UNIVERSITI PUTRA MALAYSIA

DEVELOPMENT OF ROBUST PROCEDURES FOR PARTIAL LEAST SQUARE REGRESSION WITH APPLICATION TO NEAR INFRARED SPECTRAL DATA

DIVO DHARMA SILALAHI

IPM 2021 8

DEVELOPMENT OF ROBUST PROCEDURES FOR PARTIAL LEAST SQUARE REGRESSION WITH APPLICATION TO NEAR INFRARED SPECTRAL DATA

By

DIVO DHARMA SILALAHI

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

January 2021

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

DEVELOPMENT OF ROBUST PROCEDURES FOR PARTIAL LEAST SQUARE REGRESSION WITH APPLICATION TO NEAR INFRARED SPECTRAL DATA

By

DIVO DHARMA SILALAHI

January 2021

Chair : **Professor Habshah Binti Midi, PhD**
Institute : **Mathematical Research**

The Partial Least Square Regression (PLSR) is a multivariate method commonly used to build a predictive model of Near Infrared (NIR) spectral data. Based on our experience, several weaknesses of the PLSR have been identified with respect to its robustness issues in the pre-processing and in-processing when outliers and High Leverage Points (HLP) exist in the dataset. In addressing these problems, some robust procedures for PLSR are developed.

In the pre-processing, the pretreatment procedure is needed to remove both additive and multiplicative baseline effects and to distinguish the scattering effect in the raw spectral. The existing methods are not very successful in removing those effects. Hence, a new robust Generalized Multiplicative Scatter Correction (GMSC) algorithm is proposed to correct the additive and/or multiplicative baseline effects during pre-processing spectra. The results indicate that the proposed method outperforms the existing methods in this study.

In the in-processing, the PLSR model is very sensitive to the optimal number of PLS components used in the model fitting process. Several selection procedures of the optimal number of PLS components have been developed in this regard. However, each procedure yields different result. To date, no one has been able to determine the more superior method. Hence, a Robust Reliable Weighted Average (RRWA-PLS) which does not require the selection of an optimal number of PLS is developed by employing the weighted average strategy from multiple PLSR models generated by different complexity of the PLS components. In the PLSR model there is no variable selection procedure that able to remove the irrelevant wavelengths. To fill-in the gap in the

literature, a new robust procedure in wavelength selection based on input scaling method is developed using Filter-Wrapper method. The PLSR fails to discover the nonlinear structure in the original input space. As such, the use of the classical PLSR might not be appropriate. In addition, the contamination of outliers and HLP in the dataset also might damage the whole data processing procedures. To address these problems, robust nonlinear solutions of PLSR are developed through kernel based learning by nonlinearly projecting the original input data matrix to a high dimensional feature mapping corresponding to the kernel space. The nonlinear solutions coupled with some improved robust methods such as Diagnostic Robust Generalized Potential (DRGP) method and GM6-Estimator are also introduced.

Several statistical measures such as Root Mean Squared Error (RMSE), Coefficient of Determination (R^2), Ratio of Performance to Deviation (RPD), and Standard Error (SE) are used to evaluate the superiority of the proposed methods. The results of the simulation study and two NIR spectral data sets, namely the NIR spectral of oil palm (*Elaeis guineensis* Jacq.) fresh and dried ground fruit mesocarp, show that all the proposed methods are superior compared to the existing methods in this study.

Keywords: Near Infrared, Spectral Data, Partial Least Squares, Generalized Multiplicative Scatter Correction, Average-Weighted, Number of Components, Reliability Coefficients, Variable Selection, Variable Importance Projection, Uninformative Variable Eliminations, Nonlinear, Kernel, Hilbert-Space, GM6-Estimator, Diagnostic Robust Generalized Potential.

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

MEMBANGUNKAN PROSEDUR TEGUH KEATAS KAEDAH REGRESI SEPARA KUASA DUA TERKECIL DAN APLIKASI TERHADAP DATA SPEKTRUM INFRA MERAH

Oleh

DIVO DHARMA SILALAHI

Januari 2021

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

Regresi separa kuasa dua terkecil (PLSR) adalah teknik multivariate yang biasa digunapakai untuk membina model ramalan data spektrum infra merah dekat (NIR). Berdasarkan pengalaman kami, beberapa kelemahan PLSR telah dikenal pasti dari segi isu keteguhannya dalam pra-pemprosesan dan semasa pemprosesan apabila wujud titik terpencil dan titik tuasan tinggi (HLP) dalam set data. Bagi menangani masalah ini, beberapa prosedur teguh keatas PLSR telah dibangunkan.

Dalam pra-pemprosesan, prosedur prarawatan diperlukan bagi menghapuskan kedua-dua kesan dasar aditif dan multiplikatif dan bagi membezakan kesan serakan pada spektrum mentah. Kaedah sedia ada tidak berjaya untuk menghapuskan kesan tersebut. Oleh itu, algoritma baru pembetulan serakan multiplikatif teritlak teguh (GMSC) dicadangkan bagi membetulkan kesan garis dasar aditif atau multiplikatif semasa pra-pemprosesan spektra. Keputusan menunjukkan bahawa kaedah yang dicadang mengatasi kaedah sedia ada dalam kajian.

Semasa pemprosesan, model PLSR sangat sensitif terhadap bilangan optimal komponen PLS yang digunakan dalam proses pemodelan. Beberapa prosedur pemilihan bilangan optimal komponen PLS telah dibangunkan. Walau bagaimana pun, setiap prosedur menghasilkan keputusan yang berbeza. Sehingga kini, tiada seorang pun yang berupaya menentukan kaedah yang lebih unggul. Oleh itu, kaedah purata berwajaran teguh yang boleh dipercayai (RRWA-PLS) yang tidak memerlukan pemilihan bilangan optimal PLS dibangunkan dengan menggunakan strategi purata berpemberat dari model PLSR berganda yang dijanakan oleh komponen PLS yang berbeza dan komplek. Sehingga kini, belum ada prosedur pemilihan pembolehubah yang

berupaya untuk menghapuskan gelombang yang tidak relevan dalam model PLSR. Untuk mengisi jurang dalam kesusasteraan, prosedur baru teguh dalam pemilihan gelombang berasaskan kaedah penskalaan input dibangun menggunakan kaedah penapis-pembungkus. PLSR tidak berjaya untuk mengesan struktur tak linear dalam ruang input asal. Oleh itu, penggunaan PLSR klasik berkemungkinan tidak bersesuaian. Tambahan pula, pencemaran dari titik terpecil dan titik tuasan tinggi (HLP) dalam set data boleh mengganggu keseluruhan prosedur pemprosesan data. Untuk menangani masalah ini, beberapa penyelesaian tak linear teguh keatas PLSR dibangunkan melalui pembelajaran berasaskan kernel dengan mengunjurkan secara tak linear input matrik data asal kepada pemetaan cirri dimensi tinggi yang sepadan dengan ruang kernel. Penyelesaian tak linear dengan gabungan beberapa kaedah peningkatan teguh seperti kaedah potensi teritlak teguh berdiagnostik (DRGP) dan penganggar GM6 juga diperkenalkan.

Beberapa ukuran statistik seperti ralat punca min kuasa dua (RMSE), pekali penentuan (R^2), nisbah prestasi kepada penyimpangan (RPD), dan ralat piawai (SE) digunakan untuk menilai keunggulan kaedah yang dicadangkan. Hasil kajian simulasi dan dua set data spectrum NIR sebenar, spektrum NIR dari mesokarp segar dan kering buah kelapa sawit (*Elaeis guineensis* Jacq.) menunjukkan semua kaedah yang dicadangkan lebih unggul berbanding kaedah yang sedia ada dalam kajian ini.

Kata kunci: Infra Merah Dekat, Data Spektrum, Kuasa Dua Terkecil Separa, Pembetulan Serakan Multiplikatif Teritlak, Purata Berpemberat, Bilangan Komponen, Pekali Kebolehpercayaan, Pemilihan Pembolehubah, Unjuran Kepentingan Pembolehubah, Penghapusan Pembolehubah Tak Berinformatif, Tak Linear, Kernel, Ruang Hilbert, Penganggar GM6, Potensi Teritlak Teguh Berdiagnostik.

ACKNOWLEDGEMENTS

I wish to express my deep gratitude and heartfelt appreciation to the people and institutions that became instrumental in the completion of my Doctoral degree in Applied and Computational Statistics.

Grateful acknowledgment is made to my supervisor, Prof. Dr. Habshah Binti Midi, for her intellectual insight, constant encouragement, and hospitality; and the members of the Committee: Dr. Jayanthi A/P Arasan, Dr. Mohd Shafie Bin Mustafa, and Dr. Jean Pierre Caliman for their valuable advice, constructive suggestions, sincere concern and understanding.

I would like to thank Southeast Asian Regional Center for Graduate Study and Research in Agriculture (SEARCA) for funding support in term of scholarship and research grant during my study in Universiti Putra Malaysia. Many thanks also to the former SEARCA Director, Dr. Gil C. Saguiguit Jr. and SEARCA Director, Dr. Glenn B. Gregorio. To Dr. Maria Cristeta N. Cuaresma, Maam Blessie P. Saez, and all SEARCA staff, for their consent and great cooperation to all SEARCA scholars.

I am also grateful to the PT. Sinar Mas Agribusiness Resources and Technology (PT. SMART Tbk) as my home institution for the research activities support, special equipment support, and financial support given during my study. Special thanks are also extended to Bapak Franky Oesman Widjaja and Bapak Daud Dharsono for their approval on my PhD study. To all research staff and operator of SMARTRI for cooperation and outstanding help and support to this research.

To the administrative staff, faculty member, and all colleagues of the Institute for Mathematical Research and School of Graduate School of Universiti Putra Malaysia thank you so much for all your support in helping me to meet all the requirements for the completion of my studies.

I will be forever grateful to my beloved wife, Retno Kusumaningtyas, and my beloved children, Mabini Meisha Rahma Silalahi, Rizaldi Safar Abbas Silalahi, and Maulia Puan Aleesha Silalahi for their love, patient, encouragement and spiritual support. This thesis is also extremely dedicated to my beloved father, (Alm.) Zainal Abidin Silalahi, SH and my beloved mother, Indrawati as the two real heroes in my life.

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

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

Jayanthi A/P Arasan, PhD
Associate Professor
Faculty of Science
Universiti Putra Malaysia
(Member)

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

Jean Pierre Caliman, PhD
Director
SMART Research Institute
Indonesia
(Member)

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

Date: 10 June 2021

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

BLP	Bad Leverage Point
CV	Cross Validation
D	A set of any arbitrary deleted
DRGP	Diagnostic Robust Generalize Potential
GLP	Good Leverage Point
GM	Generalized M-estimator
GMSC	Generalized Multiplicative Spectra Correction
HLP	High Leverage Points
i	Observation/spectrum for the *i* th row
j	Predictor/wavelength for the *j* th column
KPDRGP	Kernel Partial Diagnostic Robust Potential
KPRGM6	Kernel Partial Robust GM6
KPRM	Kernel Partial Robust modified M-estimator
KPRMM	Kernel Partial robust MM-estimator
LMS	Least Median Squares
LS	Least Square
LTS	Least Trimmed Squares
MAD	Median Absolute Deviation
MCD	Minimum Covariance Determinant
MCUVE	Monte Carlo Uninformative Variable Eliminations
MGT	Modified Generalized Studentized Residuals
MSC	Multiplicative Spectral Correction
MVE	Minimum Volume Ellipsoid
NIPALS	Nonlinear Iterative Partial Least Squares

NIRS	Near Infrared Spectroscopy
OPLS	Orthogonal Projections to Latent Structures
PLSR	Partial Least Square Regression
PORIM	Palm Oil Research Institute of Malaysia
PRM	Partial Robust M-Regression
R	A set of remaining
R^2	Coefficient of Determination
RKHS	Reproducing Kernel Hilbert Spaces
RLS	Reweighted Least Squares
RMSE	Root Mean Squared Error
RPD	Ratio of Performance to Deviation
RRWA	Robust Reliable Weighted Average
SE	Standard Error
SNV	Standard Normal Variate
UVE	Uninformative Variable Elimination
VIP	Variable Importance Projection
WA-PLS	Weighted Average PLS
%FFA	Percentage of Fat Fatty Acid
%OWM	Percentage of Oil to Wet Mesocarp
%ODM	Percentage of Oil to Dry Mesocarp
w_i^r	Weight to identify row weight
w_i^x	Weight to identify column weight
VIP_{pred}	VIP value in predictive components
VIP_{ortho}	VIP value in orthogonal components
$1_n, 1_{n_v}$	Vectors whose elements equal to 1, with lengths n and n_v

$\mathrm{SSY}_{comp;\, g}$	Variance of $\mathbf{y}$ explained by the gth PLS component
I	Transmitted intensity
$s(\mathbf{x}_i)$	Standard deviation of spectra for each *i* spectrum
ε	Specific extinction coefficient
$\mathbf{y}$	Single response variable
$\breve{\mathbf{C}}_v$	Shape estimates of the MVE estimators
λ_i	Sequence of eigenvalues
$\hat{\sigma}_{LTS}$	Robust scale estimates using LTS estimator
$\breve{\mathrm{R}}\mathrm{M}_i^2$	Robust mahalanobis distance that employ the *i*th elements $\mathbf{K}$ in the input space
RMD_i	Robust Mahalanobis Distance
$\breve{\mathbf{m}}_v$	Robust location of the MVE estimators
$\mathrm{med}_{L1}(\mathbf{V})$	Robust estimator
H	RKHS space
$\breve{\mathbf{X}}$	Re-weight/new scaled matrix $\mathbf{X}$
$\breve{\mathbf{y}}$	Re-weight vector $\mathbf{y}$
c_{artif}	Reliability of the artificial random noise variables
c_j	Reliability of each variable
$\psi_1(u)$	Re-descending score function
$\mathbf{X}$	Predictor variable
$\hat{\mathbf{y}}$	Prediction of calibration set
$\boldsymbol{\varphi}\boldsymbol{\varphi}^T$	Outer product of the $n \times n$ kernel Gram matrix $\mathbf{K}$
N_t	Number of subsample from training
r	Number of PLS sub-model
m	Number of predictor variable
d	Number of PLS model

n	Number of observation
$\boldsymbol{\varphi}(\cdot)$	Nonlinear mapping function in the input space
$\bar{\mathbf{x}}_i$	Mean of spectra for each each i spectrum
$\mathbf{m}$	Mean of all n spectra
$\breve{\mathbf{V}}$	Matrix of latent variables with kernel Gram matrix $\mathbf{K}$ in the input space
$\mathbf{P}$	Matrix $m \times 1$ consisting loading vector
$\boldsymbol{\varphi}$	Mapping function
MD_i^2	Mahalanobis (squared) Distance
$\theta(u)$	Loss function
$\mathbf{p}$	Loading vector
$\mathbf{q}$	Loading $l \times 1$ vector
K	Kernel function
$\mathbf{u}$	$n \times 1$ matrix of $\mathbf{y}$ block score
$\{\mathbf{x}_i\}_{i=n+1}^{n+n_v}$	Input vectors of calibration set
I_0	Initial incident intensity
$\{\varphi_i\}_{i=1}^{s}$	Infinite sequence of eigenfunctions
w_i	Generalized weight
p_{ii}^*	Generalized potentials
$\Re$	Field
F	Feature space
$\rho(z,c)$	Fair weight function
$\lVert \cdot \rVert$	Euclidean norm
$\mathbf{I}$	n - dimensional identity matrix
$\Delta\lambda$	Difference between the λ values of adjacent data points

Ω	Diagonal matrix with size $m \times m$
w_{ii}	Diagonal elements in matrix $\mathbf{W}$
w_j^{*x}	Column based weight obtained from robust GMSC
b_i	Coefficient parameter of OLS
$\{f_n\}_{n=1}^{\infty}$	Cauchy sequence
λ_j	Band in j wavelength
A	Absorbance of solution
$\langle f,g \rangle$	A complete inner product space
$\mathbf{v}_g$	$n \times 1$ column vector of scores $\mathbf{x}_j$ in $\mathbf{X}$
$\mathbf{w}_j$	$m \times 1$ vector of weight for $\mathbf{X}$
$\mathbf{f}$	$n \times 1$ vector of residual in $\mathbf{y}$
$\mathbf{b}_{inner}$	$l \times 1$ vector of regression coefficient as solution using OLS
$\mathbf{a}$	$l \times 1$ vector coefficient
$\mathbf{g}$	$n \times 1$ matrix of residual in the inner relation
$\boldsymbol{\varphi}$	$n \times s$ matrix of mapped space data
$\mathbf{K}_v$	$n_v \times n$ kernel matrix of validation set
$\mathbf{K}$	$n \times n$ kernel Gram matrix
$\hat{\mathbf{b}}_{PLSR}$	l dimensional vector of regression coefficient
$\{\mathbf{x}_j\}_{j=1}^{n}$	Input vectors of validation set
$\boldsymbol{\gamma}$	Inner relation coefficient of latent regression equation
$\mathbf{V}$	$n \times l$ matrix of the $n \times 1$ vector $\mathbf{v}_g$
$\mathbf{E}$	$n \times m$ matrix of residual in outer relation for predictor $\mathbf{X}$
$\breve{\mathbf{f}}$	$n \times 1$ vector of residual in mixed relation

CHAPTER 1

INTRODUCTION

1.1 Background and Purposes

The Near Infrared Spectroscopy (NIRS) technology has been attracting much attention as secondary analytical tool for chemical analysis of agricultural products. It has been proven that it is rapid, chemical-free, non destructive, reliable, and requires less (even no) sample preparation. It offers the opportunity for the agricultural industry to increase their productivity particularly for quality inspection. In the oil palm (*Elaeis guineensis* Jacq.) industry, this quality inspection is very important which corresponds to the evaluation on the final product of the breeding program and cultivation practice.

The NIRS requires a spectrometer to produce sufficient information called NIR spectrum. It is resulted from the interaction between physical properties of the sample with the optical light of electromagnetic. Practically, the NIR spectral data consist of a large amount of spectrum that leads to a high-dimensional problem. This is due to the situation where a huge number of n observations and wavelength ranges (as m predictors) are employed in the dataset. This high-dimensional may suffers to a potential risk of multicollinearity and heterocedasticity. The NIR spectral are also very often composed of complex overtone, noise, and overlapping peaks with related to the sample condition and instrument performance. These bring the parallel shift, slope and intensity effect and path length difference in the spectra baseline. In addition, the risk of contamination from outliers and High Leverage Points (HLP) in the spectral dataset may decrease the fitted model accuracy. Therefore, a well-assign of robust pretreatment procedure coupled with multivariate analysis is highly suggested.

In multivariate analysis, the Partial Least Square Regression (PLSR) (Wold, 1973) is a statistical standard procedure to build the predictive model of NIR spectral data. It summarizes the variability in both the predictor ($\mathbf{X}$) and response ($\mathbf{y}$) variables into a new smaller set of uncorrelated variables called latent variables or PLS components. The PLSR keeps a maximize covariance of the highly collinear original predictors to create the latent variables and regress these to the dependent variable.

The PLSR has the ability to identify the unwanted samples in the dataset (Xu *et al.*, 2011), to handle the multicollinearity and heterocedasticity effects (Haenlein & Kaplan, 2004), and practically distribution-free assumption (Wold, 1980;

Manne, 1987). It does not matter whatever data distribution is, it is also opposed violations of independence, collinear, and small sample size that are known as major requirement assumptions in classical regression. Aside of its benefits, some studies have reported the weakness due to its robustness issues. The fitted model performs poorly when outliers and HLP exist in a dataset (Kerkri *et al*., 2018). The model is sensitive to the number of PLS components used in the fitting process (Wiklund *et al.*, 2007). Each time the dataset are updated, the re-calculation on the number of components used in the model is required and often yields to different accuracy. There is still no variable selection procedure applied to prevent the irrelevant wavelengths which may impair the model accuracy (Mehmood *et al.*, 2012; Wang *et al.*, 2016). The method fails to discover the nonlinear structure in the original input space, whereby the irregular data space problem still appear in the dataset (Qin & McAvoy, 1992; Rosipal & Trejo, 2001). In addition, the contamination from HLP comprises Good Leverage Point (GLP) and Bad Leverage Point (BLP). The GLP are not significant because they are still near to the fitted regression line, and they can increase the efficiency of an estimate (Midi *et al.*, 2009; Bagheri & Midi, 2015). On the other hand, the BLP are far from the majority pattern of the data; they have significant damage on the computed values of various estimates (Bagheri & Midi, 2015; Alguraibawi *et al.*, 2015). The contamination of outliers and BLP in the dataset should be eliminated during the fitting process. Therefore some improvements on the PLSR method including the robust procedures both in the pre-processing and in-processing spectra are introduced to overcome these problems.

1.2 Importance and Motivation of the Study

In the pre-processing of PLSR, several pretreatment methods are considered such as the Standard Normal Variate (SNV) (Barnes *et al.*, 1989), Multiplicative Spectral Correction (MSC) (Geladi *et al.*, 1985), and the Detrending (Barnes *et al.*, 1989). These methods are often treated in combination to the Derivative method (Owen, 1995) as smoothing procedure. The SNV uses only row-oriented individual spectra transformation through its mean and standard deviation in the standardization. The MSC includes the entire spectra to remove the baseline effect both translation and offset in the spectra. The Detrending applies the subtraction using polynomial fit to remove the baseline shift. In many cases, the Detrending has similarity with the Derivative that could be treated in parallel with the SNV or MSC. It is observed that these pretreatment methods are non-robust since outliers, HLP, and uninformative predictors are not taken care off in the scatter correction process. The weakness of these methods has inspired us to propose a new robust Generalized Multiplicative Spectra Correction (GMSC) method. The proposed method is expected to be able to correct the additive and/or multiplicative baseline effects during pre-processing spectra. This GMSC is based on the row-column weights includes the ability to remove or to reduce the effect of outliers and HLP. Moreover, this method is also able to downgrade the influence of uninformative predictors during the pretreatment.

The PLSR model is sensitive to the number of components used in the fitted model. The Weighted Average (WA) strategy then is suggested as alternative to prevent the sensitivity in the classical PLS. The classical PLS model uses cross-validation approach with one-sigma heuristic (Hastie *et al.*, 2009) to determine the optimum number of components used. The re-calculation on the number of components is done each time the calibration dataset are updated. To encounter this, the WA-PLS method (Hastie *et al.*, 2009) was reviewed. The method uses averaging strategy to incorporate all the possible complexity of the model. In fact some irrelevant variables are still included during the fitting model that may affect model accuracy. The shortcoming of this method has motivated us to develop a Robust Reliable Weighted Average Partial Least Square (RRWA-PLS). The method utilizes the weighted average strategy from multiple PLSR models with different complexity of the PLS components. Two weighting schemes are employed namely the trimmed version (20%) of the standard error prediction (SEP) based on the re-sampling of k-fold Cross Validation (CV) and the reliability values of each predictor variables.

Several variable selection methods attempts to remove irrelevant variables in the PLSR. A proper selection method is crucial to prevent the PLSR from processing certain number of irrelevant wavelengths during model fitting process. Some existing variable selection methods such as filter and wrapper methods are used as the scaling matrix for input variable $\mathbf{X}$. The classical Variable Importance Projection (VIP) (Wold *et al.*, 1993) is a famous filter method that uses the weighted combination over all component variables of the squared PLSR. However, it does not include the projection to the orthogonal components (Galindo-Prieto *et al.*, 2014). In the wrapper method, the Monte Carlo Uninformative Variable Elimination (MCUVE) (Cai *et al.*, 2008) constructs the reliability of each wavelength through the fraction between the mean and standard deviation of PLS regression coefficient. It is suspected that these methods are easily affected by outliers. Their works have motivated us to propose an improvised input scaling method which is based on the Filter-Wrapper method. The method combines the superiority of modified VIP using Orthogonal Projections to Latent Structures (OPLS) (Galindo-Prieto *et al.*, 2014) and the Monte Carlo Uninformative Variable Eliminations (MCUVE) to scale the wavelength variable as input factor for PLSR. Moreover, a new robust reliability coefficient and new robust cut-off criterion are introduced in the procedure.

The collected spectra are very often composes of complex overtone and many overlapping peaks which may lead to misinterpretation because of its significant nonlinear characteristics. Using linear solution might not be appropriate. Moreover, with the high-dimension of dataset due to large number of observations and data points will impact to the multicollinearity problem. This also will increase the risk of contamination from multiple outliers and HLP. The multiple linear regression methods are considered not suitable to fit these problems. In order to deal with these irregular data space problem, this has encouraged us to apply the nonlinear solution for PLSR. The solution deploys the kernel based learning by nonlinearly projecting the original input data matrix to a high dimensional feature space corresponding to a Reproducing Kernel Hilbert Spaces (RKHS) (Aronszajn, 1950). Here, the performance of existing non-kernel of classical PLSR and robust PLSR using modified M-estimator (Serneel *et al.*, 2005), MM-estimator, modified GM6-estimator, and Diagnostic Robust Generalize Potential (DRGP) are compared with their kernel version.

Aside on handling irregular data space problem using kernel solution of RKHS, the elimination on outliers and BLP is required to prevent a serious damage on the parameter estimate. In relation to this, there are many studies (see Atkinson, 1994; Imon, 2002; Seneels *et al.*, 2015; Jia *et al.*, 2010) have been conducted to identify the outliers and HLP, but none of them has ability to classify the HLP into good or bad. This has encouraged us to introduce a new method by considering only the outliers and BLP in the elimination. The improvement on bounded influence and high breakdown-point (with close to 50%) robust procedure of GM6-estimator (Coakley & Hettmanspreger, 1993) are introduced. The proposed method accommodates several robust approaches on the initial weight in GM6-estimator to remove both outliers and BLP in the dataset.

The desirability indices (Trautmann, 2004) using several statistical measures are presented to evaluate the superiority of the proposed methods. The measure involves the Root Mean Squared Error (RMSE), Coefficient of Determination (R^2), Ratio of Performance to Deviation (RPD), and Standard Error (SE) based on the differentiation of the actual values against their prediction values. The existing and proposed methods with related to the study are reviewed clearly in each chapter together with their application using artificial data and NIR spectral dataset. Monte Carlo simulation is then utilized in the artificial data to evaluate the stability of the proposed methods. This study provides a development and important contribution to tackle the challenges of scientific big data particularly for process control in the vibrational spectroscopy technique.

1.3 Objective of Thesis

In summary, the foremost objective of our study can be outlined with the following objectives:

1. Developing a new robust pretreatment method that is resistant to outliers and HLP that able to downgrade the influence of uninformative predictors in the NIR spectral data.
2. Establishing a new PLSR model based on modified weighted average strategy with less sensitivity to the optimum number of PLS components used in the model fitting.
3. Improving the wavelength selection method in the PLSR model using input scaling strategy based on the reliability coefficient of filter-wrapper method.

4. Formulating a robust nonlinear solution to the PLSR method using kernel based learning of RKHS in handling the irregular data space in the input data matrix.
5. Improving the robust solution to the nonlinear PLSR method with high resistant to the outliers and BLP in the dataset.

1.4 Scope and Limitation of Study

The PLSR is the commonly used algorithm to solve a partial least squares regression problem for high dimensional data, when the number of predictors (m) is larger than the sample size (n). It can also be used for big and low dimensional data where the number of predictors (m) is smaller than the sample size (n). In this thesis, the focus of the study is for big data when large data points are considered whereby the number of predictors and sample size are very huge. The NIR spectral data are categorized as big data because the dataset has a large number of n observations and m predictors. The PLSR method is the common solution to reduce the dimension into smaller new latent variables called components scores.

The development of PLSR model on NIR spectral data is still limited particularly in the area of robust statistics. This is probably due to the high cost of the NIRS instrument that hinders the development of such methods. Consequently, the limitations of the PLSR are not getting much attention. The existing software still employs the classical methods for analyzing the pre-processing and in-processing NIR spectral data. These have motivated us to develop new algorithm with main objectives are to minimize all the complexities found in the classical PLSR model.

We focus our study in the pre-processing and in-processing the NIR spectral data, since these process are the most important to the outcome of the post-processing. Post-processing is related to the use of fitted PLSR model to the routine laboratory analysis and quality control procedures.

In this thesis, the NIR spectral data are just an application of the proposed techniques. The techniques are applicable in the situation where number of observations is greater than the number of predictors.

1.5 Outline of the Thesis

In accordance to the objective and scopes of study, the contents of this thesis are organized into seven chapters. The thesis chapters are structured so that the objectives of the study are apparent in the sequence outline.

Chapter Two: This chapter discusses about the literature reviews of the PLSR model, pretreatment of NIR spectral data, and in-processing such as weighted average PLSR model, variable selection, and kernel based learning of RKHS. The important of existing robust methods for estimation parameter in the presence of outliers and HLP are also reviewed. The desirability index as statistical measures used to evaluate the superiority of the methods and the fundamental of NIRS spectral data is also discussed in the rest of this chapter.

Chapter Three: This chapter evaluates the performance of the existing pretreatment methods: SNV, Detrend, SNV with Detrend, MSC, and MSC with Detrend. Our developed method called Generalized Multiplicative Spectra Correction (GMSC) is discussed in detail. The superiority of the proposed GMSC is also evaluated by combining the method with the Detrend and Derivative algorithm.

Chapter Four: This chapter evaluates the performance of our proposed Robust Reliable Weighted Average PLS (RRWA-PLS) with the classical WA-PLS and the improvised weight of classical WA-PLS which is called as MWA-PLS.

Chapter Five: This chapter discusses about our new procedure of wavelength selection in the PLSR model called modified VIP-MCUVE (mod-VIP-MCUVE). The method uses input scaling strategy based on reliability coefficient of filter-wrapper method. The existing of classical VIP and MCUVE method and the auto scaling in classical PLSR are also included in the evaluation.

Chapter Six: This chapter deals with the development of robust PLSR which is based on the improvised MM-estimator, improvised GM6-estimator, and Diagnostic Robust Generalize Potential (DRGP). These methods are compared with the classical PLSR and the existing improvised M-estimator. The existing kernel versions on classical PLSR called Kernel PLSR (KPLS) and improvised M-estimator (KPRM) are also reviewed with the kernel version on DRGP (KPDRGP). These kernel versions are used to evaluate their performance in handling the irregular data space that may happen in the NIR spectral dataset.

Chapter Seven: In this chapter, the improvement on the robust procedure of kernel solution in the Chapter Six is extended by removing only the outliers and BLP in the dataset. The proposed methods called as Kernel Partial Robust GM6-estimator (KPRGM6) and Kernel Partial Robust Modified GM6-estimator (KPRMGM6) are presented. The superiority of the proposed robust methods is compared with the non-robust KPLS.

Chapter Eight: This chapter provides the general conclusions of the studies and the recommendations for future research.

Bennett, K. P., & Embrechts, M. J. (2003). An optimization perspective on kernel partial least squares regression. *Nato Science Series sub series III computer and systems sciences*, *190*, 227-250.

Berg, R. A. (2006). Hoefsloot HCJ van den, Westerhuis JA, Smilde AK, Werf MJ, van der: Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics*, *7*, 142.

Blanco, M., & Villarroya, I. N. I. R. (2002). NIR spectroscopy: a rapid-response analytical tool. *TrAC Trends in Analytical Chemistry*, *21*(4), 240-250.

Butkutë, B. (2005). Effect of the calibration model on the correlation between spectral data and nitrogen content in various agricultural objects. *Chemija*, *16*(1).

Cai, W., Li, Y., & Shao, X. (2008). A variable selection method based on uninformative variable elimination for multivariate calibration of near-infrared spectra. *Chemometrics and intelligent laboratory systems*, *90*(2), 188-194.

Carroll, R. J., & Welsh, A. H. (1988). A note on asymmetry and robustness in linear regression. *The American Statistician*, *42*(4), 285-287.

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.

Centner, V., Verdú-Andrés, J., Walczak, B., Jouan-Rimbaud, D., Despagne, F., Pasti, L. & De Noord, O. E. (2000). Comparison of multivariate calibration techniques applied to experimental NIR data sets. *Applied spectroscopy*, *54*(4), 608-623.

Chong, I. G., & Jun, C. H. (2005). Performance of some variable selection methods when multicollinearity is present. *Chemometrics and intelligent laboratory systems*, *78*(1-2), 103-112.

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.

Cozzolino, D., Kwiatkowski, M. J., Dambergs, R. G., Cynkar, W. U., Janik, L. J., Skouroumounis, G., & Gishen, M. (2008). Analysis of elements in wine using near infrared spectroscopy and partial least squares regression. *Talanta*, *74*(4), 711-716

Cui, C., & Fearn, T. (2017). Comparison of partial least squares regression, least squares support vector machines, and Gaussian process regression for a near infrared calibration. *Journal of Near Infrared Spectroscopy*, *25*(1), 5-14.

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.

Dardenne, P., Sinnaeve, G., & Baeten, V. (2000). Multivariate calibration and chemometrics for near infrared spectroscopy: which method?. *Journal of Near Infrared Spectroscopy*, *8*(4), 229-237.

De Haan, J., & Sturm, J. E. (2000). No Need to Run Millions of Regressions. *Available at SSRN 246453, 1-12*.

De Jong, S. (1993). SIMPLS: an alternative approach to partial least squares regression. *Chemometrics and intelligent laboratory systems*, *18*(3), 251-263.

Du, Z., & Wiens, D. P. (2000). Jackknifing, weighting, diagnostics and variance estimation in generalized M-estimation.*Statistics & probability letters*, *46*(3), 287-299.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *Annal. Stat.* ***7****(1),* 1–26.

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: A Journal of the Chemometrics Society*, *13*(2), 165-184.

Fushiki, T. (2011). Estimation of prediction error by using K-fold cross-validation. *Statistics and Computing*, *21*(2), 137-146.

Galindo-Prieto, B., Eriksson, L., & Trygg, J. (2014). Variable influence on projection (VIP) for orthogonal projections to latent structures (OPLS). *Journal of Chemometrics*, *28*(8), 623-632.

Garcia, H., & Filzmoser, P. (2011). Multivariate Statistical Analysis using the R package chemometrics. *Vienna: Austria*.

Garthwaite, P. H. (1994). An interpretation of partial least squares. *Journal of the American Statistical Association*, *89*(425), 122-127.

Geladi, P., MacDougall, D., & Martens, H. (1985). Linearization and scatter-correction for near-infrared reflectance spectra of meat. Applied spectroscopy, 39(3), 491-500.

Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. *Statistics in medicine*, *27*(15), 2865-2873.

Gholizadeh, A., Borůvka, L., Saberioon, M. M., Kozák, J., Vašát, R., & Němeček, K. (2015). Comparing different data preprocessing methods for monitoring soil heavy metals based on soil spectral features. *Soil and Water Research*, *10*(4), 218-227.

Gidskehaug, L., Anderssen, E., Flatberg, A., & Alsberg, B. K. (2007). A framework for significance analysis of gene expression data using dimension reduction methods. *BMC bioinformatics*, *8*(1), 346.

Gómez-Carracedo, M. P., Andrade, J. M., Rutledge, D. N., & Faber, N. M. (2007). Selecting the optimum number of partial least squares components for the calibration of attenuated total reflectance-mid-infrared spectra of undesigned kerosene samples. *Analytica chimica acta*, *585*(2), 253-265.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, *3*(Mar), 1157-1182.

Guzmán, E., Baeten, V., Pierna, J. A. F., & García-Mesa, J. A. (2011). Application of low-resolution Raman spectroscopy for the analysis of oxidized olive oil. *Food Control*, *22*(12), 2036-2040.

Habshah, M., Norazan, M. R., & Rahmatullah Imon, A. H. M. (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.

Haenlein, M., & Kaplan, A. M. (2004). A beginner's guide to partial least squares analysis. *Understanding statistics*, *3*(4), 283-297.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). Linear models: robust estimation. *Robust Statistics: The Approach Based on Influence Functions*, 307-341.

Han, Q. J., Wu, H. L., Cai, C. B., Xu, L., & Yu, R. Q. (2008). An ensemble of Monte Carlo uninformative variable elimination for wavelength selection. *Analytica chimica acta*, *612*(2), 121-125.

Handschin, E., Schweppe, F. C., Kohlas, J., & Fiechter, A. A. F. A. (1975). Bad data analysis for power system state estimation.*IEEE Transactions on Power Apparatus and Systems*, *94*(2), 329-337.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction, *Springer Series in Statistics*.

He, Y., Zhang, Y., Pereira, A. G., Gómez, A. H., & Wang, J. (2005). Nondestructive determination of tomato fruit quality characteristics using VIS/NIR spectroscopy technique.*International Journal of Information Technology*, *11*(11), 97-108.

Hira, Z. M., & Gillies, D. F. (2015). A review of feature selection and feature extraction methods applied on microarray data. *Advances in bioinformatics*, *2015*.

Hourant, P., Baeten, V., Morales, M. T., Meurens, M., & Aparicio, R. (2000). Oil and fat classification by selected bands of near-infrared spectroscopy. *Applied spectroscopy*, *54*(8), 1168-1174.

Huber, P. J. (1973). Robust regression: asymptotics, conjectures and Monte Carlo. *Annals of statistics*, *1*(5), 799-821.

Huber, P. J. (2011). Robust statistics. In International Encyclopedia of Statistical Science (pp. 1248-1251). Springer Berlin Heidelberg.

Hubert, M., & Branden, K. V. (2003). Robust methods for partial least squares regression. *Journal of Chemometrics: A Journal of the Chemometrics Society*, *17*(10), 537-549.

Ilari, J. L., Martens, H., & Isaksson, T. (1988). Determination of particle size in powders by scatter correction in diffuse near-infrared reflectance. *Applied Spectroscopy*, *42*(5), 722-728.

Imon, A. H. M. R. (2002). Identifying multiple high leverage points in linear regression. *Journal of Statistical Studies*, *3*, 207-218.

Imon, A. H. M. R. (2005). Identifying multiple influential observations in linear regression. *Journal of Applied statistics*, *32*(9), 929-946.

Imon, A. H. M. R., & Khan, M. A. I. (2003). A Solution to the problem of multicollinearity caused by the presence of multiple high leverage points. *Int. J. Stat. Sci*, *2*, 37-50.

Jia, R. D., Mao, Z. Z., Chang, Y. Q., & Zhang, S. N. (2010). Kernel partial robust M-regression as a flexible robust nonlinear modeling technique. *Chemometrics and Intelligent Laboratory Systems*, *100*(2), 91-98.

Kasemsumran, S., Thanapase, W., Punsuvon, V., & Ozaki, Y. (2012). A feasibility study on non-destructive determination of oil content in palm fruits by visible–near infrared spectroscopy. *Journal of Near Infrared Spectroscopy*, *20*(6), 687-694.

Kerkri, A., Allal, J., & Zarrouk, Z. (2018). Robust Nonlinear Partial Least Squares Regression Using the BACON Algorithm. *Journal of Applied Mathematics*, *2018*.

Kim, J., Kiss, B., & Lee, D. (2016). An adaptive unscented Kalman filtering approach using selective scaling. In *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 000784-000789). IEEE.

Kim, S., Okajima, R., Kano, M., & Hasebe, S. (2013). Development of soft-sensor using locally weighted PLS with adaptive similarity measure. *Chemometrics and Intelligent Laboratory Systems*, *124*, 43-49.

Kokaly, R. F., & Clark, R. N. (1999). Spectroscopic determination of leaf biochemistry using band-depth analysis of absorption features and stepwise multiple linear regression. *Remote sensing of environment*, *67*(3), 267-287.

Kvalheim, O. M., Arneberg, R., Grung, B., & Rajalahti, T. (2018). Determination of optimum number of components in partial least squares regression from distributions of the root-mean-squared error obtained by Monte Carlo resampling. *Journal of Chemometrics*, *32*(4), e2993.

Leardi, R., & Gonzalez, A. L. (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.

Lee, C., Polari, J. J., Kramer, K. E., & Wang, S. C. (2018). Near-Infrared (NIR) Spectrometry as a Fast and Reliable Tool for Fat and Moisture Analyses in Olives. *ACS omega*, *3*(11), 16081-16088.

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

Levasseur-Garcia, C. (2018). Updated overview of infrared spectroscopy methods for detecting mycotoxins on cereals (corn, wheat, and barley). *Toxins*, *10*(1), 38.

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.

Liu, Y., & Zhou, Y. (2013). Quantification of the Soluble Solids Content of Intact Apples by Vis–NIR Transmittance Spectroscopy and the LS-SVM Method. *Spectroscopy, (28), 32-43.*

Liu, Y., Sun, X., Zhou, J., Zhang, H., & Yang, C. (2010). Linear and nonlinear multivariate regressions for determination sugar content of intact Gannan navel orange by Vis–NIR diffuse reflectance spectroscopy. *Mathematical and Computer Modelling*, *51*(11), 1438-1443.

Lodhi, H., & Yamanishi, Y. (Eds.). (2010). *Chemoinformatics and Advanced Machine Learning Perspectives: Complex Computational Methods and Collaborative Techniques: Complex Computational Methods and Collaborative Techniques*. IGI Global.

Ma, X., Zhang, Y., Cao, H., Zhang, S., & Zhou, Y. (2018). Nonlinear regression with high-dimensional space mapping for blood component spectral quantitative analysis. *Journal of Spectroscopy*, *2018*.

Mallows, C.L. (1975). On some topics in robustness. Technical memorandum, Bell Telephone Laboratories, Murray Hill, N.J.

Manne, R. (1987). Analysis of two partial-least-squares algorithms for multivariate calibration. *Chemometrics and Intelligent Laboratory Systems*, *2*(1-3), 187-197.

Mark, H. (1989). Chemometrics in near-infrared spectroscopy. *Analytica chimica acta*, *223*, 75-93.

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

Martens, H., & Naes, T. (1992). *Multivariate calibration*. John Wiley & Sons.

McLeod, G., Clelland, K., Tapp, H., Kemsley, E. K., Wilson, R. H., Poulter, G., Coombs, D & Hewitt, C. J. (2009). A comparison of variate pre-selection methods for use in partial least squares regression: A case study on NIR spectroscopy applied to monitoring beer fermentation. *Journal of food engineering*, *90*(2), 300-307.

Mehmood, T., Martens, H., Sæbø, S., Warringer, J., & Snipen, L. (2011). A Partial Least Squares based algorithm for parsimonious variable selection. *Algorithms for Molecular Biology*, *6*(1), 27.

Midi, H. (1998). Robust Estimation of a Linearized Nonlinear Regression Model with Heteroscedastic Errors: A Simulation Study. *Pertanika Journal Science & Technology, 6(1)*, 23-35.

Midi, H., Hendi, H. T., Arasan, J., & Uraibi, H. (2020). Fast and Robust Diagnostic Technique for the Detection of High Leverage Points. *Pertanika Journal of Science & Technology*,*28*(4).

Midi, H., Norazan, M. R., & Rahmatullah Imon, A. H. M. (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.

Mika, S., Rätsch, G., & Müller, K. R. (2001). A mathematical programming approach to the kernel fisher algorithm. *Advances in neural information processing systems*, 591-597.

Naes, T., Isaksson, T., & Kowalski, B. (1990). Locally weighted regression and scatter correction for near-infrared reflectance data. *Analytical Chemistry*, *62*(7), 664-673.

Natrella, M. G. (1963). *Experimental Statistics Handbook 91*. US Government Printing Office.

Norris, K. H., & Williams, P. C. (1984). Optimization of Mathematical Treatments of Raw Near-Infrared Signal in the.*Cereal Chem*, *61*(2), 158-165.

Orso, A., Shi, N., & Harrold, M. J. (2004). Scaling regression testing to large software systems. *ACM SIGSOFT Software Engineering Notes*, *29*(6), 241-251.

Oussama, A., Elabadi, F., Platikanov, S., Kzaiber, F., & Tauler, R. (2012). Detection of olive oil adulteration using FT-IR spectroscopy and PLS with variable importance of projection (VIP) scores. *Journal of the American Oil Chemists' Society*, *89*(10), 1807-1812.

Owen, A. J. (1995). Uses of derivative spectroscopy. Agilent Technologies, 8.

Palermo, G., Piraino, P., & Zucht, H. D. (2009). Performance of PLS regression coefficients in selecting variables for each response of a multivariate PLS for omics-type data. *Advances and applications in bioinformatics and chemistry: AABC*, *2*, 57.

Paz, P., Sánchez, M. T., Pérez-Marín, D., Guerrero, J. E., & Garrido-Varo, A. (2008). Nondestructive determination of total soluble solid content and firmness in plums using near-infrared reflectance spectroscopy. *Journal of agricultural and food chemistry*, *56*(8), 2565-2570.

Preda, C. (2007). Regression models for functional data by reproducing kernel Hilbert spaces methods. *Journal of statistical planning and inference*, *137*(3), 829-840.

Qin, S. J., & McAvoy, T. J. (1992). Nonlinear PLS modeling using neural networks. *Computers & Chemical Engineering*, *16*(4), 379-391.

Rännar, S., Lindgren, F., Geladi, P., & Wold, S. (1994). A PLS kernel algorithm for data sets with many variables and fewer objects. Part 1: Theory and algorithm. *Journal of Chemometrics*, *8*(2), 111-125.

Rao, V., Soh, A.C., Corley, R.H.V., Lee, C.H., Rajanaidu, N. (1983). Critical Reexamination of the Method of Bunch Quality Analysis in Oil Palm Breeding. *PORIM Occasional Paper*. 1983. Available online: https://agris.fao.org/agris-search/search.do?recordID=US201302543052 (accessed on 13 October 2020).

Reunanen, J. (2003). Overfitting in making comparisons between variable selection methods. *Journal of Machine Learning Research*, *3*(Mar), 1371-1382.

Rinnan, Å., van den Berg, F., & Engelsen, S. B. (2009). Review of the most common pre-processing techniques for near-infrared spectra. *TrAC Trends in Analytical Chemistry*, *28*(10), 1201-1222.

Rodriguez-Saona, L. E., Fry, F. S., McLaughlin, M. A., & Calvey, E. M. (2001). Rapid analysis of sugars in fruit juices by FT-NIR spectroscopy. *Carbohydrate Research*, *336*(1), 63-74.

Roggo, Y., Chalus, P., Maurer, L., Lema-Martinez, C., Edmond, A., & Jent, N. (2007). A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies. *Journal of pharmaceutical and biomedical analysis*, *44*(3), 683-700.

Rosipal, R. (2011). Nonlinear partial least squares an overview. *Chemoinformatics and advanced machine learning perspectives: complex computational methods and collaborative techniques*, 169-189.

Rosipal, R., & Trejo, L. J. (2001). Kernel partial least squares regression in reproducing kernel hilbert space. *Journal of machine learning research*, *2*(Dec), 97-123.

Rousseeuw, P. J. (1983). Regression techniques with high breakdown point. *The Institute of Mathematical Statistics Bulletin*, *12*, 155.

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

Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. *Mathematical statistics and applications*,*8*(283-297), 37.

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

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

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection* (Vol. 589). John wiley & sons.

Rousseeuw, P. J., & Leroy, A. M. (2005). Robust regression and outlier detection (Vol. 589). John wiley & sons.

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

Saccenti, E., Westerhuis, J. A., Smilde, A. K., van der Werf, M. J., Hageman, J. A., & Hendriks, M. M. (2011). Simplivariate models:

uncovering the underlying biology in functional genomics data. *PloS one*, *6*(6), e20747.

Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *bioinformatics*, *23*(19), 2507-2517.

Sánchez, M. T., De la Haba, M. J., Benítez-López, M., Fernández-Novales, J., Garrido-Varo, A., & Pérez-Marín, D. (2012). Non-destructive characterization and quality control of intact strawberries based on NIR spectral data. *Journal of Food engineering*, *110*(1), 102-108.

Schowengerdt, R. A. (1997). Remote Sensing Models and Methods for Image Processing, Academic Press. *San Diego*.

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.

Shao, Y., He, Y., Bao, Y., & Mao, J. (2009). Near-infrared spectroscopy for classification of oranges and prediction of the sugar content. *International Journal of Food Properties*, *12*(3), 644-658.

Shenk, J. S., Westerhaus, M. O., & Berzaghi, P. (1997). Investigation of a LOCAL calibration procedure for near infrared instruments. *Journal of Near Infrared Spectroscopy*, *5*(4), 223-232.

Siegel, A. F. (1982). Robust regression using repeated medians. *Biometrika*, *69*(1), 242-244.

Siew, W. L., Tan, Y. A., & Tang, T. S. (1995). Methods of Test for Palm Oil and Palm Oil Products: Compiled by Siew Wai Lin, Tang Thin Sue, Tan Yew Ai. Palm Oil Research Institute of Malaysia.

Sindhwani, V., Minh, H. W., & Lozano, A. C. (2012). Scalable matrix-valued kernel learning and high-dimensional nonlinear causal inference. *stat*, *1050*, 17.

Sindhwani, V., Quang, M. H., & Lozano, A. C. (2013). Scalable matrix-valued kernel learning for high-dimensional nonlinear multivariate regression and granger causality. *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence,* 586-595.

Song, W., Wang, H., Maguire, P., & Nibouche, O. (2017). Local Partial Least Square classifier in high dimensionality classification. *Neurocomputing*, *234*, 126-136.

Stuart, B. (2004). Infrared Spectroscopy: Fundamentals and Applications. John Wiley & Sons. *Inc., USA*, pp.167–185.

Sudarno, Silalahi, D. D., Risman, T., Widyastuti, B. L., Davrieux, F., Yuan, Y. Y., & Caliman, J. P. (2017). Rapid determination of oil content in dried-ground oil palm mesocarp and kernel using near infrared spectroscopy. *Journal of Near Infrared Spectroscopy*, *25*(5), 338-347.

Tran, T., Szymańska, E., Gerretzen, J., Buydens, L., Afanador, N. L., & Blanchet, L. (2017). Weight randomization test for the selection of the number of components in PLS models. *Journal of Chemometrics*, *31*(5), e2887.

Trautmann, H. (2004). *The desirability index as an instrument for multivariate process control* (No. 2004, 43). Technical Report/Universität Dortmund, SFB 475 Komplexitätsreduktion in Multivariaten Datenstrukturen.

Trygg, J., & Wold, S. (2002). Orthogonal projections to latent structures (O-PLS). *Journal of Chemometrics: A Journal of the Chemometrics Society*, *16*(3), 119-128.

van den Berg, R. A., Hoefsloot, H. C., Westerhuis, J. A., Smilde, A. K., & van der Werf, M. J. (2006). Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC genomics*, *7*(1), 142.

van der Voet, H. (1994). Comparing the predictive accuracy of models using a simple randomization test. *Chemometrics and intelligent laboratory systems*, *25*(2), 313-323.

Vandeginste, B. G., Massart, D. L., Buydens, L. M., Lewi, P. J., Smeyers-Verbeke, J., & Jong, S. D. (1998). Handbook of chemometrics and qualimetrics. *Elsevier Science Inc*.

Viscarra Rossel, R. A., & Lark, R. M. (2009). Improved analysis and modeling of soil diffuse reflectance spectra using wavelets. *European Journal of Soil Science*, *60*(3), 453-464.

Wand, M. P., & Jones, M. C. (1994). *Kernel smoothing*. Crc Press.

Wang, L., Lee, F. S., Wang, X., & He, Y. (2006). Feasibility study of quantifying and discriminating soybean oil adulteration in camellia oils by attenuated total reflectance MIR and fiber optic diffuse reflectance NIR. *Food chemistry*, *95*(3), 529-536.

Wang, S., Tang, J., Liu, H. (2016). Feature selection. In Encyclopedia of Machine Learning and Data Mining. *Springer Science + Business Media*: New York, NY, USA, pp. 1–9.

Wang, X., & Zhou, G. (2010). Study on Pretreatment Algorithm of Near Infrared Spectroscopy. In *International Conference on Computer and Computing Technologies in Agriculture* (pp. 623-632). Springer, Berlin, Heidelberg.

Wang, Z. X., He, Q. P., & Wang, J. (2015). Comparison of variable selection methods for PLS-based soft sensor modeling. *Journal of Process Control*, *26*, 56-72.

Wiklund, S., Nilsson, D., Eriksson, L., Sjöström, M., Wold, S., & Faber, K. (2007). A randomization test for PLS component selection. *Journal of Chemometrics: A Journal of the Chemometrics Society*, *21*(10-11), 427-439.

Wold, H. (1973). In: P.R. Krishnaiah (Ed.), Multivariate Analysis, vol. III, Academic Press: New York. pp. 383– 407.

Wold, H. (1973). Nonlinear iterative partial least squares (NIPALS) modelling: some current developments. In *Multivariate analysis–III* (pp. 383-407). Academic Press.

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.

Wold, H. (1980). Model construction and evaluation when theoretical knowledge is scarce: Theory and application of partial least squares. In *Evaluation of econometric models* (pp. 47-74).

Wold, S., Antti, H., Lindgren, F., & Öhman, J. (1998). Orthogonal signal correction of near-infrared spectra.*Chemometrics and Intelligent laboratory systems*, *44*(1), 175-185.

Wold, S., Johansson, E., Cocchi, M. (1993). PLS - Partial Least-Squares Projections to Latent Structures. In 3D QSAR in Drug Design. Theory, Methods and Applications. Kubinyi, H., (Eds). *ESCOM Science Publishers. B. V.: Leiden*, 523-550.

Wold, S., Sjöström, M., Eriksson, L. (2001). "PLS-regression: a basic tool of chemometrics". Chemometrics and Intelligent Laboratory Systems. 58 (2): 109–130.

Wu, W., Walczak, B., Massart, D. L., Prebble, K. A., & Last, I. R. (1995). Spectral transformation and wavelength selection in near-infrared spectra classification. *Analytica Chimica Acta*, *315*(3), 243-255.

Xu, L., Cai, C. B., & Deng, D. H. (2011). Multivariate quality control solved by one-class partial least squares regression: identification of adulterated peanut oils by mid-infrared spectroscopy. *Journal of Chemometrics*, *25*(10), 568-574.

Yang, H., Griffiths, P. R., & Tate, J. D. (2003). Comparison of partial least squares regression and multi-layer neural networks for quantification of nonlinear systems and application to gas phase Fourier transform infrared spectra. *Analytica Chimica Acta*, *489*(2), 125-136.

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

Yu, C., & Yao, W. (2017). Robust linear regression: A review and comparison. *Communications in Statistics-Simulation and Computation*, *46*(8), 6261-6282.

Zhang, M. H., Xu, Q. S., & Massart, D. L. (2004). Averaged and weighted average partial least squares. *Analytica chimica acta*, *504*(2), 279-289.

Zhang, T. (2001). An introduction to support vector machines and other kernel-based learning methods. *AI Magazine*, *22*(2), 103.

BIODATA OF STUDENT

Divo Dharma Silalahi was born on 24th of November 1988 in Pangkalpinang, Bangka-Belitung Province, Indonesia. He is a Senior Researcher at SMART Research Institute (SMARTRI), the Research and Development Division of PT. Sinarmas Agribusiness Resources Technology (PT.SMART.Tbk) which is one of the largest, publicly-listed and integrated palm-based consumer companies in Indonesia. His specializations are robust statistics, chemometrics, near infrared spectroscopy, hyperspectral imaging, computer vision, and database programming. He has been starting in implementing the NIRS technology for research and development in palm oil industry since 2011. He holds a Bachelor degree in Statistics from Diponegoro University, Indonesia in 2009, and then in 2015 He earned Master's degree with Major in Statistics and Minor in Mathematics from University of the Philippines Los Banos, Philippines. Right after, in 2016 He was admitted in the Universiti Putra Malaysia under the PhD in Applied and Computational Statistics program with research focuses on robust statistics. He received numbers of awards such national young scientist award, best paper during international conference and some prestigious scholarship from PT. SMART TBK for his Masteral degree and from Southeast Asian Regional Center for Graduate Study and Research in Agriculture (SEARCA) for his Doctoral degree. He also has published number of paper in some high quality Journal (Indexed in Scopus and Web of Science) during his career and study.

LIST OF PUBLICATIONS

Journals

Silalahi, D. D., Midi, H., Arasan, J., Mustafa, M. S., & Caliman, J. P. (2018). Robust generalized multiplicative scatter correction algorithm on pretreatment of near infrared spectral data. *Vibrational Spectroscopy*, *97*, 55-65. (Published on 12 May 2018). Indexed in JCR – Q2

Silalahi, D. D., Midi, H., Arasan, J., Mustafa, M. S., & Caliman, J. P. Automated Fitting Process Using Robust Reliable Weighted Average on Near Infrared Spectral Data Analysis. *Symmetry*, *12*(12), 2099. (Published on 17 December 2020). Indexed in JCR – Q2

Silalahi, D. D., Midi, H., Arasan, J., Mustafa, M. S., & Caliman, J. P. (2020). Robust Wavelength Selection Using Filter-Wrapper Method and Input Scaling on Near Infrared Spectral Data. *Sensors*, *20*(17), 5001. (Published on 3 September 2020). Indexed in JCR– Q1

Silalahi, D. D., Midi, H., Arasan, J., Mustafa, M. S., & Caliman, J. P. (2020). Kernel partial diagnostic robust potential to handle high-dimensional and irregular data space on near infrared spectral data. *Heliyon*, *6*(1), e03176. (Published on 2 January 2020). Indexed in Scopus

Silalahi, D. D., Midi, H., Arasan, J., Mustafa, M. S., & Caliman, J. P. (2021). Kernel partial robust modified GM6 with high resistance to multiple outliers and bad leverage points. *Symmetry*, *13*(4), 547. (Published on 26 March 2021). Indexed in JCR – Q2

Silalahi, D. D., Midi, H., Arasan, J., Mustafa, M. S., & Caliman, J. P. (2020). Nonlinear partial robust MM and GM6 regression using kernel space on near infrared spectral data. *Thai Journal of Mathematics*. (Accepted for publication on 9 January 2020). Indexed in Scopus

Proceedings

Silalahi, D. D., & Midi, H. (2017, January). Considering a non-polynomial basis for local kernel regression problem. *In AIP Conference Proceedings.* AIP Publishing LLC. *1795, pp. 020024.*

LIST OF PRESENTATIONS

Silalahi, D. D., & Midi, H. (2016, November). A Promising Reliable, Precise and Robust Non Destructive Assessment Technique on Oil Palm Plantation using Near Infrared: A Review and Modeling Challenges. *In 2nd University Consortium Graduate Forum (UCGF), University of the Philippines Los Banos*. Philippines.

Silalahi, D. D., & Midi, H. (2018, October). Method Selection via Robust Version of Some Scatter Corrections on Pretreatment of Near Infrared Spectral Data. *International Conference on Mathematics: Pure, Applied and Computation, Institute Teknologi Sepuluh Nopember.* Indonesia.

Silalahi, D. D., Midi, H., Arasan, J., Mustafa, M. S., & Caliman, J. P. (2018, December). Nonlinear partial robust GM6 regression using kernel space on near infrared spectral data. *In 14th IMT-GT International Conference of Mathematics, Statistics and Its Applications, Thaksin University.* Thailand.

Silalahi, D. D., Midi, H., Arasan, J., Mustafa, M. S., & Caliman, J. P. (2019, August). Robust Wavelength Selection using Input Scaling of Filter-Wrapper Methods on Near Infrared Spectral Data of Oil Palm Fruit Mesocarp. *In the 62nd ISI World Statistics Congress.* Malaysia.

Silalahi, D. D., Midi, H., Arasan, J., Mustafa, M. S., & Caliman, J. P. (2019, September). Robust Reliable Weighted Average Partial Least Square Regression on Near Infrared Spectral Data. *In the 19th International Council of Near Infrared Spectroscopy Conference 2019 (NIR).* Australia.

UNIVERSITI PUTRA MALAYSIA

STATUS CONFIRMATION FOR THESIS / PROJECT REPORT AND COPYRIGHT

ACADEMIC SESSION : _______________

TITLE OF THESIS / PROJECT REPORT :

____

____

____

NAME OF STUDENT : ____

I acknowledge that the copyright and other intellectual property in the thesis/project report belonged to Universiti Putra Malaysia and I agree to allow this thesis/project report to be placed at the library under the following terms:

1. This thesis/project report is the property of Universiti Putra Malaysia.
2. The library of Universiti Putra Malaysia has the right to make copies for educational purposes only.
3. The library of Universiti Putra Malaysia is allowed to make copies of this thesis for academic exchange.

I declare that this thesis is classified as :

*Please tick (√)

☐	**CONFIDENTIAL**	(Contain confidential information under Official Secret Act 1972).
☐	**RESTRICTED**	(Contains restricted information as specified by the organization/institution where research was done).
√	**OPEN ACCESS**	I agree that my thesis/project report to be published as hard copy or online open access.

This thesis is submitted for :

☐	**PATENT**	Embargo from_____________ until ______________ (date) (date)

Approved by:

_______________________	_______________________
(Signature of Student) New IC No/ Passport No.:	(Signature of Chairman of Supervisory Committee) Name:
Date :	Date :

[Note : If the thesis is CONFIDENTIAL or RESTRICTED, please attach with the letter from the organization/institution with period and reasons for confidentially or restricted.]