PENALIZED QUANTILE REGRESSION METHODS AND EMPIRICAL MODE DECOMPOSITION FOR IMPROVING THE ACCURACY OF THE MODEL SELECTION

ALI SALEH AL-MASSRI AMBARK

UNIVERSITI SAINS MALAYSIA

2024

PENALIZED QUANTILE REGRESSION METHODS AND EMPIRICAL MODE DECOMPOSITION FOR IMPROVING THE ACCURACY OF THE MODEL SELECTION

by

ALI SALEH AL-MASSRI AMBARK

Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosophy

July 2024

ACKNOWLEDGEMENT

In the name of Allah, Most Gracious, Most Merciful

First of all, thanks to Allah, who gives me the strength and patience to accomplish this thesis. I want to offer my heartfelt thanks and appreciation to everyone who contributed and helped with the preparation and completion of this thesis.

I would like to express my heartfelt thanks, appreciation, and gratitude to my supervisor, Assoc. Prof. Dr. Mohd Tahir Ismail, for his continual guidance, continuous support, patience, valuable comments, and suggestions during the completion of my thesis. I have greatly benefited from his incorporeal support, insightful suggestions, and extensive knowledge of statistics.

I would also like to sincerely thank my mother, wife, brothers, sisters, son Saleh, and daughter Tasneem for their prayers as a constant source of inspiration, permanent support, and encouragement. In addition to the sacrifices, they have made and the care they have continually given me. I am not forgetting in loving memory my father, who passed away in 2017. Last but not least, I would also like to thank all of my friends, particularly those who have provided me with useful suggestions and encouragement.

ALI S.A. AMBARK

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF SYMBOLS

$m(t)$	The mean envelope
$U(t)$	Upper envelope
$L(t)$	Lower envelope
$C_k(t)$	The k-th intrinsic mode function
K	Number of intrinsic mode function components
$r(t)$	Residual of the original signal decomposition
$x(t)$	Original signal
$h_q(t)$	New function (IMF) component
y_i	The i-th response variable (actual value)
β_0	The intercept
x_{ij}	The j-th predictor variable of the i-th observation
β_j	The regression coefficient of the j-th predictor variable
ε_i	Random errors at time-period i
p	Number of predictor variables
q	Repetition indicator
R_i	Partial residual
$\mathbf{y}$	Vector of the response variable
$\mathbf{X}$	Matrix of the predictor variables
$\boldsymbol{\beta}$	Vector of unknown regression coefficients
$\boldsymbol{\varepsilon}$	Vector of random errors
$\widehat{\boldsymbol{y}}$	The estimated model
$\widehat{\boldsymbol{\beta}}$	Vector of estimated regression coefficients
n	Sample size
∂	Partial derivative
X^T	Transpose of matrix of the predictor variables
$(.)^{-1}$	Inverse function
$\|.\|_2^2$	L_2-norm square
$\hat{\beta}_{OLS}$	Vector of estimated regression coefficients through the OLS method
$Q_\tau(.)$	The conditional quantile function for the τ-th conditional quantile
$\in$	Belong to
ρ_τ	The check loss function

α	The penalization parameter
$P(\beta)$	The penalty function
τ	Level of quantiles
L_2-norm	Ridge regression
$\hat{\beta}_{Ridge}$	Vector of estimated regression coefficients through the ridge regression
(s, λ)	Tuning parameter
L_1-norm	LASSO regression
$\|\beta\|_1$	L_1-norm of the vector regression coefficients
$\hat{\beta}_{LASSO}$	Vector of estimated regression coefficients through the LASSO regression
$\hat{\beta}_{Enet}$	Vector of estimated regression coefficients through the elastic net regression
$cov(.)$	The covariance
$var(.)$	The variance
R_j^2	Coefficient of Determination of the predictor obtained through predicting j^{th}
VIF_j	Variance inflation factor for the j^{th} predictor
$x_j(t)$	The j-th original signal
$C_{jk}(t)$	The k-th intrinsic mode function components of the j-th predictor
r_{jk}	The j-th residual of the original signal decomposition
β_{jk}	The k-th regression coefficient of the j-th predictor variable
D	Number of folds
λ_1	Tuning parameter of LASSO penalty
λ_2	Tuning parameter of ridge penalty
$\hat{y}_i$	The predicted value of variable y_i
t	The time domain
λ_{min}	Lambda at minimum of the MSE
λ_{1se}	Lambda at minimum of the MSE with one standard error
log	Logarithmic function
$N(0,1)$	Normal distribution with zero mean and unit variance
$X^2(2)$	chi-square distribution with 2 degrees of freedom
ρ	interpredictor correlation
$S(.,.)$	The soft-thressholding factor
$\hat{\beta}_R$	Vector of estimated regression coefficients by the Ridge method
$\hat{\beta}_L$	Vector of estimated regression coefficients by the LASSO method
$\hat{\beta}_{Enet}$	Vector of estimated regression coefficients by the Enet method

LIST OF ABBREVIATIONS

ADMM	Alternating Direction Method of Multipliers
AEnetQR	Adaptive elastic net penalty
AIC	Akaike Information Criterion
App.1	First application
App.2	Second application
BIC	Bayesian Information Criterion
CHN	China
CHN/USD	China's daily closed exchange rates against USD
CSD	Chirplet Signal Decomposition
CV	Cross-Validated
DBN	Deep Belief Networks
EMD	Empirical Mode Decomposition
EMD-QREnet	Elastic Net penalized quantile Regression based on Empirical Mode Decomposition
EMD-QRL	LASSO penalized quantile Regression based on Empirical Mode Decomposition
EMD-QRR	Ridge penalized quantile Regression based on Empirical Mode Decomposition
Enet	Elastic Net Regression
Exp.	First experiment
Exp.2	Second experiment
FT	Fourier Transform
HHT	Hilbert-Huang Transform
HQRM	Hierarchical penalty in Quantile Regression with Multiple
HSA	Hilbert Spectral Analysis
ICM	Interpredictor Correlations Matrix
IMF	Intrinsic Mode Function
JAP	Japan
JAP/USD	Japan's daily closed exchange rates against the USD
KDE	Kernel Density Estimation
KKT	Karush–Kuhn–Tucker
KPSS	Kwiatkowski Phillips Schmidt Shin
LASSO	Least Absolute Shrinkage and Selection Operator

LQR	Linear Quantile Regression
LSWA	Least Squares Wavelet Analysis
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MASE	Mean Absolute Scaled Error
MCP	Minimax Concave Penalty
MSE	Mean Square Error
MYR	Malaysia
MYR/USD	Malaysia's daily closed exchange rates against USD
NQR	Nonlinear Quantile Regression
OLS	Ordinary Least Square
PE	Prediction Error
QR	Quantile Regression
QRA	Quantile Regression Averaging
QREnet	Elastic Net penalized quantile Regression
QRL	LASSO penalized quantile Regression
QRNN	Quantile Regression Neural Network
QRR	Ridge penalized quantile Regression
QRRF	Quantile Regression Random Forest
QRSVM	Quantile Regression Support Vector Machine
R	Ridge Regression
RESET	Ramsey Regression Equation Specification Error Test
RMSE	Root Mean Square Error
RSS	Residual Sum of Squares
SCAD	Smoothly Clipped Absolute Deviation
SD	Standard Deviation
SLQR	Sampling LASSO Quantile Regression
SNCD	Semismooth Newton Coordinate Descent
SR	Stepwise Regression
TAW	Taiwan
TAW/USD	Taiwan's daily closed exchange rates against USD
USD	United States of America Dollar
VIF	Variance Inflation Factor
WD	Wigner-Ville Distribution

WSFT	Wavelet Short-term Fourier Transform
WT	Wavelet Transform
D-CV	D -fold Cross-Validation

LIST OF APPENDICES

KAEDAH REGRESI KUANTIL PENALTI DAN PENGHURAIAN MOD EMPIRIK UNTUK MENINGKATKAN KETEPATAN PEMILIHAN MODEL

ABSTRAK

Dalam kajian ini dan beberapa kajian saintifik, pemboleh ubah yang diminati sering diwakili oleh proses siri masa, dan data siri masa tersebut selalunya tidak pegun dan tidak linear, mengakibatkan ketepatan model regresi yang terhasil dan kesimpulan yang kurang boleh dipercayai. Di samping itu, kaedah kuasa dua terkecil biasa adalah sensitif kepada data terpencil dan ralat berat hujung dalam data, dan beberapa peramal mungkin mengalami masalah multikolineariti. Lebih-lebih lagi, memilih pemboleh ubah yang berkaitan apabila menyuai model regresi adalah kritikal. Oleh itu, tiga kaedah berdasarkan gabungan algoritma penguraian mod empirikal (EMD) dan regresi kuantil penalti telah dicadangkan dalam kajian ini. Algoritma EMD menguraikan data siri masa tidak pegun dan tidak linear ke dalam koleksi terhingga komponen ortogonal yang dipanggil fungsi mod intrinsik dan komponen reja. Dalam beberapa kajian, komponen ini telah digunakan sebagai pembolehubah peramal baru untuk mengkaji tingkah laku pemboleh ubah sambutan. Kajian ini bertujuan untuk mengaplikasi kaedah EMD-QRR, EMD-QR, dan EMD-QREnet yang dicadangkan untuk mengenal pasti pengaruh komponen penguraian pemboleh ubah peramal asal ke atas pemboleh ubah peramal untuk membina model yang paling sesuai dan meningkatkan ketepatan ramalan. Tambahan pula, kajian ini membincangkan isu multikolineariti antara komponen penguraian. Untuk mengesahkan prestasi ramalan kaedah yang dicadangkan, kaedah yang dicadangkan dibandingkan dengan tiga kaedah regresi sedia ada yang digunakan dalam kajian lepas. Kajian simulasi dan analisis empirikal data sebenar telah dijalankan dalam kajian ini. Untuk kajian simulasi, dua eksperimen

telah dipertimbangkan menggunakan fungsi gelombang sinus. Set data sebenar digunakan dalam contoh ilustrasi: pasaran saham harian dan kadar pertukaran. Keputusan daripada eksperimen berangka dan aplikasi data sebenar menggambarkan bahawa kaedah yang dicadangkan berprestasi lebih baik daripada kaedah lain pada kuantil yang berbeza. Penemuan juga menunjukkan bahawa kaedah yang dicadangkan mempunyai prestasi unggul dalam anggaran, pemilihan pembolehubah apabila masalah multikolineariti hadir, dan membina model akhir yang bebas daripada multikolineariti dan tahan terhadap penyimpangan atau taburan berekor berat. Oleh itu, secara keseluruhan, regresi kuantil penalti berdasarkan EMD mempunyai ketepatan yang lebih tinggi dan lebih baik daripada kaedah lain.

PENALIZED QUANTILE REGRESSION METHODS AND EMPIRICAL MODE DECOMPOSITION FOR IMPROVING THE ACCURACY OF THE MODEL SELECTION

ABSTRACT

In this study, in several scientific studies, the variables of interest are often represented by time series processes, and such time series data are frequently non-stationary and non-linear, resulting in low accuracy of the resulting regression models and less reliable conclusions. In addition, the ordinary least squares method is sensitive to outliers and heavy-tailed errors in data, and several predictors may suffer from multicollinearity problems. Moreover, selecting the relevant variables when fitting the regression model is critical. Therefore, three methods based on a combination of the empirical mode decomposition (EMD) algorithm and penalized quantile regression have been proposed in this study. The EMD algorithm decomposes the non-stationary and non-linear time series data into a finite collection of approximately orthogonal components called intrinsic mode functions and residual components. In several studies, these components have been employed as novel predictor variables to study the behaviour of the response variable. This study aims to apply the proposed EMD-QRR, EMD-QR, and EMD-QREnet methods to identify the influence of the decomposition components of the original predictor variables on the response variable to build a model that has the best fit and improve prediction accuracy. Furthermore, this study deals with the multicollinearity issue between the decomposition components. To verify the prediction performance of the proposed methods, the proposed methods are compared with three existing regression methods used in previous studies. Simulation studies and empirical analysis of the real data were

carried out in this study. For simulation studies, two experiments were considered using the sine wave function. The real datasets are applied in illustrative examples: the daily stock markets and exchange rates. Both numerical experiments and empirical results show that the proposed methods perform better than other methods at different quantiles. Additionally, the proposed methods can achieve low prediction errors and produce a model free from multicollinearity and resistant to outliers or heavy-tailed distributions compared to the existing methods. The proposed EMD-QRL and EMD-QREnet methods can select the decomposition components that have a significant impact on the response variable. Thus, overall, penalized quantile regression based on EMD has higher accuracy and is superior to other methods.

CHAPTER 1

INTRODUCTION

1.1 Background and Motivation

In real-world scenarios, whether natural or artificial, time series processes are often used to represent variables of interest, such time series are frequently non-stationary and non-linear (Masselot et al., 2018). These data are most likely both non-linear and non-stationary, The decomposition of non-stationary and non-linear time series is an important issue to consider when analysing. The decomposition of non-stationary and non-linear time series is an important issue to consider when analysing. Meanwhile, there is a dearth of analytical methods for dealing with time series data (Huang, 2005). However, several algorithms, such as the Fourier transform method (Titchmarsh, 1948), the Wigner distribution (Classen & Mecklenbrauker, 1980), and wavelet analysis (Chan, 1995), have been applied in the literature to analyse time series data (Al-Jawarneh & Ismail, 2022; Huang, 2005).

Regression analysis is a robust statistical method widely used in empirical applications in various fields such as finance, economics, environmental, social, and life sciences. Regression models are commonly estimated using the ordinary least squares (OLS) method, which estimates the conditional mean of the response variable. In other words, the relationship between the predictor variables and the response variable in the coordinate plane is assessed with a mean regression line, despite the OLS method having excellent properties such as linearity, efficiency, and unbiasedness. However, it does not provide information on other aspects of the response variable's distribution, such as ignoring distribution shape, normality of errors and robustness to violations. Moreover, it is well known for being very sensitive

to the existence of outliers or heavy-tailed distributions, which means the estimation efficiency might be reduced naturally and can lead to misleading inferences (Mendez-Civieta et al., 2021; Tian & Song, 2020; Yousif & Housain, 2021; Yuzbasi et al., 2018). In addition, when dealing with heterogeneous data in regression analysis, targeting only a mean function is often insufficient to comprehensively understand the relationship between the response and predictor variables (Hu et al., 2021). Heterogeneous data in regression refers to data that exhibit variability in their characteristics, such as different distributions, variances, or patterns across subsets of the data (Bernardi et al., 2016).

To overcome these inadequacies of classical regression, quantile regression (QR) was suggested by Koenker and Bassett (1978) as an alternative to the ordinary least squares (OLS) method. QR provides much more information about the whole conditional distribution of a response variable instead of just the average value and gives an overall evaluation of the influence of the predictor variables at various quantiles τ of the response variable (Koenker, 2005; Tian & Song, 2020; Y. Wu & Liu, 2009). Quite recently, quantile regression (QR) has grown into a fundamental and commonly used technique to examine the relationship between the response variable and the predictor variables at various quantiles of the conditional distribution function, providing more comprehensive visibility of the phenomena under study. Moreover, quantile regression provides a more comprehensive understanding of the relationship between predictor variables and the response variable by characterizing the conditional mean and the entire conditional distribution, including its location, scale, and shape. This makes it a valuable tool for analyzing data where the assumptions of OLS regression may not hold, or when you want a more nuanced understanding of the data's distributional properties.

The most notable feature of QR is that this approach is robust against outliers and insensitive to heterogeneity. Thus, it can handle non-normal errors common in several real-world applications. In addition, quantile regression provides a considerably more complete understanding of the impact of predictor variables on the various quantiles of the response variable distribution than OLS regression captures. Therefore, this method provides accurate insight into the relationship between response and predictor variables at the upper and lower tails. Also, this method provides information on the conditional distribution of the response variable's location, scale, and shape. Overall, the quantile regression can quantify the entire conditional distribution of the response variable conditional on predictors and assess the predictor variables' influences at various quantiles of the response variable (Koenker, 2005). These unique advantages attracted a great deal of interest in the literature, and quantile regression is applied in several scientific fields, such as finance, economics, social science, medicine, and growth charts (Alkenani & Msallam, 2019; Benoit et al., 2013; Mendez-Civieta et al., 2021).

Recently, the empirical mode decomposition (EMD) approach was presented by Huang et al. (1998), which is an intuitive, straight, and adaptable method for decomposing non-linear and non-stationary time series data. This approach is the first part of the Hilbert-Huang transform (HHT). In contrast to traditional techniques such as wavelet decomposition (Chan, 1995) and Fourier decomposition (Titchmarsh, 1948), EMD imposes no a priori limitations on the data and allows it to speak for itself. Despite this approach being completely derived from empirical evidence and lacking a formal mathematical foundation, it may efficiently divide a data series into distinct components, each corresponding to a particular oscillation frequency.

Since its inception, this technique has been used in a wide variety of fields, including economics (Huang et al., 2003), engineering (Yang et al., 2003), medicine (Yang et al., 2011), physics (Varadarajan & Nagarajaiah, 2004a), and environmental science (Huang et al., 1999). This technique supplies more accurate findings in many situations than traditional methods, uncovering novel patterns within the analyzed data sets (Qin et al., 2016). The EMD method decomposes the non-stationary and non-linear signal into several components called intrinsic mode functions (IMFs) and one residual. These decomposition components are of different wavelengths, amplitudes, and frequencies, which are functionally significant (Huang, 2014). These decomposition components may be utilized as novel predictor variables to study their influence on the response variable (Al-Jawarneh & Ismail, 2021; Masselot et al., 2018; Qin et al., 2016).

Parameter estimation, model selection, and variable selection are important considerations in regression analysis. Therefore, many robust regression methods for fitting multiple regression models have been proposed. Among these, the most common regularization approach is penalized least squares regression. For example, the Ridge (Hoerl & Kennard, 1970), LASSO (Tibshirani, 1996), the Elastic Net (Zou & Hastie, 2005) and so on. However, since the least squares criteria are used, none of these methods resist outliers or heavy-tailed error. Alternative robust procedures have been developed in the literature. One such appealing, robust procedure is penalized quantile regression methods that provide promising approaches for parameter estimation, model selection, and variable selection in the event of the existence of outliers or heavy-tailed errors (Ajeel & Hashem, 2020; Su & Wang, 2021).

In regression analysis, there are some problems that may have an impact on model selection prediction accuracy. Such problems are that the time series data used

in regression are often non-stationary and non-linear, the presence of outliers or heavy-tailed errors, and several predictors may suffer from multicollinearity problems. hese issues in regression may increase parameter estimations' variability, rendering the final result less dependable. Regularization techniques, such as LASSO, Ridge, and Elastic Net, are used to add a penalty term to the regression equation. This penalty helps to estimate the quantile regression coefficients, prevent overfitting and improve the model's prediction performance. The empirical mode decomposition (EMD) technique is also implemented for analysing nonlinear and non-stationary signals, providing insights into their underlying oscillatory components and patterns. The EMD method and the penalized quantile regression methods, namely, Ridge (QRR), LASSO (QRL), and elastic-net (QREnet) regression, are proposed to address these issues to enhance the predictive accuracy and improve model selection.

1.2 Problem Statement

In the case of time series data, the regression analysis assumes limitations on all variables before estimating to enhance predictive accuracy and model selection. However, five significant challenges currently exist which include the following:

a) In the regression analysis, the variables of the time series data are supposed to be linear. In other words, the relationship between the time series observations is linear. However, the assumption of linearity in time series data stemming from the real world is not always realistic.

b) In the regression analysis, time series data variables are supposed to be stationary in the sense that the data has the property that the mean, variance and autocorrelation structure do not change over time (Adarsh & Janga Reddy, 2019; Moore et al., 2018a). Numerous datasets stemming from the real world are often

non-stationary time series datasets. Some traditional methods were used to overcome this problem. However, these methods lead to some lost data features and valuable information from the original data. Existing methods are not adaptive and highly efficient, which affects the accuracy of the findings (Al-Jawarneh & Ismail, 2021, 2022; Huang, 2014).

c) Even though the OLS method has excellent properties, e.g., linearity, efficiency, and unbiasedness, it does not provide information on other aspects of the distribution of the response variable. Because its procedures typically focus on the mean of the response, it is known to be particularly sensitive to the existence of outliers or heavy-tailed distributions, and heterogeneity implying that estimate efficiency may be naturally lowered in this case. Sometimes researchers may be interested in modelling other values than the mean of the response variable, for instance, the median or other quantiles (Amin et al., 2015; Mendez-Civieta et al., 2021).

d) It is assumed that there is no dependence among the predictor variables, that is, there is no correlation between two or more predictor variables. However, if this assumption is violated, the issue of multicollinearity arises. It is impossible to quantify the unrivalled impact of a specific indicator in the case of multicollinearity. Furthermore, the regression coefficients have a huge sample variance as well as incorrect signals, which influences both inference and estimation. Thus, multicollinearity is a major issue in linear regression analysis (Ali et al., 2021).

e) Variable selection is also very important in quantile regression, similar to the linear regression model when the number of predictors is large. However, keeping irrelevant variables in the model, is undesirable because it makes the model

difficult to interpret and may impair its predictive ability. As a result, when fitting the quantile regression model, it is critical to select the relevant variables (Alkenani & Msallam, 2019; Amin et al., 2015).

To address these challenges this study introduces three novel hybrid penalized quantile regression methods based on empirical mode decomposition (EMD). These new methods aim to overcome the shortcomings of existing quantile regression models, ultimately enhancing the predictive accuracy of the model selection.

1.3 Research Objectives

The main aim of this study is to enhance the accuracy of model selection for non-linear and non-stationary time series datasets through the use of the EMD technique and penalized quantile regression techniques, with the following objectives:

a) To apply EMD multi-scale data decomposition on the non-stationary and non-linear predictors by decomposing each non-stationary and non-linear predictor into a finite set of decomposition components, which will represent the new predictor variables in this study.
b) To investigate the relationship between the decomposition components via the EMD approach and the response variable by selecting the necessary decomposition components that greatly influence the response variable in the existence of multicollinearity.
c) To improve the prediction accuracy of the model selection by selecting the necessary decomposition components.
d) To develop three hybrid methods by combining EMD and penalized quantile regression (QRR, QRL, and QREnet) to improve the prediction accuracy of the model selection and deal with multicollinearity among decomposed components.

1.4 Scope of the Study

a) The proposed methods are based on the use of decomposition components by the EMD approach as new predictor variables in a penalized quantile regression analysis to improve the performance of existing methods in the prediction accuracy of model selection.

b) Three traditional methods were used to test and validate the performance of the proposed methods.

c) This study uses two datasets (simulation and real data) to achieve these objectives. For simulation datasets, the simulation data were generated by two simulation experiments using the sine wave function. The other datasets (real data), two empirical applications will be used to evaluate the performance of the proposed methods relative to existing methods: the first application uses the daily close stock market data of three countries, namely, Japan (Nikkei Index), China (Shenzhen Component Index) as predictor variables, and Singapore (Singapore Exchange Limited) as the response variable between 5 January 2015 and 29 December 2022, and the second application is the daily exchange rates of four countries which are Taiwan (TAW), Malaysia (MYR), Japan (JAP), and China (CHN) between March 27, 2015, and December 30, 2022.

1.5 Significance of the Study

This study presents new techniques to improve the predictive accuracy of model selection. The new techniques are based on a combination of EMD, an efficient method for dealing with nonlinear nonstationary time series data, and some penalized quantile regression methods. Three new techniques are proposed, namely the EMD-QRR, EMD-QRL, and EMD-QREnet methods. The proposed techniques are accurate

and reliable in determining the principal components of the original predictor that have the greatest influence on the response variable using the EMD method. The EMD method assures the reliability of the variables' relationships in the time and frequency domains. In this way, examining the connection between the variables becomes conceivable.

This study has provided a significant contribution towards developing penalized quantile regression models to determine the decomposition components of the original time series predictors that display the most substantial effects on the response variable and deal with multicollinearity and heterogeneity among the decomposition components at different quantiles with high prediction accuracy in model selection compared with the existing methods.

1.6 Limitations of the Study

In this study, the proposed methods focus on improving the prediction accuracy of the model selection, which seeks to comprehend the relationship between the decomposition components via EMD algorithm and the response variable from a different perspective, i.e., multicollinearity and heterogeneity problems. However, time series data frequently have autocorrelation issues, such as the degree of resemblance between the present value and its preceding values. If such a problem arises, it is a new area of research. As a result, these issues will be addressed in future studies.

1.7 Organization of thesis

Chapter 1 introduces an overview of the problems that motivate this study, its objectives, scope, significance, and limitations. Chapter 2 deals with the literature review of the present study. The chapter aims to review previous studies that are

interrelated to the current study, which consists of the following: an introduction, decomposition of time series based on the Hilbert-Huang transform; the EMD algorithm, penalized quantile regression; the multicollinearity issue and the choice of the optimal tuning parameter. Chapter 3 is the research methodology for the current study. This chapter displays the proposed methods: the hybrid of EMD-QRR, EMD-QRL, and EMD-QREnet. The multicollinearity test and statistics measures of performance evaluation of the proposed methods are also described in Chapter 3. It ends with a summary. Chapter 4 deals with the simulation study of the current study, which incorporates the introduction, simulation details, simulation results of the three case studies, discussion, and conclusion. Chapter 5 deals with the empirical analysis by applying the original time series data in two applications, discussion, and conclusion. Chapter 6 deals with the conclusion of the main findings in the thesis and recommendations for future work.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter contains eight sections. The second section provides an overview of the signal decomposition methods using the Hilbert-Huang transform. The third section describes the development and application of the empirical mode decomposition (EMD) algorithm in different areas. In addition, the list of recent studies that use EMD in regression techniques is also displayed at the end of this section. The fourth section presents the Ordinary Least Square method (OLS). The fifth section displays the quantile regression method (QR) in the literature. This section also presents several recent studies that apply the QR technique with penalized regression, and a list of the several penalized quantile regression techniques used in this study is also displayed at the end of this section. The sixth section presents the multicollinearity issue. The seventh section presents the tuning parameter selection and highlights the D-fold cross-validation techniques. The last section is the summary.

2.2 Decomposition of Time Series based on Hilbert-Huang Transform

In many real-world systems, whether natural or human-made, the variables of interest are typically represented using time-series processes. These datasets are almost certainly both non-linear and non-stationary. Analyzing such complex systems is a critical challenge in various fields of study. Many algorithms have been applied in the literature for analyzing non-stationary and non-linear time-series data. For example, the Fourier transform method by Titchmarsh (1948) assumes that the data is stationary and linear, and wavelet analysis by Chan (1995) was designed for linear and stationary or non-stationary signals. Recently, Ghaderpour and Pagiatakis (2017) proposed least

squares wavelet analysis (LSWA), a technique for analyzing non-stationary and unequally spaced time series with low or high components. An essential condition for representing non-linear and non-stationary data is an adaptive basis. Here, adaptive means that the basic definition must be based on and derived from the data. The a posteriori adaptive basis provides a completely different approach from the established mathematical model for data analysis.

Therefore, a new method named the Hilbert-Huang transform (HHT) was introduced by Huang et al. (1998). It was able to satisfy the requirements of the posterior basis function necessary for adaptive data analysis. The HHT is explicitly designed for analyzing non-linear and non-stationary data. The HHT mainly comprises empirical mode decomposition (EMD) and Hilbert spectral analysis (HSA). The main principle of HHT is EMD, which decomposes non-stationary and non-linear signals into a finite and often small number of orthogonal non-overlapping time-scale signals or components. The EMD analyzes signals while keeping the time domain of the signal. Then the HHT technique will be applied to all the extracted components to obtain instantaneous frequency data (Huang, 2014).

2.3 Empirical Mode Decomposition

Empirical Mode Decomposition (EMD) was introduced by Huang et al. (1998) as a new and effective approach to decomposing a non-linear and non-stationary signal. This approach effectively decomposes complex and multiscale signals into finite collections of approximately orthogonal components, named are Intrinsic Mode Functions (IMF), and residual components through an iterative process called the sifting process (Maheshwari & Kumar, 2014; Moore et al., 2018b). The key feature of EMD is its ability to adaptively decompose signals, making it particularly suitable for

capturing the essence of a signal's dynamics. The intrinsic mode functions (IMFs) hold a special role in this process, as they represent localized instantaneous frequencies, thereby allowing for a comprehensive characterization of the signal's behaviour. This characterization is realized through the computation of the IMF's instantaneous frequency using the analytic signal method (a process known as the Hilbert-Huang transform). Specifically, the low-level IMFs encapsulate high local frequencies, while the high-level IMFs encompass low local frequencies, making EMD a versatile tool for signal analysis (Huang, 2014).

2.3.1 Intrinsic Mode Function (IMF)

IMFs represent a generally simple oscillatory mode as an alternative to the straightforward harmonic function. An IMF is defined as any function with the same number of extrema and zero crossings and whose envelopes are symmetric concerning zero. This definition ensures that a Hilbert transform behaves correctly within the IMF (Huang et al., 1998).

Each IMF component function must fulfill the following two conditions:

i. The number of local extreme values (maxima and minima) and the number of zero-crossings must be equal or differ at most by one.

ii. The local mean must be zero, defined as the mean of the upper and lower envelopes.

$$m(t) = \frac{U(t)+L(t)}{2} = 0 \quad (2.1)$$

where $m(t)$ is the mean envelope, $U(t)$ is the upper envelope, and $L(t)$ is the lower envelope.

The first condition seems required for oscillation data; the second condition requires the symmetric upper and lower envelopes of IMF, as the IMF component is decomposed from the original data; finding the real envelopes is quite difficult due to the data's non-linear and non-stationary nature. Only a few functions have known envelopes, such as the constant amplitude sinusoidal function (Al-Jawarneh et al., 2021; Lu, 2007).

2.3.2 Sifting Process

Huang et al. (1998) invented the EMD method to break up the original data into a series of IMF via a process named the Sifting process of EMD. The idea is to separate the data into slow-varying local mean and fast-varying symmetric oscillation parts. The resulting oscillations are designated as IMFs, while the local mean constitutes the residue. This residue is then utilized as the input data for further decomposition, and this iterative process continues until no additional oscillatory components can be extracted from the remaining residue. Through this method, the original data, $x(t)$ can be constructed back as in Equation (2.2) (Huang et al., 1998; Huang, 2014; Lu, 2007).

$$x(t) = \sum_{k=1}^{K} C_k(t) + r(t). \quad (2.2)$$

Here $x(t)$ indicates the original signal, $r(t)$ represents the residue of the original signal decomposition, and $C_k(t)$ represents the k-th intrinsic mode function (IMF).

Initially, due to the upper and lower envelopes being unknown, on each step of the decomposition, to approximate the envelopes and obtain the IMF and residue, a

repetitive sifting process is applied as follows:(Awajan et al., 2019; Huang et al., 1998; Huang, 2014)

Step 1: The original signal $x(t)$ is entered for the sifting process on the assumption that the value of the two repetition indicators is $(q = 1, k = 1)$.

Step 2: Identify all local extrema, including minima and maxima of a time series signal $x(t)$. For more illustration, see Figure 2.1.

Figure 2.1 displays an example of step 2. The black line is the original signal $x(t)$, while the red circle point on the upper line is the local maximum. On the other hand, the blue circle point on the lower line symbolizes the local minimum.

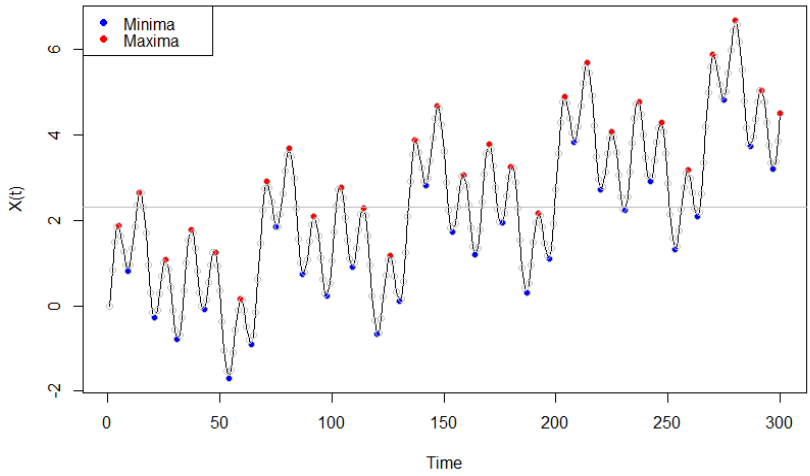

Figure 2.1 Local extreme of the original signal $x(t)$.

Step 3: Produce Envelope. Connect all local extrema with a cubic spline line to generate the upper envelope $U_q(t)$ and lower envelope $L_q(t)$, respectively.

Step 4: The mean envelope, $m_q(t)$, is determined by the mean of the upper and lower envelopes by using Equation (2.3).

$$m_q(t) = \frac{U_q(t)+L_q(t)}{2} \quad (2.3)$$

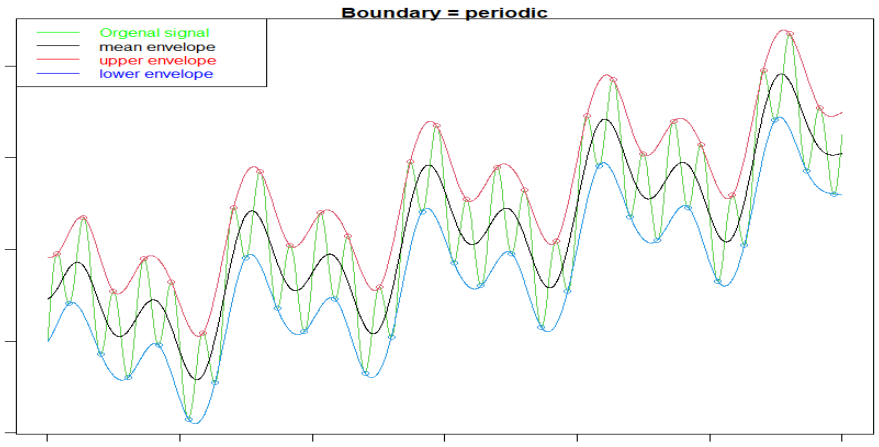

Figure 2.2 Upper, lower, and mean envelopes of the original signal $x(t)$.

Figure 2.2 displays an example of steps 3 and 4. The green line represents the original signal $x(t)$, the red line represents the upper envelope $U_q(t)$, and the blue line lower envelope $L_q(t)$, as indicated in step 3. The black line represents the mean envelope of the upper and lower envelopes, as explained in Step 4.

Step 5: Subtract the mean envelope $m_q(t)$ from the original time series $x(t)$ to obtain the component $h_q(t)$ as shown in Equation (2.4).

$$h_q(t) = x(t) - m_q(t) \tag{2.4}$$

Step 6: Check whether series $h_q(t)$ is an IMF or not, according to IMF conditions (presented in Section 2.3.1).

i. If not an IMF, substitute the function $h_q(t)$ with $x(t)$, update the iteration indicator such that it equals $q = q + 1$, and repeat the sifting process, which consists of step 2 to step 5 until $h_q(t)$ meets the conditions of IMF.

ii. If the function $h_q(t)$ is an IMF according to the definition of IMF, then $h_q(t) = C_k(t)$, saves the $C_k(t)$ result obtained and go to step 7.

Step 7: Calculate the residual using the IMF and the signal $x(t)$ as on the formula:

$$r_k(t) = x(t) - h_q(t) \tag{2.5}$$

Check whether the residue function $r_k(t)$ is a monotonic or constant function or satisfies the stopping criterion of the standard deviation (SD_q), which needs a small

normalized squared difference between two successive sifting operations. The difference is defined in equation (2.6).

Save the residue and all the IMFs obtained, and the sifting process stops.

If the residue is not a monotonic or constant function, substitute the $r_k(t)$ with $x(t)$ and repeat the sifting process which consists of step 2 to step 7 with k=k + 1 until $r_k(t)$ is a monotonic or constant function or satisfie///.s stopping criterion SD_q.

$$SD_q = \sum_{t=0}^{T} \frac{(h_{q-1}(t) - h_q(t))^2}{h_{q-1}^2(t)} \tag{2.1}$$

The refinement process (steps 2 to 7) needed to extract the IMF and residual components, requires a certain number of iterations and is named a sifting process. All the steps are summarized in Figure 2.3, the EMD method tree graph, and Figure 2.4 is taken from (Al-Jawarneh & Ismail, 2022) and is a flowchart for the sifting process.

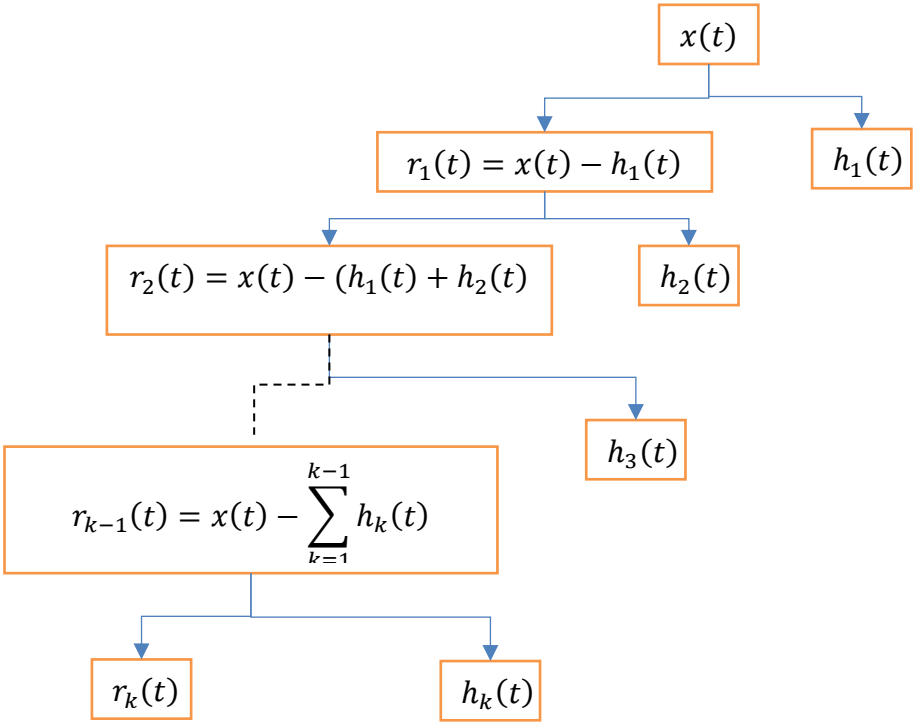

Figure 2.3 The EMD Decomposition tree

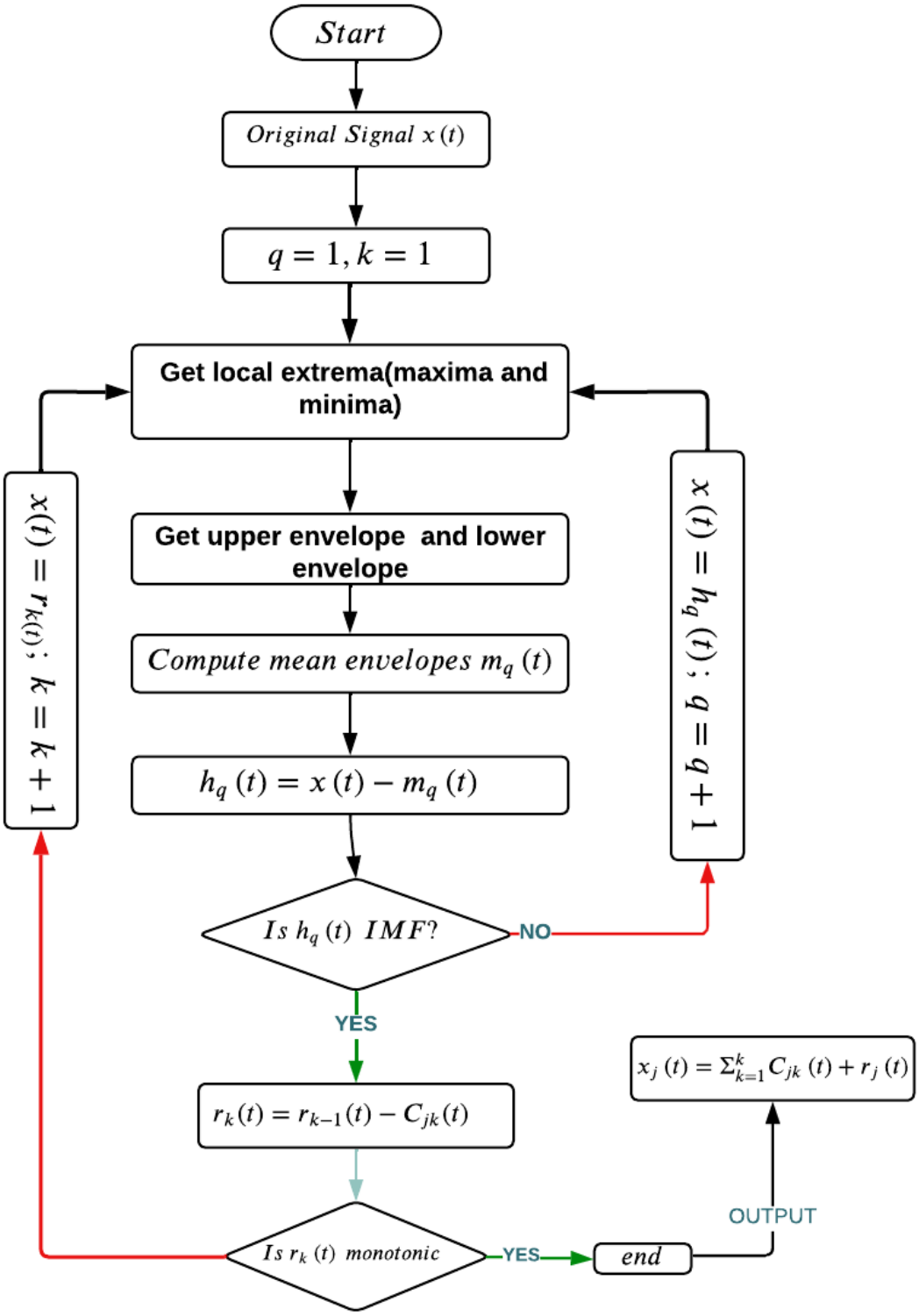

Figure 2.4 Flowchart for the sifting process

2.3.3 Previous Applications, Extensions and Limitations of the EMD

The EMD has been widely used in several scientific fields, such as economics (Hossain & Ismail, 2020), financial time series (Jaber et al., 2014), medicine (Masselot et al., 2018), physics (Varadarajan & Nagarajaiah, 2004), mechanical engineering (Zhang et al., 2010), electronic engineering (Suvasini et al., 2015), civil and construction engineering (OBrien et al., 2017), short-term traffic speed (Wang et al., 2016; Zheng et al., 2017), and environmental science (Wei et al., 2018). This technique supplies more accurate findings in many situations than traditional methods, uncovering novel patterns within the analyzed data.

However, although the EMD approach has proven highly effective and specialized in analysing non-linear and non-stationary signals, its application has some limitations. For instance, most phases during the sifting process are not mathematically determined and lack a reliable mathematical theory, which has become a major problem restricting the application of EMD (Awajan et al., 2018; Liu & Chen, 2019, Flandrin et al., 2004; Moore et al., 2018b). Several research works have proposed theoretical assumptions about the EMD procedure, predominately defined as algorithmic steps (Wu et al., 2001). The number of IMF components extracted from a signal equals log2N. Furthermore, Peel et al. (2005) argued that the average period might be computed for each IMF by 2×N. Kizhner et al. (2006) presented several theoretical basics for the EMD algorithm by presenting three assumptions on the EMD sifting process. However, the theoretical component of EMD remains poorly, as Rilling and Flandrin (2006) described. Li et al. (2017) used the differential operation to solve the mode mixing of the IMF components. The proposed methodologies in this study can overcome these limitations with basic EMD.

Many studies have provided an extension of the EMD method. Among the studies are He et al. (2017), which suggested that 3D EMD decomposes a volume into three-dimensional IMFs. Then Vatchev and Sharpley (2008) demonstrated that any function with simple zeros and extrema could be decomposed into two or fewer weak IMFs. Wu and Huang (2009) presented the ensemble EMD (EEMD) as an extension of EMD. Yeh (2012) proposed a method for computing complicated bi-dimensional EMD that may be applied to analyze two-dimensional signals. Rehman and Mandic (2010) have also successfully developed and applied the extensions to multivariate works. (Torres et al. (2011) proposed the EEMD algorithm, which enables an exact reconstruction of the original signal and better spectral separation of the modes.

Many researchers have compared the EMD method with other technical decomposition methods. The results of the studies showed that the EMD algorithm exhibited high accuracy in dealing with non-stationary and non-linear signals compared with other technical decomposition methods in various fields. Wang et al. (2011) compared the EMD method and wavelet decomposition (WD) in non-stationary and non-linear time series data analysis. Lu et al. (2013) compared the EMD method and the method with chirplet signal decomposition (CSD) method used for ultrasonic signal feature extraction. Ghosh et al. (2014) compared the EMD with Fourier transform (FT), wavelet short-term Fourier transform (WSFT), and wavelet transform (WT) to denoise an electrocardiogram signal.

2.3.4 Statistical Regression Methods combined with EMD

Many studies have combined the EMD algorithm with other established statistical regression or forecasting methods. The decomposition components obtained via EMD are used as new predictor variables to predict their behaviors and impacts on

other response variables by utilizing appropriate models or in various statistical situations, such as, Yang et al. (2011) studied the ordinary least squares (OLS) regression analysis and forward stepwise regression (SR) methods based on the EMD approach by using decomposition components via EMD of the weather variables (predictor variables), which are the pressure, temperature, humidity, sunshine duration and maximal wind speed, and (response) variables which is headache incidence. They used the same methodology to examine the associations between air pollution, weather, and unemployment variables (predictor variables) and the decomposition components via EMD of the suicide variable (response variable). The residual (trend) component was removed to avoid spurious regression and multicollinearity.

Shen et al. (2012) used a combination of ridge regression with the ensemble EMD method to reduce decomposition error and address the mode mixing problem. Then Shen and Lee (2012) applied LASSO regression based on Ensemble EMD (EEMD) to reduce the errors caused by the outliers on the ultrasound imaging for the blood flow velocity dataset. After that, Chu et al. (2018) proposed a new method for combining Lasso regression and deep belief networks (DBN) with ensemble empirical mode decomposition (EEMD) to investigate the relationship between multiscale climate predictors and the decomposition components of nonstationary and nonlinear monthly streamflow on the Tennessee River in the USA. They found that their proposed model can significantly improve the accuracy of monthly streamflow forecasting.

Adarsh (2016) applied multivariate EMD and stepwise linear regression (SLR) to predict the monthly rainfall (response variable) with decomposition components of the four predictor variables, namely, mean sea level pressure, relative humidity, surface temperature, and wind velocity in the Kerala meteorological subdivision in

India. At the same time, Adarsh et al. (2018) used the same methodology to estimate the relationship between the reference evapotranspiration (response variable) and the four predictor variables, such as solar radiation, air temperature, relative humidity, and wind velocity (predictor variables). Later Adarsh and Janga Reddy (2019) applied the same methodology with various variables.

Qin et al. (2016) presented a LASSO regression based on the EMD method for choosing decomposed components that exhibit the most substantial effects on the response variable. This method is compared with the OLS and Ridge methods based on the EMD. Both numerical experiments and applications on the two Chinese stock markets are applied in this study. In comparison, Masselot et al. (2018) applied LASSO regression and multivariate EMD to choose the decomposition components that have a substantial influence on the response variable/variables in the two models proposed, one of which decomposes the predictor variables only and the other decomposes the predictor and response variables via EMD. The proposed methodology is applied to study the relationship between weather (a response variable) and cardiovascular mortality (predictors variables) in Montreal, Canada. These decomposition components obtained via the EMD algorithm are used as new predictor variables to predict their effects and behaviours about the response variable.

Recently, Al-Jawarneh et al. (2020) presented elastic net regression based on the EMD for selecting decomposed components that exhibit the most potent effects on the response variable, and multicollinearity between the decomposition components was dealt with. A numerical experiment and actual time series data were applied to the two Chinese stock markets (the Shanghai Composite Index and the Shenzhen Component Index). Al-Jawarneh et al. (2021) and Al-Jawarneh and Ismail (2021) used the same methodology but with multivariate predictors with different numerical

experiments and real data. A more recent study by Alsayed (2022) employed the elastic net regression method based on empirical mode decomposition to address precisely the non-stationary and non-linearity characteristics of the variables, and it can also tackle the multicollinearity between the predictors to check the behavior and trend of the Turkish stock market (XU100). The predictors include COVID-19 infected cases and financial country-level variables named credit default swap, foreign exchange rate USD/TL, and TL reference interest rate.

2.4 Classical Linear Regression

Consider the following multiple linear regression model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i$$
$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i \quad (2.7)$$

where $i = 1,2,3, \ldots \ldots, n$, $j = 1,2, \ldots \ldots p$, y_i is response variable, x_{ij} represents a set of predictor variables that could be associated with the response variable, β_0 is the intercept, β_j indicate the regression coefficient of the x_{ij} and ε_i are random errors where $E(\varepsilon) = 0$, $Var(\varepsilon) = \sigma^2$, and the errors are uncorrelated (Montgomery et al., 2012). Equation (2.7) can be expressed in the matrix form as follows:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (2.8)$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \ \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}, \ \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

$\mathbf{y}$ is a $(n \times 1)$ vector of observations on the response variable, $\mathbf{X}$ is a $(n \times p)$ matrix of observations on the predictor variables (a p-dimensional vector of predictor variables), $\boldsymbol{\beta}$ is a $(p \times 1)$ vector of unknown regression coefficients, $\boldsymbol{\varepsilon}$ is a $(n \times 1)$

vector of random errors, that supposed to be normally distributed with $E(\varepsilon) = 0$ and , $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\boldsymbol{T}}) = \sigma^2\mathbf{I}_{\boldsymbol{n}}$.

The Ordinary Least Squares (OLS) method is a widely employed technique for estimating regression coefficients. The OLS estimates the coefficients by minimizing the residual sum of squares (RSS). Consequently, the estimated model $\hat{\mathbf{y}}$ for the true model $\mathbf{y}$ in Equation (2.8) is obtained as follows: (Melkumova & Shatskikh, 2017; Wetherill & Seber, 1977).

$$\hat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} \tag{2.9}$$

Thus, the sum of the squared (RSS) differences between the actual $\mathbf{y}$ and estimated $\hat{\mathbf{y}}$ values in the matrix form given as follows:

$$L = \sum_{i=1}^{n} \boldsymbol{\varepsilon}_{\mathbf{i}}^{\mathbf{2}} = \boldsymbol{\varepsilon}^{\boldsymbol{T}}\boldsymbol{\varepsilon} = \left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)^{T}\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) \tag{2.10}$$

Differentiate L Equation (2.10) in terms of the unknown parameters and equal the derivatives to zero, which is:

$$\left.\frac{\partial L}{\partial \boldsymbol{\beta}}\right|_{\widehat{\boldsymbol{\beta}}} = \mathbf{2X^{T}X}\widehat{\boldsymbol{\beta}} - \mathbf{2X^{T}y} = 0$$

$$\mathbf{X^{T}X}\widehat{\boldsymbol{\beta}} = \mathbf{X^{T}y} \tag{2.11}$$

By using Equations (2.11), the least squares estimator of β can be calculated using the following formula

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X^{T}X})^{-\mathbf{1}}\mathbf{X^{T}y} \tag{2.12}$$

Assuming that the inverse matrix $(\mathbf{X^{T}X})$ is a full rank matrix. The $(\mathbf{X^{T}X})^{-1}$ matrix will always exist if the regressors are linearly independent, that is, if no column of the X matrix is a linear combination of the other columns. A unique solution for the regression coefficients is obtained from equation (2.12) (Douglas et al., 2012).