Table 5-7 provides the vector lengths for the Y vector sorted by the vector length for 5 iterations. Figure 5-3 is the histogram of the vector lengths for all the 305 subjects.

Table 5-8: After 500 iterations the subjects with the largest 5% of the lengths sorted by ID for the Y vector

id	Total
1	2
4	1
50	500
52	500
69	42
102	500
107	437
116	236
126	499
136	498
151	75
171	500
175	63
189	425
194	499
211	500
232	264
238	500
243	500
244	500
255	500
261	458
301	1

Table 5-8 lists the subjects with the largest vector lengths sorted by ID for the MRI data example for the Y vector. The total column gives the number of times the subject had vector lengths in the largest 5%.

Table 5-9: Comparison of PLSR results for the MRI example data

	n	RMSEP-second component
The entire dataset	305	13.29972
Outlier free data	281	10.77068

Table 5-9 provides a comparison of the Root Mean Squared Error Predicted from the PLSR for the entire dataset and the dataset without the outliers for the MRI data example.

5.5 DISCUSSION

The resampling by half-means method appears to accurately identifies outliers for large ($n > p$) and small ($n < p$) datasets. This is an advantage of the resampling by half-means method when compared to the leading minimum Covariance determinant method where it is valid only for large ($n > p$) datasets. The outlier cutoff method used in all the results is the largest 5% of the lengths calculated.

Simulation results on the Table 5-1 presents the % of the times a subject is identified as an outlier during the 1000 trials. Within each trial the method is run 500 times. There are 10 subjects that were identified as outliers 100% of the times. Two more subjects were identified greater than 99% of the times. Therefore these outliers were identified consistently. In Table 5-2 the diagnostic measures from the PLSR on the entire dataset and the outlier free dataset are presented. The RMSEP decreases from 13.29972 to 10.7869 on average, thereby improving the PLSR estimates. The small sample simulation study identified the two outliers introduced to the dataset with significantly larger lengths in all the 1000 trials (Table 5-3, only 2 trials shown). The

PLSR diagnostic measures are given for the small sample simulation study in Table 5-4. The RMSEP reduces from 15.26 to 12.01 when the outlier free dataset is used for the PLSR estimation.

Next the results are given for the example dataset of 305 subjects. Table 5-5 gives the lengths calculated for the subjects for 5 iterations and there are two large values indicating possible extreme outliers and the Figure 5-2 plot 1 gives the histogram for lengths for all the subjects. The next two plots gives the histograms of the lengths with the extreme lengths removed one at a time. The last plot indicates possibly more outliers among the data. In Table 5-6, the largest 5% is used as the outlier cutoff and the number of times the subjects had lengths belonging to the largest 5% for the 500 iterations of the method. The subjects who belong to the largest 5% of the lengths were selected as outliers. Thus 12 subjects were identified as possible outliers from the predictors. Table 5-7 gives us the lengths calculated for the response variable for the 305 subjects and the Figure 5-3 gives the histogram for the lengths and shows the existence of possible outliers. Table 5-8 gives the possible outliers identified using the largest 5% as the cutoff for 500 iterations of the method in the response variable. Eight subjects were identified in the largest 5% of the lengths for all 500 iterations. Three more subjects were in the largest 5% of the lengths for over 95% of the 500 iterations. Table 5-9 gives the Root Mean Square Error Predicted (RMSEP) for the PLSR for the entire data set and for the PLSR estimates calculated using the outlier free dataset used to calculate the correlation matrix and the means used in PLSR thus improving PLSR estimates. This is shown by the reduction in RMSEP.

Simulation results for the large datasets ($n > p$) and small datasets ($n < p$) the RHM method identified the possible outliers and thereby improved the PLSR estimates. This was also seen when RHM method implemented on the MRI data example of the 305 subjects.

6.0 SUMMARY

Our method 1, the repeated Minimum Covariance Determinant with Mahalanobis Squared Distance for multivariate outlier detection was successful in identifying possible outliers and possible masked outliers, which are outliers that did not get identified as possible outliers due to extreme outliers in the data. Method was applied to MRI brain volume region data for 305 subjects with one outcome variable DSST. Application of the cross-covariance matrix and the means of the then outlier free dataset on the PLSR provide a more robust regression model that had a reduced root mean squared error predicted. The combination of the repeated Minimum Covariance Determinant on Mahalanobis Squared Distance, for multivariate outlier detection and PLSR was successful in developing a more robust prediction model. The limitation of the method is it can be only used on large ($n > p$) datasets, where n is the number of observations and p is the number of variables. Therefore method 1, suitable for MRI brain volume data but might not be suitable for MRI brain voxel data. Method 1 can be used on Positron Emission Tomography (PET) data, and other non-imaging applications as long as the dataset confirms to $n > p$.

Method 2 can be used on large ($n > p$) or small ($n < p$) datasets thereby overcoming the limitation of method 1. The Resampling by Half Means was introduced in 1998 in Analytical Chemistry [27] and we introduced this method to imaging data. We also used the median and the median absolute deviation in the place of the mean and the standard deviation to obtain more consistent results. The success of the application of this method for small samples ($n < p$) in MRI brain volume data was important. Once the possible outliers were identified the cross-covariance matrix and the means were used in the PLSR model to develop a more robust prediction model. The results from the PLSR showed that the prediction model was more robust and had a reduced

root mean squared error predicted. It is important to note that method 2 can be applied to brain voxel data where the number of observation/subjects will be smaller than the number of voxels, and thereby developing more accurate PLSR model.

BIBLIOGRAPHY

1. Wold H. "Nonlinear Estimation by Iterative Least Squares Procedures." *Research Papers in Statistics* (1966): 411-444.
2. Geladi, P., Kowalski B. R. "Partial Least-Squares Regression: a Tutorial." *Analytica Chimica Acta* 185 (1986): 1-17.
3. Wold S., Martens H., Russwurm H. (Eds.), "Food Research and Data Analysis." Applied Science Publishers, London, (1983).
4. Wold S., Kowalski B. (Ed.), "Mathematics and Statistics in Chemistry." Reidel, Dordrecht, Chemometrics: 1984.
5. Abdi H. "Partial Least Squares Regression (PLS regression)." *Encyclopedia for Research Methods for the Social Sciences* (2003): 792-795.
6. Baker K. "Singular Value Decomposition Tutorial." 2005. at [www.ling.ohio-state.edu/~kbaker/pubs/Singular Value Decomposition Tutorial.pdf](http://www.ling.ohio-state.edu/~kbaker/pubs/Singular%20Value%20Decomposition%20Tutorial.pdf) (2005)
7. McIntosh A. R., et al. "Spatial Pattern Analysis of Functional Brain Images Using Partial Least Squares." *Neuroimage* 3.3 (1996): 143-157.
8. Hubert M., Vandenberghe B. K. "Robust Methods for Partial Least Squares Regression." *Journal of Chemometrics* 17.10 (2003): 537-549.
9. Wold H., "Estimation of Principal Components and Related Models by Iterative Least Squares." *Multivariate Analysis*, Academic Press, New York, (1966), 391-420.
10. De Jong S., SIMPLS: "An Alternative Approach to Partial Least Squares Regression." *ChemometricsIntell. Lab. Syst.* (1993); 18:251-263.
11. Wakelin I. N., Macfie H. J. H. "A Robust PLS Procedure." *Journal of Chemometrics* 6.4 (1992): 189-198.
12. Cummins D. J., Andrews C.W. "Iteratively Reweighted Partial Least Squares: A Performance Analysis by Monte Carlo Simulation." *J. Chemometrics* (1995); 9:489-507.
13. Pell R. J. "Multiple Outlier Detection for Multivariate Calibration Using Robust Statistical Techniques." *Chemometrics Intell. Lab. Syst.* (2000); 52:87-104.
14. Gil J. A., Romera R. "On Robust Partial Least Squares (PLS) methods." *J. Chemometrics* (1998); 12:365-378.

15. Stahel W. A. "Robust Estimation: Infinitesimal Optimality and Covariance Matrix Estimators." Ph.D. thesis, ETH, Zurich, (1981).
16. Donoho D. L. "Breakdown Properties of Multivariate Location Estimators," Ph.D. Qualifying paper, Harvard University, (1982).
17. Hubert M, Verboven S. "A robust PCR Method for High-Dimensional Regressors." *Chemometrics* (2003); 17:438–452.
18. Hubert, Mia, Rousseeuw P.J., Branden K. V. "ROBPCA: A New Approach to Robust Principal Component Analysis." *Technometrics* 47.1 (2005): 64-79.
19. Hubert, Mia, Debruyne M. "Minimum Covariance Determinant." *Wiley Interdisciplinary Reviews: Computational Statistics* 2.1 (2009): 36-43.
20. Rousseeuw, P. J., Van Driessen K. "A Fast Algorithm for the Minimum Covariance Determinant Estimator." *Technometrics* 41.3 (1999): 212-223.
21. Li G, Chen Z. "Projection-Pursuit Approach to Robust Dispersion and Principal Components: Primary Theory and Monte-Carlo." *J. Am. Statist. Assoc.* (1985); 80:759–766.
22. Hubert M, Rousseeuw P.J, Verboven S. "A Fast Method for Robust Principal Components with Applications to Chemometrics." *Chemometrics Intell. Lab. Syst.* (2002); 60:101–111.
23. Rousseeuw, P. J., "Least Median of Squares Regression," *Journal of the American Statistical Association*, (1984); 79, 871-880.
24. Woodruff, D. L., Rocke, D. M., "Heuristic Search Algorithms for the Minimum Volume Ellipsoid," *Journal of Computational and Graphical Statistics*, (1993): 2, 69-95.
25. Butler, R. W., Davies, P. L., Jhun, M., "Asymptotics for the Minimum Covariance Determinant Estimator," *The Annals of Statistics*, (1993): 21, 1385-1400.
26. Davies, L., "The Asymptotics of Rousseeuw's Minimum Volume Ellipsoid Estimator," *The Annals of Statistics*, (1992): 20, 1828-1843.
27. Egan, William J., Morgan S. L. "Outlier Detection in Multivariate Analytical Chemical Data." *Analytical chemistry* 70.11 (1998): 2372-2379.
28. Hawkins, Douglas M. "Identification of Outliers." Vol. 11. London: Chapman and Hall, (1980).
29. Wold, H. "Soft Modelling: The Basic Design and Some Extensions." *Systems Under Indirect Observation, PartII* (1982): 36-37.

30. Wold S., Ruhe A., Wold H., Dunn W., “The Collinearity Problem in Linear Regression. The PLS Approach to Generalized Inverse.” *SIAM J. Sci. Stat. Comput.*, 5(1984), pp. 735-743.
31. Joreskog K., Wold H., eds., “Systems Under Indirect Observation.” Part 1, North Holland, Amsterdam, 1982.
32. Mahalanobis, Chandra P. “On the Generalized Distance in Statistics.” *Proceedings of the National Institute of Science in India* (1936): 2(1):49-55.
33. De Maesschalck R., Jouan-Rimbaud D., Massart, D. L. “The Mahalanobis Distance.” *Chemometrics and Intelligent laboratory Systems* (2000); 50:1-18.
34. Hardin J., Rocke D. “The Distribution of Robust Distances.” <http://www.cipic.ucdavis.edu/~dmrocke/preprints.html> (2002).
35. R Core Team, R: “A Language and Environment for Statistical Computing.” R Foundation for Statistical Computing, Vienna, Austria. (2012), <http://www.R-project.org>.