ENHANCING MODEL SELECTION BASED ON PENALIZED REGRESSION METHODS AND EMPIRICAL MODE DECOMPOSITION

ABDULLAH SULEIMAN SALEH AL JAWARNEH

UNIVERSITI SAINS MALAYSIA

2021

ENHANCING MODEL SELECTION BASED ON PENALIZED REGRESSION METHODS AND EMPIRICAL MODE DECOMPOSITION

by

ABDULLAH SULEIMAN SALEH AL JAWARNEH

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

February 2021

ACKNOWLEDGEMENT

First of all, I thank Allah, the Almighty, for his care and protection. I would like to express my sincere gratitude and appreciation to all those who assisted and contributed to the preparation and completion of this thesis.

I extend my sincere respect and gratitude to my supervisor Dr. Mohd Tahir Ismail for agreeing to be my supervisor, for his assistance, patience, support, cooperation and care. I will never forget the huge efforts he made to read the drafts of my work and the insightful comments and suggestions he provided, which enriched the quality of the thesis.

My special thanks and respect go to the members of the examining committee for their efforts in reading and discussing this thesis. I am very grateful to the Dean, lecturers, the staff of the department and all postgraduate students at the School of Mathematical Sciences, Universiti Sains Malaysia. All of them have contributed in a way or another for the success of this work. I also wish to thank Dr. Ahmad Awajan at the Department of Mathematics, Al-Hussein Bin Talal University, Jordan, for his help in this research project.

My warm thank and sincere appreciation go to my beloved father Suleiman, mother Fatimah, wife Heba Nasr, son Shadi and daughter Selena for their encouragement, support and patience during a hard time. I will be extremely grateful to them for the rest of my life. My thanks also go to my dear sisters and brothers for their help and moral support during my study.

Abdullah Al Jawarneh

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

adLASSO	Adaptive Least Absolute Shrinkage and Selection Operator
adLASSO-EMD	Adaptive LASSO based on Empirical Mode Decomposition
AIC	Akaike Information Criterion
App.1	First application
App.2	Second application
App.3	Third application
AR	Autoregressive
ARIMA	Autoregressive integrated moving average
BE	Backward Elimination
BIC	Bayes Information Criterion
CCOD	Cyclical Coordinate Descent Algorithm
CHN	China
CHN/USD	Daily closed exchange rates of the China against USD
CV	Cross-Validated
D	Dimension
DBN	Deep Belief Networks
D-CV	D-fold Cross-Validation
ELNET	Elastic Net
ELNET-EMD	Elastic Net Regression based on Empirical Mode Decomposition
EMD	Empirical Mode Decomposition
Exp.1	First experiment
Exp.2	Second experiment
FSR	Forward Stepwise Regression
FT	Fourier Transform
HHT	Hilbert-Huang Transform
HTA	Hilbert Transform Analysis
IC	Information Criterion
iid	Independent and Identically Distributed
IMF	Intrinsic Mode Function
JAP	Japan
JAP/USD	Daily closed exchange rates of the Japan against USD

LASSO	Least Absolute Shrinkage and Selection Operator
LASSO-EMD	LASSO Regression based on Empirical Mode Decomposition
LE	Local Extrema
MA	Moving average
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MASE	Mean Absolute Scaled Error
MCP	Minimax Concave Penalty
MCP-EMD	Minimax Concave Penalty based on Empirical Mode Decomposition
MSE	Mean Square Error
MYR	Malaysia
MYR/USD	Daily closed exchange rates of the Malaysia against USD
Num. of V.S.	Number of Variable Selections
OLS	Ordinary Least-Square
OLS-EMD	Ordinary Least-Square based on Empirical Mode Decomposition
PCM	Pearson's Correlation Matrix
PE	Prediction Error
RMSE	Root Mean Square Error
RR	Ridge Regression
RR-EMD	Ridge Regression based on Empirical Mode Decomposition
RSS	Residual Sum of Squares
S&P 500 index	Stock performance of 500 large companies posted on stock exchanges in the US
SCAD	Smoothly Clipped Absolute Deviation
SCAD-EMD	Smoothly Clipped Absolute Deviation based on Empirical Mode Decomposition
SR	Stepwise Regressions
SR-EMD	Stepwise Regression based on Empirical Mode Decomposition
SVR	Support Vector Regression
TAW	Taiwan
TAW/USD	Daily closed exchange rates of the Taiwan against USD
USD	United States of America Dollar
VIF	Variance Inflation Factor
WT	Wavelet Transform
ZC	Zero-Crossing

LIST OF SYMBOLS

$C_k(t)$	The k-th intrinsic mode function component
K	Number of intrinsic mode function components
$R(t)$	Residual components
$Num(LE)$	Number of the local extrema in the signal
$Num(ZC)$	Number of the zero-crossing in the signal
$m(t)$	Envelope mean
$u(t)$	Upper envelop
$l(t)$	Lower envelop
$x(t)$	Original signal
t	Time domain
$C_{K+1}(t)$	Residual component
q	Repetition indicator
$h_q(t)$	New function (IMF) component
$R_0(t)$	Original signal
SD_q	Standard deviation “stoppage criterion”
log	Logarithmic function
N	Length of the original signal
✓	Yes
×	No
y_i	Actual value at time-period i “the ith response variable”
β_0	Constant term (intercept)
x_{ij}	The j-th predictor variable of the i-th observation
β_j	Regression coefficient of the j-th predictor variable
ε_i	Random error term at time-period i

n	Sample size
p	Number of predictor variables
$\mathbf{y}$	Vector of the response variable
$\mathbf{X}$	Matrix of the predictor variables
$\boldsymbol{\beta}$	Vector of the regression coefficients
$\boldsymbol{\varepsilon}$	Vector of random observation errors
$\widehat{\boldsymbol{y}}$	Estimated model
$\in$	Belong to
R	Real number
$\widehat{\boldsymbol{\beta}}$	Vector of estimated regression coefficients
$\hat{y}_i$	Estimated value of the variable y_i at time-period i
$\boldsymbol{\varepsilon}^t$	Transpose of the vector of random observation errors
$\mathbf{X}^t$	Transpose of the predictor variables matrix
∂	Partial derivative
$(.)^{-1}$	Inverse function
$\|.\|$	Norm function
$\|.\|_2^2$	L_2-norm square
$det(.)$	Determent of the matrix
$\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}}$	Vector of estimated regression coefficients by the OLS method
$E(.)$	Expected mean
$\forall$	For all
$Var(.)$	Variance
$Cov(.)$	Covariance
$N(0, 1)$	Normal distribution of zero mean and unit variance
$\bar{y}$	Sample mean of the response variable y
$\bar{x}_j$	Sample mean of the j-th predictor variable x_j

S_y	Standard deviation of response variable y
S_{x_j}	Standard deviation of the j-th predictor variable x_j
$\mathbf{V}$	Vector of the response variable after standardized
$\mathbf{Z}$	Matrix of the predictor variables after standardized
$\widehat{\boldsymbol{\beta}}^{\mathrm{RR}}$	Vector of estimated regression coefficients by the RR method
$\|\boldsymbol{\beta}\|_2^2$	L_2-norm square of the vector regression coefficients $\boldsymbol{\beta}$
$RP_\lambda(\boldsymbol{\beta})$	Ridge penalty function
τ; λ	Tuning parameter
$RP_\lambda'(\boldsymbol{\beta})$	First differentiable of the ridge penalty function
$RP_\lambda''(\boldsymbol{\beta})$	Second differentiable of the ridge penalty function
$\rightarrow$	Converge to
∞	Infinity
$\mathbf{I}$	Identity matrix
$\widehat{\boldsymbol{\beta}}^{\mathrm{LASSO}}$	Vector of estimated regression coefficients by the LASSO method
$\|.\|$	Absolute value function
$\|\boldsymbol{\beta}\|_1$	L_1-norm of the vector regression coefficients $\boldsymbol{\beta}$
$LP_\lambda(\boldsymbol{\beta})$	LASSO penalty function
$\mathbf{r}_\mathrm{f}$	Partial residual
$\widehat{\mathbf{V}}$	Estimated model after standardized
$\mathbf{Z_j}$	The j-th predictor variable
$\mathbf{Z_f}$	All the other predictor variables in the model except the j-th predictor variable
$sign(\hat{\beta}_j)$	Sign of coefficient $\hat{\beta}_j$ is ± 1
$S(\,,\,)$	Soft-thresholding function
$\widehat{\boldsymbol{\beta}}^{\mathrm{adLASSO}}$	Vector of estimated regression coefficients by the adLASSO method
$AP_{\lambda,\ell}(\boldsymbol{\beta})$	adLASSO penalty function

ω_j	Adaptive data-driven weights of the *j*-th predictor
$\hat{\beta}_j^{init}$	Initial estimate of the regression coefficient β_j
ℓ	The positive constant value greater than zero
$\mathcal{A}$	The set of regression coefficients $\boldsymbol{\beta}$ which have an important index effect
$\mathcal{A}_n$	The set of $\boldsymbol{\beta}$ that is estimated using the adLASSO method
$\xrightarrow{d}$	Converge to distribution
$\boldsymbol{\Sigma}$	Covariance matrix
$\widehat{\boldsymbol{\beta}}^{\text{SCAD}}$	Vector of estimated regression coefficients by the SCAD method
$SP_{\lambda,\gamma}(\beta_j)$	SCAD non-convex penalty function
γ	Tuning parameter used to control the degree of concavity
$SP'_{\lambda,\gamma}(\beta_j)$	First differentiable of the SCAD non-convex penalty function
$T^{SCAD}(\,,\,)$	SCAD thresholding operator function
$\widehat{\boldsymbol{\beta}}^{\text{MCP}}$	Vector of estimated regression coefficients by the MCP method
$MP_{\lambda,\gamma}(\beta_j)$	MCP non-convex penalty function
$MP'_{\lambda,\gamma}(\beta_j)$	First differentiable of the MCP non-convex penalty function
$T^{MCP}(\,,\,)$	MCP thresholding operator function
L_2-penalty	Ridge regression
L_1-penalty	LASSO regression
$\widehat{\boldsymbol{\beta}}^{\text{ELNET}}$	Vector of estimated regression coefficients by the ELNET method
$EP_{\lambda}(\boldsymbol{\beta})$	ELNET penalty function
λ_1	Tuning parameter of LASSO method
λ_2	Tuning parameter of RR method
α	Regularization parameter
μ	Sample mean
ρ	Pearson's correlation

$Bias^2$	Bias squared term
S	Number of tuning parameters
λ_{opt}	Optimal tuning parameter
D	Number of folds
$x_j(t)$	The *j*-th original signal
$C_{j,\,k}(t)$	The *k*-th intrinsic mode function components of the *j*-th predictor
$R_j(t)$	Residual component of the *j*-th predictor
C_i, C_{ii}	Significant predictor variable
C_f	Insignificant predictor variable
λ_{opt}^{RR}	Optimal tuning parameter value for the RR method
λ_{opt}^{LASSSO}	Optimal tuning parameter value for the LASSO method
ω_k^{RR-EMD}	Adaptive weights of the *k*-th component based on the RR-EMD method
$\lambda_{opt}^{adLASSSO}$	Optimal tuning parameter value for the adLASSO method
λ_{opt}^{SCAD}	Optimal tuning parameter value for the SCAD method
λ_{opt}^{MCP}	Optimal tuning parameter value for the MCP method
α_{opt}	Optimal regularization parameter value
λ_{opt}^{ELNET}	Optimal tuning parameter value for the ELNET method
R^2	Coefficient of determination
y_{i-1}	The actual value from the prior time-period
min	Minimum of the *MSE*
$1se$	One-standard-error
λ_{min}	Lambda at the minimum of the *MSE*
λ_{1se}	Lambda at minimum of the *MSE* with one standard error

LIST OF APPENDICES

MENINGKATKAN PEMILIHAN MODEL BERDASARKAN KAEDAH REGRESI PENALTI DAN PENGHURAIAN MOD EMPIRIK

ABSTRAK

Dalam tesis ini, kaedah regularisasi penalti iaitu, penyimpangan mutlak dipotong terlicin (SCAD), penyusutan mutlak terkecil dan operator pemilihan tersuai (adLASSO), penalti cekung minimax (MCP), dan kaedah regresi jaringan Elastik (ELNET) telah digunakan. Kaedah tersebut digabungkan dengan bahagian pertama jelmaan Hilbert-Huang, iaitu algoritma penghuraian mod empirik (EMD) secara berasingan. Algoritma EMD digunakan untuk menguraikan set data siri masa tidak pegun dan tak linear menjadi satu set komponen penguraian ortogonal yang terhingga yang mana merangkumi sekumpulan fungsi mod intrinsik (IMF) dan komponen reja. Komponen-komponen ini telah digunakan dalam beberapa kajian sebagai pemboleh ubah peramal baru untuk meramalkan tingkah laku pemboleh ubah sambutan. Kaedah regularisasi penalti adalah teknik statistik yang digunakan untuk mengatur dan memilih pemboleh ubah peramal yang diperlukan yang mempunyai pengaruh besar pada pemboleh ubah sambutan, untuk menghasilkan model yang konsisten dari segi pemilihan pemboleh ubah dan penganggaran normal asimptot, dan untuk mengatasi masalah multikekolinearan apabila wujud antara pemboleh ubah peramal. Objektif utama kajian ini adalah untuk menerapkan kaedah SCAD-EMD, adLASSO-EMD, MCP-EMD dan ELNET-EMD yang dicadangkan untuk menentukan kesan komponen penguraian pemboleh ubah peramal siri masa univariat / multivariat yang asal pada pemboleh ubah sambutan dan mengatasi multikekolinearan antara komponen penguraian bagi meningkatkan ketepatan ramalan untuk membina model yang sesuai. Teknik yang dicadangkan dibandingkan dengan empat kaedah regresi tradisional yang

digunakan dalam kajian sebelumnya. Dua jenis aplikasi, iaitu, eksperimen berangka dan set data sebenar yang melibatkan masalah tak pegun dan tidak linear telah diterapkan. Dua contoh eksperimen berangka menggunakan pemboleh ubah peramal asal univariat dan multivariat oleh fungsi gelombang sinus digunakan. Manakala set data sebenar yang diterapkan dalam contoh ilustrasi, merangkumi kadar pertukaran harian dan pasaran saham harian untuk beberapa negara. Dapatan berdasarkan pengukuran ralat menunjukkan bahawa kaedah yang dicadangkan mengatasi kaedah persaingan lain dalam eksperimen berangka dan aplikasi set data sebenar. Terutama, kaedah ELNET-EMD yang dicadangkan mempunyai keupayaan untuk mengenal pasti komponen penguraian yang mempunyai keertian paling besar pada pemboleh ubah sambutan walaupun terdapat korelasi tinggi antara komponen penguraian dengan ketepatan ramalan yang tinggi. Manakala kaedah adLASSO-EMD, SCAD-EMD, dan MCP-EMD yang dicadangkan mempunyai keupayaan untuk menghasilkan model yang konsisten dengan mengurangkan ralat ramalan berbanding dengan kaedah tradisional. Algoritma EMD menjadikan hubungan antara pemboleh ubah lebih dipercayai dalam domain masa dan frekuensi secara serentak.

ENHANCING MODEL SELECTION BASED ON PENALIZED REGRESSION METHODS AND EMPIRICAL MODE DECOMPOSITION

ABSTRACT

In this study, the penalized regularization methods, namely, the smoothly clipped absolute deviation (SCAD), adaptive least absolute shrinkage and selection operator (adLASSO) regression, minimax concave penalty (MCP) and elastic net (ELNET) regression, are adopted. Those methods are combined with the first part of the Hilbert–Huang transformation, namely, the empirical mode decomposition (EMD) algorithm. The EMD algorithm is employed to decompose the nonstationary and nonlinear time series dataset into a finite set of orthogonal decomposition components, which includes a set of intrinsic mode function and residual components. These components have been used in several studies as new predictor variables to predict the behaviour of the response variable. The penalized regularization methods are statistical techniques used to regularize and select the necessary predictor variables that have substantial effects on the response variable. These methods are also utilized to produce a consistent model in terms of variable selection and asymptotically normal estimates and address the multicollinearity problem when it exists between the predictor variables. This study aims to apply the proposed SCAD-EMD, adLASSO-EMD, MCP-EMD and ELNET-EMD methods to determine the effect of the decomposition components of the original univariate/multivariate time series predictor variable(s) on the response variable. Moreover, this study tackles the multicollinearity between the decomposition components to enhance the prediction accuracy for creating a fitting model. The proposed techniques are compared with four traditional regression methods employed in the previous study. Two types of applications, namely,

numerical experiments and actual real datasets that involve nonstationary and nonlinear problems, are applied. The two numerical experiment examples using the univariate and multivariate original predictor variable(s) by the sine wave function is used. The actual real datasets are applied in illustrative examples, which include the daily exchange rates and daily stock market for several countries. Results based on the error measures show that the proposed methods outperform the other competitive methods in the numerical experiment and actual dataset applications. The proposed ELNET-EMD method can identify the decomposition components that have a great significance on the response variable despite the high correlation between the decomposition components with high prediction accuracy. By contrast, the proposed adLASSO-EMD, SCAD-EMD and MCP-EMD methods can produce a consistent model with less prediction error compared with those of the traditional methods. The EMD algorithm makes the relationship between the variables reliable in terms of time and frequency domains.

CHAPTER 1

INTRODUCTION

1.1 Background and Motivation

The regression analysis is a robust statistical method for studying the relationship between the predictor and the response variables to obtain a fit model with significant predictor variables for improving the products and services in various fields. However, several challenges, such as the nature of time series variables used in the study, affect the work of regression analysis and prediction accuracy. These variables often belong to nonstationary time series datasets and the correlation may exist between two or more predictor variables; accordingly, the variance of the estimated coefficients increases and the estimated parameters are inaccurate (Cho *et al.*, 2010, Masselot *et al.*, 2018). A large number of predictor variables also exist in the model; consequently, such a model is difficult to interpret (Qin *et al.*, 2016). In regression analysis, these problems lead to bias in selecting fit models and can mislead the conclusions of the studies (Hoover, 2003, Masselot *et al.*, 2018).

Most real-life data appear as nonstationary and nonlinear time series datasets. The decomposition of nonstationary and nonlinear time series is an important issue that needs to be considered in conducting an analysis. Analytical methodologies employed with these time series datasets are scarce. However, several traditional decomposition methods (such as Fourier decomposition by Titchmarsh (1948) and wavelet decomposition method by Chan (1994)) and modification methods (such as the differentiation method) can be applied. Nevertheless, these methods lead to several problems, such as the loss of valuable information (Parsons *et al.*, 2000, Huang, 2014, Bendjama and Boucherit, 2016).

Huang *et al.* in 1998 studied the distorted nonstationary and nonlinear time series signals without moving from time domain to frequency domain. Where the decomposition components extracted via the decomposing method keep the information in the time domain. This approach led to a new technical method called the empirical mode decomposition (EMD) proposed by Huang *et al.* (1998). The EMD method does not require a precondition for the time series dataset, unlike the traditional Fourier and wavelet decomposition methods (Huang, 2014). The practical principle of the EMD aims to separate the nonstationary and nonlinear signal into a finite set of orthogonal nonoverlapping timescale components, namely, intrinsic mode function (IMF) and residual components (Huang, 2014). Each of these components includes particular information frequencies found within the original time series dataset (Huang *et al.*, 1998, Moore *et al.*, 2018, Liu and Chen, 2019). Therefore, these decomposition components can be used as new predictor variables to predict their effects and behaviors about another response variable by using the suitable regression methods (Qin *et al.*, 2016, Masselot *et al.*, 2018, Adarsh and Janga Reddy, 2019).

In regression analysis, the ordinary least square (OLS) regression estimation has several drawbacks, such as the presence of multicollinearity between the predictor variables, low prediction accuracy and difficulty in reducing the number of predictor variables (Jadhav *et al.*, 2014). Few statistical methods can deal with reducing the number of predictors in the model when multicollinearity exists. Consequently, many researchers continue in developing hybrid regression models to deal with these issues by improving the OLS method (Montgomery *et al.*, 2012), such as the stepwise regression (SR) and the penalized regularization method, namely, the ridge regression (RR) (Hoerl and Kennard, 1970) and the least absolute shrinkage and selection operator (LASSO) regression (Tibshirani, 1996).

However, the results obtained by the SR method lacked the reliability in selecting the optimal model (Smith, 2018). Meanwhile, the RR method still cannot deal with the reduction of the predictor numbers; hence, the unnecessary predictor variables will still exist in the final model (Tibshirani, 1996, Zou and Hastie, 2005). The LASSO method is inconsistent for variable selection and dealing with multicollinearity (Fan and Li, 2001, Zou and Hastie, 2005, Zou, 2006). Therefore, the smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001), adaptive LASSO (adLASSO) (Zou, 2006), minimax concave penalty (MCP) (Zhang, 2010) and elastic net (ELNET) methods (Zou and Hastie, 2005) were proposed. These four methods represent a new development penalized regularization method for improving the model interpretability and identifying relevant variables.

The behaviour of the variables in regression analysis, such as nonstationary, multicollinearity problem and the large number of the predictors in the model may affect the prediction accuracy in model selection. This situation indicates several difficulties in the regression analysis with these existing issues, such as determining the necessary predictors and improving the prediction accuracy of the model selection. These issues are significant in model selection; thus, researchers and decision-makers must consider them. The EMD method and the new penalized regularization methods, namely, adLASSO, SCAD, MCP and ELNET regression analysis, are proposed to address these gaps to enhance the model selection.

1.2 Problem Statement

In the time-series dataset case, the regression analysis assumes restrictions on all the variables before estimation to achieve the reliability and prediction accuracy of the model selection. At present, the following four major problems exist:

1) In the time series dataset, the variable of the time series dataset is assumed to be linear. Specifically, the relationship between the observations of the time series is linear (Moore *et al.*, 2018). However, the linearity assumption of the real-life variable appears as a nonlinear time series dataset.

2) In the regression analysis, the variable of time series datasets is assumed to be stationary. Specifically, the properties of the dataset, such as mean, variance and autocorrelation, do not depend on the time when the series is observed (Moore *et al.*, 2018, Liu and Chen, 2019, Adarsh and Janga Reddy, 2019). The real-life data mostly appear as nonstationary time series datasets. Several methods for converting and modifying these datasets are applied. However, these methods lead to the loss of valuable information and features of the original dataset. Moreover, the traditional decomposition methods are not highly efficient, thereby affecting the accuracy of the results (Parsons *et al.*, 2000, Huang, 2014, Bendjama and Boucherit, 2016).

3) The predictor variables are assumed to be free from multicollinearity. Multicollinearity occurs when two or more predictor variables contain the same information. In mathematical terms, a high correlation exists between the predictor variables when the matrix of predictors $\mathbf{X}^t\mathbf{X}$ is close to singular (i.e. the determinant of the matrix is close to the zero value) (Cho *et al.*, 2010, Jadhav *et al.*, 2014) . This phenomenon will have an effect on the variance, an estimate of the parameters and a wrong sign of coefficients. These phenomena mislead the fit model selection (Alin, 2010, Daoud, 2017). Several ways are used to solve multicollinearity, such as dropping one or more predictors, which have been highly correlated from the model. However, these predictors are

often important. This situation leads to bias in the estimation (El-Dereny and Rashwan, 2011, Daoud, 2017).

4) Variable selection or reduction of the number of predictors: The existence of a large number of predictor variables in the model leads to an overfitting case. Theoretically, a minimum of ten predictors in the regression model can cause an overfitting case. This situation leads to the difficulty in determining the optimal variables, which have an effect on the response variable and interpreting the final model and bad prediction accuracy (Jadhav *et al*., 2014). Several traditional methods, such as the SR method, are used. However, this method lacks reliability because of the dropping and retaining of the predictor variables, thereby affecting the model selection (Smith, 2018).

These problems motivated this study to develop new four hybrid penalized regression methods. The proposed methods that are developed in this study will overcome the problems with existing regression models, thereby further improving the prediction accuracy of the model selection.

1.3 Research Objectives

This study aims to improve the accuracy of the model selection for nonstationary time series datasets on the basis of the EMD method and the newly penalized regression approaches, namely, SCAD, adLASSO, MCP and ELNET, with the following objectives:

1) To decompose the nonstationary and nonlinear predictors into a finite set of components while maintaining its property using EMD algorithm so that it is estimable.

2) To propose and develop four penalty regression methods (SCAD, adLASSO, MCP, ELNET) based on the EMD method to enhance the prediction accuracy of the model selection.

3) To demonstrate that the proposed penalty regression methods using the extracted EMD data with and without multicollinearity outperform the traditional regression methods.

1.4 Scope of the Study

This study aims to select the necessary decomposition components via the EMD method of the nonstationary and nonlinear time series predictors on the response variable. Moreover, this study aims to tackle the multicollinearity problem that exists between the decomposition components and enhance the prediction accuracy. Four new regression technique methods are examined: (1) adLASSO regression between the response variable and all the decomposed components via the EMD of the original predictor(s) (adLASSO-EMD); (2) SCAD method between the response variable and all the decomposed components via the EMD of the original predictor(s) (SCAD-EMD); (3) MCP method between the response variable and all the decomposed components via the EMD of the original predictor(s) (MCP-EMD); and (4) ELNET regression between the response variable and all the decomposed components via the EMD of the original predictor(s) (ELNET-EMD).

Numerical experiments and actual real datasets are applied to provide a convenient framework for achieving the objectives of this study. The numerical experiments are represented by two experiments: the first experiment is by using the univariate original predictor; the second experiment utilizes the multivariate original predictors. The two experiments are simulated by using the sine wave function. The

actual real datasets are applied in three economic time series applications: the first application is on the daily exchange rates and stock market of Malaysia. The second and third applications are on the daily stock market and the daily exchange rates for several countries.

1.5 Limitation of the Study

In this study, the proposed methods aim to understand the relationship between the decomposition components and the response variable from a different point of view, that is, a multicollinearity problem. These methods also aim to reduce the number of decomposition components in the model to enhance the production accuracy. However, the time series data often have autocorrelation problems, such as the degree of similarity among the current value and its prior values. If such problem occurs, then they are another area of research problems. Therefore, these problems will be considered in future studies.

1.6 Significance of the Study

This study aims to apply the proposed four penalized regression technique methods on the basis of the EMD methods, namely, adLASSO-EMD, SCAD-EMD, MCP-EMD and ELNET-EMD, to enhance the model selection with high production accuracy. The numerical experiments and application results that involve nonstationary and nonlinear problems show that the proposed methods can efficiently identify the decomposition components that have the largest influence on the response variable with high prediction accuracy compared with the traditional methods. The proposed ELNET-EMD method has achieved optimal results. In terms of time–frequency domain, the result showed that the proposed methods are reliable and accurate in identifying the main component(s) via the EMD algorithm of the original

predictor that has a great effect on the response variable. The EMD algorithm makes the relationship between the variables reliable in terms of the time and frequency domains. Accordingly, the study of the relationship between the variables becomes possible in this sense.

1.7 Thesis Organization

This thesis is organized as follows: Chapter 2 presents a brief review on the related literature of the decomposition methods and a description of the EMD algorithm. This chapter also provides a description of the various penalized regression techniques that are either used or compared with the proposed regression methods, standardization and multicollinearity problem. The choosing of the optimal tuning parameter also presents in this chapter.

Chapter 3 presents the proposed regression methods, namely, the hybrid of adLASSO-EMD, SCAD-EMD, MCP-EMD and ELNET-EMD, and the current regression methods, such as OLS-EMD, SR-EMD, RR-EMD and LASSO-EMD. The multicollinearity test and the test criteria are also presented in Chapter 3.

Chapter 4 applies and discusses the proposed methods through numerical experiments by sine wave function and the analytical results. Chapter 5 illustrates the proposed methods by applying the actual real time series dataset in three applications, provides analysis and discusses the findings. Chapter 6 elaborates on the conclusions of the main results in the thesis, contribution and findings of the study and suggestion for future work.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter contains nine sections. The second section offers a brief explanation of the signal decomposition methods by Hilbert-Huang transform. The third section reviews the description and overview of the empirical mode decomposition (EMD) method, such as the developments and applications in different areas. A list of the recent studies that apply EMD in established statistical regression methods is also presented at the end of this section. The fourth section describes the ordinary least-square regression method. The fifth section presents the standardization data. The sixth section describes the various penalized regression techniques that are either used or compared with the proposed regression methods in this study. These techniques include the ridge regression, the least absolute shrinkage and selection operator regression, the adaptive least absolute shrinkage and selection operator regression, the smoothly clipped absolute deviation, the minimax concave penalty and the elastic net regression. The seventh section describes the multicollinearity problem. The eighth section dispenses the choosing of the tuning parameter and highlights the *D*-fold cross-validation methods. The last section is the summary.

2.2 Time Series Decomposition based on Hilbert-Huang Transform

In various fields of studies, such as medicine, physics, economics and environmental science, most variables are time series datasets. These datasets are nonstationary and nonlinear. The decomposition of the nonstationary and nonlinear time series are important issues that need to be considered in conducting an analysis.

However, analytical methodologies employed to deal with these time series datasets are scarce.

The traditional decomposition methods assume that the time series dataset should be either stationary or linear before the analysis. These methods include the Fourier transform (FT) method proposed by Titchmarsh (1948) and wavelet transform (WT) method proposed by Chan (1994). The FT method is interested in the stationary and linear signals. However, this method is not suitable for nonstationary and nonlinearity time series signals (Bendjama and Boucherit, 2016). By comparison, the WT method is interested in the stationary/nonstationary and linear signals. However, this method is not suitable for nonlinearity signal analysis (Gröchenig, 2013, Bendjama and Boucherit, 2016). Many studies have been interested in developing methods for the signal decomposition. To date, the signal decomposition still receives great attention from many researchers.

Recently, researchers have been interested in distorted signals, which are represented as nonlinear and nonstationary signals. The traditional signal analysis provided minimal options on the analysis of signals. For instance, the signal should be either linear or stationary/nonstationary. Huang *et al.* in 1998 provided a new approach that can simultaneously deal with nonstationary and nonlinear signals without moving from the time domain to the frequency domain; the information will be maintained in the time domain; thus, this process is called the Hilbert-Huang transform (HHT) (Huang *et al.*, 1998). Such a process is different from the traditional methods.

The HHT includes two terms: EMD and Hilbert transform analysis (HTA) (Huang, 2014). The main principle to apply of the HHT depends on the concept of

EMD after this process ends. The results extracted from EMD will be applied to the HTA method; the EMD method represents the central idea of the HHT (Huang, 2014).

The practical principle of the EMD aims to separate the nonstationary and nonlinear signals into a finite set of orthogonal nonoverlapping time-scale signals or components. The EMD method has separated the signals in the time domain to ensure that the length of the decomposition components has the same length to that of the original signal. In the next step, the HTA method will be applied to all the decomposition components extracted by the EMD method to compute the instantaneous frequency data (Huang *et al.*, 1998, Huang, 2014). The next subsection will discuss the principle of the EMD algorithm analyses by applying the sifting process. This subsection will also describe the intrinsic mode function and its conditions, the applications of the EMD method and the regression analysis on the basis of EMD.

2.3 Empirical Mode Decomposition (EMD)

The EMD is a new analytical method proposed by Huang *et al.* in 1998, and it is the first part of HHT (Huang *et al.*, 1998). The EMD approach aims to decompose the nonstationary and nonlinear signal into a finite set of nearly orthogonal decomposition components, thereby keeping the time domain of the signal constant. These decomposition components are called the intrinsic mode function components and one residual component represents the trend of the original signal (Moore *et al.*, 2018, Bokde *et al.*, 2019). Each component is different in terms of its physical form (i.e. wavelength and frequency). Every component includes information about the frequencies found within the original time series dataset (Huang, 2014, Qin *et al.*, 2016, Liu and Chen, 2019). The EMD method is adaptive, intuitive and highly efficient

in dealing with nonstationary and nonlinear signals (Huang, 2014). The principle of the EMD analyzes the original signal via an iterative process called the sifting process (Huang, 2014, Awajan *et al.*, 2019, Liu and Chen, 2019).

In this section, the principle of the EMD algorithm, which analyzes by applying sifting process, description of the intrinsic mode function and its conditions, the applications of the EMD method and regression analysis on the basis of EMD, will be presented.

2.3.1 Intrinsic mode function

The intrinsic mode function (IMF) represents a new orthogonal design for the signals resulting from the division of the main original primary signal by using the EMD algorithm. The IMF $\{C_k(t); k = 1, 2, \ldots, K\}$ and residual $R(t)$ components have different and easy physical significant meanings (Huang *et al.*, 1998).

The IMF component function that satisfies two conditions (Huang, 2014, Awajan *et al.*, 2019, Bokde *et al.*, 2019) are as follows:

i. Over the whole length of a signal, the numbers of local extrema (maximum and minimum) and the number of zero-crossings (ZCs) must be either equal or differ at most by one:

$$|Num(LE) - Num(ZC)| \leq 1, \tag{2.1}$$

where $Num(LE)$ represents the number of local extrema (LE) and $Num(ZC)$ represents the number of ZCs.

ii. At any point on a signal, the envelop mean value between the upper envelop defined by the local maximum and the lower envelop defined by local minimum is equal to zero:

$$m(t) = \frac{u(t) + l(t)}{2} = 0, \tag{2.2}$$

where $m(t)$ is the mean envelop, $u(t)$ is the upper envelop and $l(t)$ is the lower envelop.

The first condition indicates that each IMF has only one local maximum or local minimum between two consecutive ZCs or vice versa. The second condition implies that all the IMFs are stationary, which makes the analytical process highly flexible in using these components (Raghuram *et al.*, 2012, Huang, 2014, Awajan *et al.*, 2019).

2.3.2 Sifting process

The iterative algorithm process of the EMD method for extracting all the IMF components and one residual component is called sifting process (Huang, 2014). The sifting process separates the original signal into a finite set of the orthogonal decomposition components of a nonoverlapping timescale (Ur Rehman *et al.*, 2013, Bokde *et al.*, 2019). Thus, the original signal is the linear combination of the finite set IMF components and one residual component that is extracted via EMD can be constructed back as in Equation (2.3).

$$x(t) = \sum_{k=1}^{K} C_k(t) + R(t), \tag{2.3}$$

where $x(t)$ is the original signal; K is the number of IMF components; $C_k(t)$ represents the *k*-th IMF $\{k = 1, 2, \ldots, K\}$; and $R(t)$ represents the final component, which is the residual or trend of the original signal. The residual component has been seen as the K+1 IMF, which means the $C_{K+1}(t)$ component.

The sifting process to decompose the original signal will be presented in six steps as follows (Huang *et al.*, 1998, Huang, 2014, Awajan *et al.*, 2019):

Step 1: Insert the original signal $x(t)$ for the sifting assuming that the repetition indicator value is equal to one ($q = 1$, $k = 1$).

Step 2: Identify all the local extremum value (local maximum and local minimum) of the original signal $x(t)$. An illustrative example is shown in Figure 2.1.

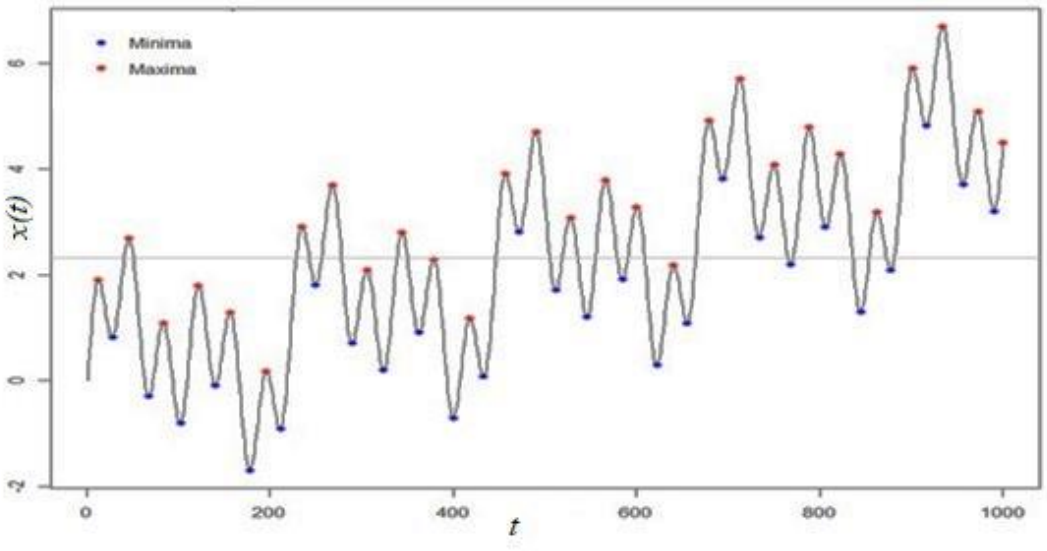

Figure 2.1 Local extreme (maxima and minima) of the original signal $x(t)$

Figure 2.1 explains an example of step 2. The black line represents the original signal $x(t)$, whereas the red circle point on the upper line represents the local maximum. By contrast, the blue circle point on the lower line represents the local minimum.

Step 3: All local maximum and minimum are separately connected to create an upper envelop $u_q(t)$ and a lower envelop $l_q(t)$, respectively, by using the cubic spline curve. The original signal must be between these envelops. An illustrative example is shown in Figure 2.2.

Step 4: The mean envelop $m_q(t)$ value between the upper and lower envelops is determined to create a new line curve, which represents the mean envelop by using Equation (2.4). An illustrative example is shown in Figure 2.2.

$$m_q(t) = \frac{u_q(t) + l_q(t)}{2}. \tag{2.4}$$

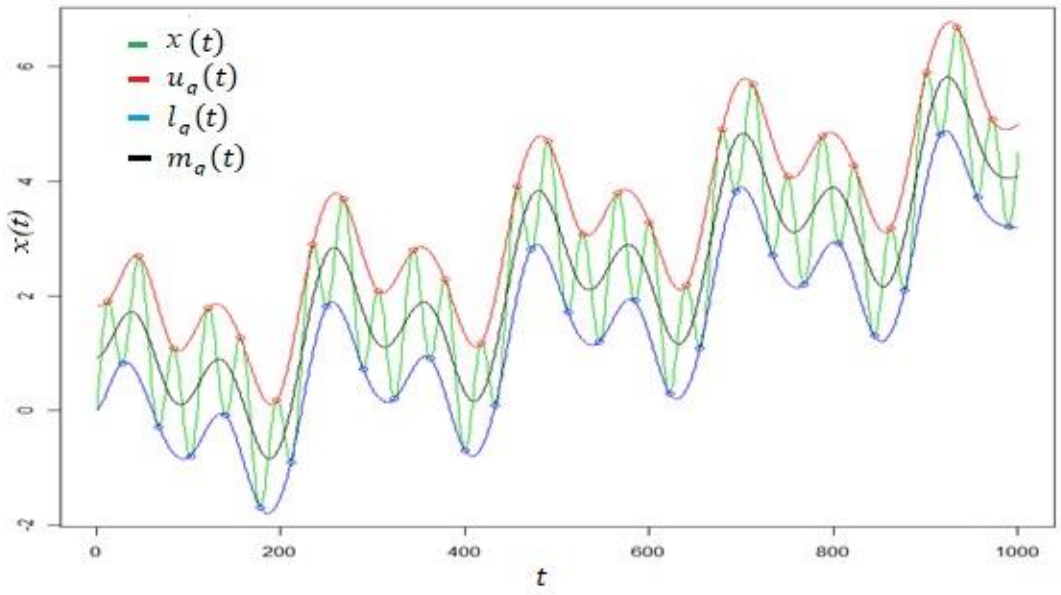

Figure 2.2 Upper and lower envelops of the original signal $x(t)$

Figure 2.2 explains an example of steps 3 and 4. The green line represents the original signal $x(t)$, the red line represents the upper envelop $u(t)$ and the blue line represents the lower envelop $l(t)$, which is explained by step 3. The black line is the mean envelop $m(t)$ between the upper and lower envelops, which is explained by step 4.

Step 5: Calculate the new function $h_q(t)$ component, which is the difference between the original signal $x(t)$ and the mean envelop $m_q(t)$ value, as shown in the following equation:

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

Check if the new function $h_q(t)$ satisfies the conditions of IMF (Section 2.3.1).

YES: Then $h_q(t) = C_k(t)$, where $C_k(t)$ is the k-th IMF $\{k = 1, 2, \ldots, K\}$; thereafter, save the $C_k(t)$ and continue to step 6.

NO: Then, we replace $h_q(t)$ with $x(t)$ and repeat the operation from step 2 with repetition indicator $q = q + 1$.

Step 6: Calculate the residual component $R_k(t)$ as in the following formula:

$$R_k(t) = R_{k-1}(t) - C_k(t); \quad R_0(t) = x(t). \tag{2.6}$$

Check if the residual component $R_k(t)$ is a monotonic or constant function that cannot extract additional IMF components or satisfies the stoppage criterion of the standard deviation (SD_q) for two consecutive successive siftings of the results (Huang, 2014), as shown in the following equation:

$$SD_q = \sum_{t=0}^{T} \frac{\left(h_{q-1}(t) - h_q(t)\right)^2}{{h_{q-1}}^2(t)}; \quad 0.2 \leq SD_q \leq 0.3. \tag{2.7}$$

YES: Save all $C_k(t)$; $k = 1, 2, \ldots, K$ and $R_k(t)$ components and then end the sifting.

NO: Replace $R_k(t)$ with $x(t)$ and then repeat the operations from step 2 on residue $R_k(t)$ with $k = k + 1$ until $R_k(t)$ has a monotonic function or satisfies stoppage criterion SD_q.

The sifting to extract of the IMF and residual components are summarized by the graphical plots by using the tree graph of the EMD algorithm in Figure 2.3 and the illustrative example in Figure 2.4 as follows:

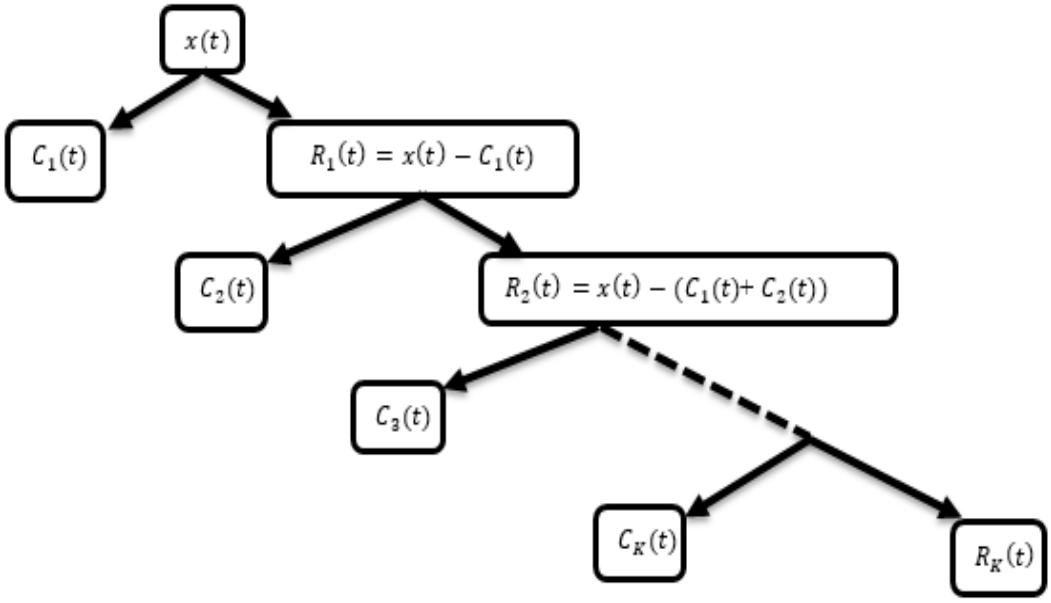

Figure 2.3 Tree graph of the EMD

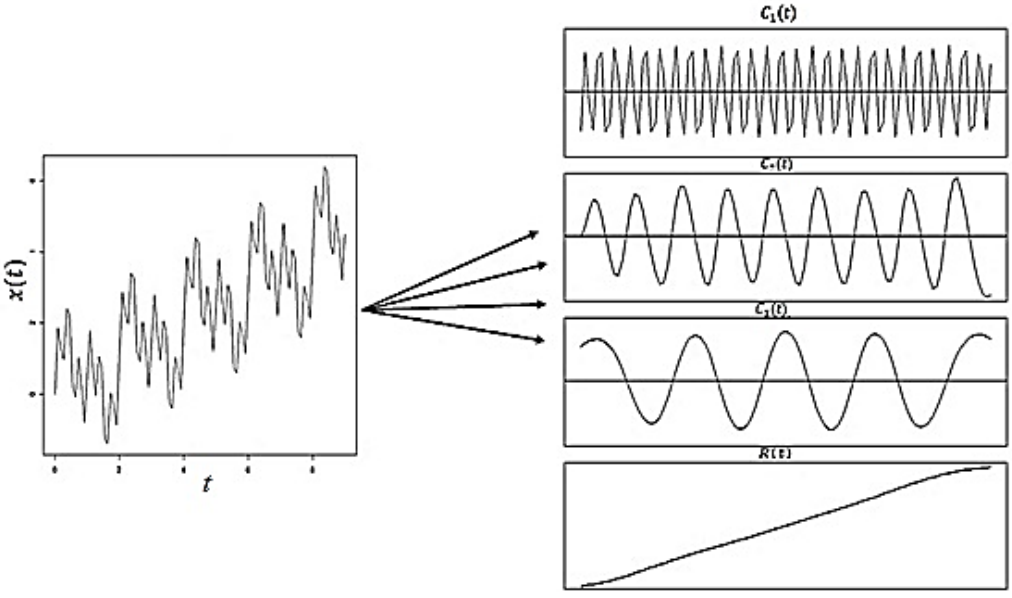

Figure 2.4 Decomposition of the original signal $x(t)$ via EMD

2.3.3 Applications, limitations and extensions of the EMD

The EMD has been widely used in various fields of science, such as medicine (Kanoga and Mitsukura, 2017), 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), economics (Jaber *et al.*, 2014, Hossain and Ismail, 2020), and environmental science (Basha *et al.*, 2015, Naik *et al.*, 2018).

However, the EMD method has some limitations in its application. For instance, during sifting process, most steps are not mathematically established (Awajan *et al.*, 2018, Liu and Chen, 2019) and have no existing theory (Flandrin *et al.*, 2004). The second limitation is related to mode mixing (Huang *et al.*, 2003). Many studies have proposed theoretical assumptions for this method, which are shown in the definition of the EMD method as algorithmic steps. Kizhner *et al.* (2006) suggested several theoretical elements for the method by proposing three hypotheses on the sifting. Wu *et al.* (2001) proved that the number of IMF components that are extracted from a signal is approximately equal to $log_2(N)$, where N is the length of the signal. Amirthanathan *et al.* (2005) showed that the average period of each IMF can be calculated by this formula $[\frac{2N}{Num(ZC)}]$ by using the number of ZCs and length of the signal. Wu *et al.* (2010) solved the limiation of mode mixing by increasing the amount of iteration process with the additional mathematical operators. Li *et al.* (2017) added the differential operation into the decomposition of the IMF components to solve mode mixing. The proposed methodologies in this study can overcome these limitations with the basic EMD method.

Many researchers provided an extension of the EMD algorithm to flexibly use it in various application areas. Xu *et al.* (2006) presented a two-dimensions (2D) EMD with finite elements for data analysis. Wu and Huang (2009) proposed the ensemble EMD by the added ensemble of white noise to the signal. Rehman and Mandic (2010)

presented the multivariate EMD. Torres *et al.* (2011) proposed the complete ensemble EMD by adding the particular noise at each step of decomposition. Dragomiretskiy and Zosso (2013) suggested the variational mode decomposition. He *et al.* (2017) proposed the 3D EMD. Zhang *et al.* (2017) proposed the noise-assisted multivariate EMD.

Several studies compared the EMD algorithm and other technical decomposition methods. The study results show that the EMD algorithm exhibited a high accuracy in dealing with nonstationary and nonlinear signals in various fields. Wang *et al.* (2011) compared the EMD algorithm and WT method by using nonstationary and nonlinear time series data. Lu *et al.* (2013) compared the EMD algorithm and the chirplet signal decomposition for ultrasonic imaging. Ghosh *et al.* (2014) compared the EMD with the FT method, short term FT method and WT used to denoise an electrocardiogram signal.

A comparative summary among the EMD, WT and FT methods for decomposing the signals (Huang, 2014) is illustrated in Table 2.1. This table explains that the performance of the EMD method does not require pre-conditions for the dataset unlike the FT and WT methods.

Table 2.1 Comparative of decomposition methods

Method	Linear	Nonlinear	Stationary	Nonstationary
FT	✓	×	✓	×
WT	✓	×	✓	✓
EMD	✓	✓	✓	✓

2.3.4 Statistical regression methods based on EMD

Recent studies have focused on using the EMD method combined with other established statistical regression or forecasting methods. This method has been successfully applied in several scientific fields. The decomposition components can

be used as new predictor variables to predict their effects and behaviors about other response variables by using suitable models or in different statistical situations, such as forecasting studies. For an illustrative example:

Yang *et al.* (2011a) applied the ordinary least-squares (OLS) regression analysis and forward stepwise regression (SR) methods on the basis of the EMD method. They studied the temporal associations between headache incidence (response) variable and decomposition components (predictor variables) via EMD of the weather variables, which are the pressure, temperature, humidity, sunshine duration and maximal wind speed. In this study, the residual (trend) component was removed to avoid the multicollinearity problem. Yang *et al.* (2011b) used the same methodology to study the effect of the set of predictor variables, namely, the pool of air pollution, unemployment and weather, on the decomposition components via EMD of the response variable, such as suicide.

Shen and Lee (2012) studied the least absolute shrinkage and selection operator (LASSO) regression on the basis of the ensemble EMD studied to reduce the effect of outliers on the ultrasound imaging for the blood flow velocity dataset. Shen *et al.* (2012) applied ridge regression (RR) on the basis of the ensemble EMD to achieve less decomposition error and solve the mode mixing problem. Kopsinis and McLaughlin (2008) used the smoothly clipped absolute deviation (SCAD) thresholding rule to enhance the denoising performance in the decomposition components via the EMD method.

Qin *et al.* (2016) applied the LASSO regression on the basis of the decomposition components via the EMD method of the univariate original predictor applied to choose the decomposed components that have most effect on the response

variable. Their method was compared with the OLS and RR methods on the basis of EMD. Numerical experiments and applications on the two Chinese stock markets (Shanghai Composite Index and Shenzhen Component Index) were applied.

Hu and Si (2013) and Zhao *et al.* (2018) employed SRs on the basis of multivariate EMD to predict and explore the relationship among the decomposition components between soil water and the environmental factors. Adarsh *et al.* (2018) applied SR on the basis of EMD and multivariate EMD to estimate the relationship between the decomposition components of the response variable reference evapotranspiration with decomposition components of the four predictor variables, namely, solar radiation, air temperature, relative humidity and wind velocity. Adarsh (2016) and Adarsh and Janga Reddy (2019) used the same methodology with different variables.

Naik *et al.* (2018) presented the kernel RR on the basis of the decomposition components via EMD method to study the significance of the decomposition components for short-term wind speed and wind power prediction on wind farms in Wyoming State of western United States. The multivariate EMD model with the kernel RR is designed to achieve significantly accurate drought forecasts of the three different agricultural sites in Pakistan (Ali *et al.*, 2019).

Masselot *et al.* (2018) proposed the LASSO regression based on noise-assisted multivariate EMD to select the decomposition components that have significant effects on the response variable/variables in two cases. The first case is to decompose the predictor variables only via EMD, while the second case involves decomposing the predictors and response variable via EMD. Their proposed method is applied by using

the multiscale and nonstationary weather variables and cardiovascular mortality in Canada “Montreal”.

In the forecasting field, Lee and Ouarda (2010) proposed the EMD with K-nearest neighbor resampling and block bootstrapping are combined to predict climatic oscillations. Chen *et al.* (2012) proposed the backpropagation neural network combined with the EMD model to predict the tourism demand. Ren *et al.* (2014) applied the EMD method and the versions of EMD (i.e. ensemble EMD and complete ensemble EMD) with the support vector regression (SVR) and artificial neural network for forecasting the nonstationary and nonlinear wind speed. Zhu *et al.* (2017) proposed the EMD method on the basis of the least-squares SVR to forecast the nonstationary and nonlinear carbon price. Nava *et al.* (2018) proposed the EMD method with SVR to improve the forecasting financial data by using the stock performance of 500 large companies posted on stock exchanges in the US (S&P 500 index).

Chu *et al.* (2018) proposed the LASSO regression and deep belief networks (DBN) on the basis of the ensemble EMD method to study the relationship between multiscale climatic predictors and the decomposition components of nonstationary and nonlinear monthly streamflow on the Tennessee River. This task was achieved by applying the LASSO regression to select the predictors and used DBN to create a forecasting model.

2.4 Ordinary Least-Square Regression (OLS)

This study considers a model structure of the general linear regression model, which creates the relationship between the response and the predictor variables and is derived as follows:

$$y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{ip}\beta_p + \varepsilon_i.$$

Then,

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i, \tag{2.8}$$

where i =1, 2, …, n; and j = 1, 2, …, p is the number of predictor variables, where y_i is the *i*-th response variable; β_0 is a constant term that represents the intercept; x_{ij} is *j*-th predictor variable of the *i*-th observation (i.e. the *i*-th level of the *j*-th predictor variable); β_j is the regression coefficient of the *j*-th predictor variable, which represents the average effect on y_i of per one unit change in *j*-th predictor variable x_{ij}; and ε_i is the random error term (Montgomery *et al.*, 2012, Melkumova and Shatskikh, 2017). Equation (2.8) can be written in matrix form as follows:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.9}$$

where $x_0 = 1_n$, $\mathbf{y} \in R^n$ is a vector of the response variable, $\mathbf{X} \in R^{n\times(p+1)}$ is a matrix of the predictor variables, $\boldsymbol{\beta} \in R^{(p+1)}$ is a vector of the regression coefficients and $\boldsymbol{\varepsilon} \in R^n$ is a vector of random observation errors.

The regression analysis aims to estimate the vector of the regression coefficient $\boldsymbol{\beta}$ for estimating the effect of the predictor variables $\mathbf{X}$ on the response variable $\mathbf{y}$. The vector of regression coefficient $\boldsymbol{\beta}$ in Equation (2.9) is unknown. Accordingly, the traditional ordinary least-square (OLS) is the commonly used estimation method for estimating unknown regression coefficients because of its simplicity. This method can

be applied without any distributional. The OLS estimator is unbiased and has a low variance (James *et al.*, 2013, Efron and Hastie, 2016). The OLS method aims to minimize the residual sum of squares (*RSS*). Thus, the estimated model $\hat{\mathbf{y}}$ for the true model $\mathbf{y}$ in Equation (2.9) is derived as follows:

$$\hat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}. \tag{2.10}$$

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

$$RSS = \boldsymbol{\varepsilon}^t\boldsymbol{\varepsilon} = \left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)^t\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)$$

Then,

$$RSS = \left\|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right\|_2^2 = \sum_{i=1}^{n}\left(y_i - x_{ij}\hat{\beta}_j\right)^2, \tag{2.11}$$

where $\|\,.\,\|_2^2$ is called L_2-norm square. We differentiate *RSS* Equation (2.11) with respect to the unknown parameters $\boldsymbol{\beta}$ and equate the derivatives to zero that is:

$$2\mathbf{X}^t\mathbf{X}\widehat{\boldsymbol{\beta}} - 2\mathbf{X}^t\mathbf{y} = 0$$

$$\mathbf{X}^t\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}^t\mathbf{y}.$$

Then, the optimal solution is as follows:

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

where $\mathbf{X}^t$ represents the transpose of the predictor variables matrix $\mathbf{X}$ and $(\,.\,)^{-1}$ represents the inverse function. If the determinant $\mathbf{X}^t\mathbf{X}$ is nonzero (i.e. $det(\mathbf{X}^t\mathbf{X}) > 0$), then all rows and columns are linearly independent, the matrix is of full rank and a unique solution to the normal equation is obtained. Equation (2.12) represents the OLS regression of $\boldsymbol{\beta}$ (Montgomery *et al.*, 2012, James *et al.*, 2013).