MODIFIED NON-TRANSFORMED PRINCIPAL COMPONENT AND ADAPTIVE PENALIZED HIGH DIMENSION FOR GROUPING EFFECT OF STOCK MARKET PRICE

YUSRINA ANDU

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Doctor of Philosophy

Faculty of Science
Universiti Teknologi Malaysia

APRIL 2020

DEDICATION

This thesis is dedicated to my late father, my mother and mother-in-law as well as my husband and kids, for their never-ending love, patience, support and prayers.

ACKNOWLEDGEMENT

All praise to Allah S.W.T for giving me enormous strength and full blessing to complete my PhD study. First and foremost, my greatest appreciation goes out to my supervisor, Prof Dr. Muhammad Hisyam Lee whom always give me good advice, encouragement and also guidance along my PhD study. His vision, attention to details, curiosity in undertaking new knowledge and research passion has been an inspiration to me. Certainly, under his supervision, it has become an interesting PhD journey for me and the things that I have learned throughout this journey will surely benefit me for the rest of my life.

I absolutely must not forget the guidance, advice and time of my co-supervisor, Prof. Dr. Zakariya Yahya Algamal. His knowledge in advance statistics are much useful for my future career. I would also like to thank you Mr. Syed Mohd Anwar from CICT UTM-HPC for allowing me to access high performance computing that enables me to complete one part of my PhD study. To my fellow graduate, Tengku Salbiah, thank you for the friendship. As well as, a big thank you to Ministry of Education Malaysia and Universiti Malaysia Kelantan for funding my PhD study.

Last but not least, a special thanks to my mother, Mdm Mariana Hj Fadzil and mother-in-law, Hajah Inoon Binti Abu for their du'a. My sincere appreciation also extends to my beloved husband, Shazani bin Sarijan whom also struggle to finish his PhD together with myself. Alhamdulillah and thank you for his high level of patience, meaningful guidance, tenacious endurance as well as encouragement throughout our PhD study. To my twin children, Ahmad Ayden and Ahmad Alyan (5 years old), and my lovely daughter, Adelia Inara (3 years old), you are my real strength in my life!

ABSTRACT

Nonstationary time series is complex and difficult to be modelled. Many researchers resolved it by transforming it into stationary time series. However, loss of generality will occur which make its inference more difficult. To overcome this, therefore a modified non-transformed approach is proposed using generalized dynamic principal component on the nonstationary series. On the other hand, the selection of informative variables is more importance, especially when the number of explanatory variables is larger than the number of observations. This is pertinent in order to achieve a better model interpretation of the highly correlated variables. Thus, the penalized likelihood methods are mostly adapted since they are able to perform variable selection and model estimation concomitantly. Nevertheless, the scarceness in the consistency of variable selection, encouragement of grouping effects and robustness can be found in the majority of these methods. Therefore, to overcome these shortcomings, several improvements in the high dimensional penalized methods are proposed in this study. The performance of homogenous variable selection was improved using ordered homogeneity pursuit least absolute shrinkage and selection operator method. An initial weight which is distance correlation is proposed in the adaptive elastic net to encourage grouping effects between highly correlated variables in high dimension data. Furthermore, this proposed method also has the capability to improve the robustness in the regression model, especially when outliers are presence in the response variable or there is a heavy-tailed distribution in the error. Three algorithms were developed for the simulation of the modified non-transformed principal component and proposed adaptive penalized high dimension methods. In this study, the effectiveness of all the modified and proposed methods was examined through three simulation studies and also through the application of stock market price. It is known that the studies that perform variable selection, encouraging grouping effects and robustness in high dimensional stock market price using the statistical approach is still scarce. In conclusion, the modified and proposed high dimensional penalized methods provide much better performance results both for the simulation and real data application as compared to their counterpart.

ABSTRAK

Siri masa tak pegun adalah kompleks dan sukar untuk dimodelkan. Kebanyakkan penyelidik menyelesaikannya dengan melakukan penjelmaaan ke atas siri masa pegun. Walaubagaimanapun kehilangan keaslian akan berlaku yang menjadikan kesimpulannya lebih sukar. Bagi mengatasi permasalahan ini, maka pendekatan pengubahan tanpa transformasi diusulkan menggunakan komponen utama dinamik teritlak terhadap siri tak pegun. Sementara itu, pemilihan pembolehubah yang informatif adalah lebih utama lebih-lebih lagi apabila bilangan pembolehubah penerang melebihi bilangan cerapan. Ini penting untuk memperolehi penjelasan model yang lebih baik daripada pembolehubah-pembolehubah yang berkorelasi tinggi. Maka, kaedah kebolehjadian terhukum seringkali diadaptasi kerana keupayaannya untuk melakukan pemilihan pembolehubah dan pengganggaran model secara serentak. Tidak kira apapun, kekurangan dalam konsistensi pemilihan pembolehubah, penggalakkan kesan pengelompokkan dan keteguhan boleh ditemui di dalam sebahagian besar kaedah ini. Oleh yang demikian, bagi mengatasi kekurangan yang dinyatakan ini, beberapa penambahbaikan di dalam kaedah terhukum berdimensi tinggi telah dicadangkan dalam kajian ini. Prestasi pemilihan pembolehubah yang homogen telah ditambahbaik menggunakan kaedah bertertib homogen mengejar pengecutan mutlak terkecil dan pemilihan operator. Pemberat awal iaitu korelasi jarak diusulkan di dalam elastik jaring mudah suai untuk menggalakkan kesan pengelompokkan di antara pembolehubah yang berkorelasi tinggi di dalam data yang berdimensi tinggi. Tambahan pula, kaedah yang diusulkan ini juga boleh menambahbaik keteguhan yang terdapat di dalam model regresi, terutamanya sama ada apabila wujud kehadiran pencilan dalam pembolehubah sambutan atau terdapat taburan hujung yang berat pada bahagian ralat. Tiga algoritma telah dibangunkan terhadap kaedah pengubahan tanpa transformasi komponen utama dan kaedah terhukum mudah suai berdimensi tinggi yang telah diusulkan. Di dalam kajian ini, keberkesanan semua kaedah yang diubahsuai dan diusulkan diperiksa melalui tiga kajian simulasi dan pengaplikasian menggunakan harga pasaran saham. Diketahui bahawa kajian yang menjalankan pemilihan pembolehubah, penggalakkan kesan pengelompokkan dan keteguhan bagi harga pasaran saham berdimensi tinggi menggunakan pendekatan statistik masih lagi kurang. Kesimpulannya, kaedah pengubahan dan kaedah terhukum berdimensi tinggi yang telah diusulkan menunjukkan keputusan yang lebih baik di dalam kedua-dua simulasi dan aplikasi data sebenar berbanding dengan kaedah-kaedah yang dibandingkan.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ADF	-	Augmented Dickey-Fuller
AEN	-	Adaptive Elastic Net
AER	-	Adaptive Elastic Net with Ridge
AEDC	-	Adaptive Elastic Net with Distance Correlation
AIC	-	Akaike Information Criteria
ARMA	-	Autoregressive Moving Average
BDPC	-	Brillinger Dynamic Principal Component
BIC	-	Bayesian Information Criteria
CV	-	Cross-validation
DC	-	Distance Correlation
DC-SIS	-	Distance Correlation with Sure Independence Screening
HPC	-	High Performance Computing
iid	-	Independently identically distributed
KPSS	-	Kwiatkowski-Phillips-Schmidt-Shin
LASSO	-	Least Absolute Shrinkage and Selection Operator
M-GDPC	-	Modified Generalized Dynamic Principal Component
MSE	-	Mean Squared Error
OHPL	-	Ordered Homogeneity Pursuit LASSO
OLS	-	Ordinary Least Squares
OPC	-	Ordinary Principal Component
PLR	-	Penalized Linear Regression
PP	-	Phillips-Perron
RMSE	-	Root Mean Square Error
SIS	-	Sure Independence Screening

LIST OF SYMBOLS

α	-	tuning parameter for penalized method
β	-	tuning parameter
C	-	Matrix of sample covariance
c_h	-	Inverse Fourier transform of the principal component of the cross spectral matrices for each frequency
δ_j	-	Difference term
Δ	-	Changes in stock market price series
d	-	Dimension of the data
ε_t	-	Error terms
f_t, j	-	First principal component
γ	-	Test statistics for the ADF test
$\Gamma(\cdot)$	-	Gamma complete function
h	-	Dynamic principal component
I	-	Identity matrix
k	-	Lags
λ_C	-	Eigenvalues of C
λ	-	Tuning parameter of penalized method for cross validation setup
L_1	-	LASSO penalty
L_2	-	Ridge penalty
L	-	Maximum values of the likelihood function for the model
m	-	Number of column in non-transformed matrices
σ^2	-	Variance
$'$	-	Transpose
n	-	Number of sample size
ψ	-	Complex value function
$P(\cdot)$	-	Penalty term
p	-	Number of observations
Q_2	-	Coefficient of determination
q	-	Number of estimated parameters in the model

$\mathbb{R}$	-	Real number
T	-	Number of rows in non-transformed matrices
t	-	time
$\mathbf{X}$	-	Data matrix
$\boldsymbol{x}$	-	Explanatory variables
y	-	Response variable
$\boldsymbol{z}_t$	-	Vector time series
$\bar{\mathrm{z}}$	-	Estimation of mean
$\hat{\beta}_{\mathcal{A}}$	-	Oracle concept in adaptive penalized Theorem 3.1
$\mathcal{A}$	-	Intrinsic dimension size of the underlying penalized model

CHAPTER 1

INTRODUCTION

1.1 Background of the Study

Nonstationary time series is complex and difficult to be modelled. Many researchers resolved it by transforming it into stationary time series. However, loss of generality will occur which make its inference more difficult. In addition, the originality of data with nonstationary pattern might be affected during the transformation process. Thus, there is a possible reduction of information loss, especially without performing the stationary transformation. It is worthy of note that the direct application using statistical approaches are still limited. Hence, it is pertinent to propose a non-transformed approach via principal component (Peña and Yohai, 2016). Nonetheless, the nonstationary pattern might not be easily observed. Therefore, the duration of the series needs to deliberately choose as the non-stationarity characteristic in the real application could be distinguished.

Principal component analysis is one of the established dimension reduction techniques. Indeed, principal component analysis is quite straightforward, able to reduce dimensions and concomitantly preserve the variability in the data. Overall, this technique can detect possible structures in the relationships between variables through dimensionality reduction by obtaining the information in the data from different variables that are captured by the components (Jolliffe, 2002).

Principal component analysis discovers the orthogonal linear combinations of the p original variables, of which a small number (less than p) explains most of the variation among the original variables by creating a smaller number of uncorrelated components through maximizing the variance (Lansangan and Barrios, 2009). Nonetheless, the application of principal component analysis as a description tool is often restricted by the interpretation of its first few principal components. Generally,

the principal component analysis is primarily affected by its dependencies in a given input data matrix. The technique is also affected by the variance (column) dependencies minimally. If the input data was observed over time, each column of the data matrix is considered as time series. Meanwhile, the temporal dependence in the data is summarized in the diagonal (variances) and off-diagonal (cross-covariances) elements of the variance-covariance matrix. Certainly, the principal components could still be defined properly if the columns of the time series data are stationary. This is because ill-conditioning is not visible in the variance-covariance matrix. Contrariwise, the series may be identified as correlations of the columns in a nonstationary time series.

It is worth noting that the principal component analysis usually combines those having similar variability pattern and loadings to indicate the equal importance of the component variables. Based on this principle, it shows that the first few components have taken the average of the variables and subsequently causes the principal component analysis unable to achieve dimension reduction. One of the main concerns in dimensionality reduction is interpretability. Besides, most of the current techniques used for component generating from time-dependent variables follow the assumption of stationarity of time series (Jolliffe, 2002; Lansangan and Barrios, 2009).

Indeed, in monitoring, several indicators are used to ascertain an appropriate assessment of the state of the phenomenon. However, when intervention occurs and pushing the indicators to drift, it would subsequently cause non-stationarity in the series. Due to the varying patterns among the indicators, hence, it is essential to perform a direct application on nonstationary time series so that the state of the phenomenon can benefit from a better interpretation. Since no distribution assumptions are required, the principal component is therefore, an applicable explanatory method for diverse types of data, including nonstationary (Jolliffe and Cadima, 2016).

Throughout the years, several studies have been carried out on nonstationary time series using the principal component (Galiaskarov et al., 2017; Kazor et al., 2016; Lansangan and Barrios, 2009; Zhao and Shang, 2016). One of the wide-ranged

applications of principal component methods in finance is in the financial markets. Among the examples of principal component analysis include forecasting the daily S&P 500 stock market returns (Zhong and Enke, 2017), analyzing dynamic interactions in the complex patterns of global financial indices time series (Nobi and Lee, 2016) and examining multivariate non-stationary of daily indices in the different sectors of China stock market (Zhao and Shang, 2016).

Modelling the relationship between explanatory and response variables using the statistical approach is essential for a wide range of studies. Nonetheless, the multicollinearity might be a problem in high dimensional data where it provides incorrect inference about relationships between those variables. Multicollinearity is commonly defined as a situation where the explanatory variables have an exact or approximately linear relationship (Melkumova and Shatskikh, 2017). It is worth mentioning that the multicollinearity can cause important variables under-represented due to insufficient information in the data to identify the outcomes among highly correlated variables (Yue et al., 2019). As a result, it will affect prediction accuracy (Zou and Hastie, 2005; Daye and Jeng, 2009). It is also often observed that the matrix $\boldsymbol{X}'\boldsymbol{X}$ is singular in high dimension data, as the $\boldsymbol{X}$ matrix have more columns than rows. The ordinary least square (OLS) estimator can still be obtained, but, it is highly likely will perform poorly. Meanwhile, In practice, regression coefficients may be statistically insignificant due to having incorrect signs and meaningful statistical inference (Dorugade, 2016). Therefore, to overcome these limitations, it is pertinent to address the statistical issues in high dimension data, for examples estimation instability, prediction, model overfit and interpretation (Algamal et al., 2016; Pourahmadi, 2013).

On the other hand, When the number of explanatory variables is greater than that of observations, the high dimensional sample data is assumed to be independent. Thus, in the case of high dimensional data, selecting a subset of informative variables is more crucial rather than the stationarity of the data (Bühlmann and Van De Geer, 2011). It is worthy to note that many explanatory variables are typically added to lower the possibility of model deviation. Nevertheless, having more explanatory variables may not be relevant as this decreases the accuracy of prediction and model

interpretation. Hence, the reduction technique is performed as this technique able to select the optimal number of explanatory variables that contain relevant information as well as improve the statistical model simultaneously. Thus, the technique will subsequently yield a model with better performance in prediction and interpretation power (Algamal et al., 2016).

It is noting that in high dimensional data, the conventional variable selections, such as Akaike information criteria (AIC), Mallow's Cp, and Bayesian information criteria (BIC) are not appropriate due to its high computation time (Bühlmann and Van De Geer, 2011; Chen and Chen, 2011; Lin et al., 2017). As a result, modelling high dimension data using these conventional methods may not achieve a better interpretation of the relationship between the explanatory and response variables. Most of the previous studies have used OLS as an estimation method. However, this estimator becomes unreliable in the existence of multicollinearity among explanatory variables. Besides, the computation of the OLS estimator could not be carried out if the number of explanatory variables exceeds the number of response variables (Kurnaz et al., 2018). Hence, to overcome these difficulties, penalized likelihood methods are frequently adapted.

Many statistical studies adapt penalized methods due to their capability of selecting significant variables and estimating regression coefficients concomitantly in high dimension data. Indeed, it is useful to obtain a better estimation of the prediction error to avoid overfitting of the model (Algamal and Lee, 2015a). Therefore, a penalty term is added to the likelihood function of the penalized methods since it can control the model complexity. Furthermore, it can be considered as a variable selection provider because when some constraints are introduced on the parameters, it may cause some of the parameters to be precisely zero.

The performance of penalized likelihood methods relies on the penalty term value which is, a compromise between the selected model bias and variance (Hastie et al., 2009; Fan and Lv, 2010; Fan and Tang, 2013). If a small penalty value is chosen, more explanatory variables with lower bias are selected. However, they may have higher variance. Contrariwise, if the large penalty amount is chosen, less variance will

be obtained, which subsequently leads to the selection of less explanatory variables but with higher bias. Thus, it is pertinent to determine the penalty term value to improve the accuracy of the prediction as well as obtaining a better model interpretation of the high dimensional data (James et al., 2013; Algamal and Lee, 2015).

One of the earliest penalized methods is ridge regression (Hoerl and Kennard, 1970). Ridge regression is among the widely used penalized method to overcome the multicollinearity problem that is usually present between variables in highly correlated data (Arashi and Roozbeh, 2019; Dong et al., 2018; Ijaz et al., 2019; Rabier et al., 2019). Hoerl and Kennard (1970) who introduced this method showed that regression coefficients can never equal to zero although they may converge towards zero. This result is obtained by adding an L_2-penalty in the sum of squares residual which also causes bias in the estimated parameters as well as increase the regression coefficients variances. Although the ridge regression can be applied to high dimensional data, it might however be restricted in terms of performing the variable selection. It shows that the interpretation of high dimensional models is not easily obtained (Tibshirani, 1996).

To overcome the ridge regression limitation, Tibshirani (1996) has introduced another penalized method namely least absolute shrinkage and selection operator (LASSO). LASSO utilizes L_1-penalty instead of L_2-penalty in which the variable selection can be performed by assigning zero values to some explanatory variable coefficients. Certainly, LASSO has garnered the attention of many researchers, particularly those conducting high dimensional studies. Yet, it also has some flaws where the selection of explanatory variables is usually less than the number of observations. As well as, this method attempts to choose only one variable among the highly correlated explanatory variables. The method also has no oracle properties where the probability of choosing the right explanatory variables set that have nonzero coefficients converges to one. Besides, if the zero coefficients were known beforehand, it will have a similar means and covariances as the asymptotically normal nonzero coefficients estimators. Notwithstanding these drawbacks, LASSO has become the

baseline of many penalized methods and has elucidated several extensions in diverse practical applications.

Zou and Hastie (2005) have then proposed elastic net method to overcome the limitation of the LASSO by combining both the L_1- and L_2- penalties. Indeed, both elastic net LASSO methods can perform variable selection and model estimation concurrently. However, both methods may lack in variable selection consistency resulting in poor model performance and low prediction power. Therefore, a recent approach of homogeneity is proposed specifically in highly correlated variables. The fundamental of homogeneity method was introduced by Ke et al. (2015) by dividing regression coefficients into several groups according to their regression coefficients values. For example, the values of regression coefficients in the same groups are similar or close, whereas those in different groups are significantly different from each other. Sparsity is a special concept of homogeneity where a large number of groups is entirely made up of zero coefficients. It is worth mentioning that the successful detection of homogeneity in the model causes the regression model able to recognize the original structure of the data and also increase its predictive performance. Thus, one of the most recent developed methods on homogeneity is the ordered homogeneity pursuit lasso (OHPL) which improves the limitation of the LASSO method (Lin et al., 2017).

Besides performing variable selection, another important issue in employing high dimensional data is encouraging the grouping effects between highly correlated variables. A high dimensional stock market price data set has many variables that are often larger than the number of observations. As for the explanatory variables for stock market prices, the correlations between them can be high due to the sharing of similar trading patterns. These patterns can be considered as having similar group structures. An ideal selection of stock market price approach could be obtained by removing the trivial stock market price and automatically include the whole groups into the model once one stock market price is selected among them. This is commonly known as the grouped selection property. Thus, to employ LASSO is not an ideal approach because when $p > n$, only n explanatory variables at most can be selected (Efron et al., 2004). It also lacks the capability to provide grouping information. Hence, the elastic net

method was then introduced to overcome the limitations of LASSO in variable selection and encouraging grouping effects between variables in highly correlated variables. Nevertheless, the elastic net method is not consistent in its selection of variables due to the lack of oracle property. Therefore, the adaptive elastic net method was developed and adapted in the present study.

On the other hand, the idea of adapting OLS as initial weights may be compelling. Unfortunately, this method could not be adapted in high dimension data. Therefore, the adoption of the penalized method is more suitable as the initial weight to encourage grouping effects between variables. It is regarded that the elastic net often performs much better than LASSO method in the selection of correlated explanatory variables and prediction accuracy (Zhou, 2013). Nevertheless, the elastic net method is also lack of oracle properties, particularly in the consistency of variable selection. Thus, the adaptive elastic net method was introduced by Zou and Zhang (2009) to overcome these drawbacks by using elastic net estimates as initial weights. However, applying the adaptive elastic net method with the elastic net as initial weight may cause lower precision. As well as, some significant variables may be incorrectly assigned the smaller weight values in the initial estimator. As a result, important variables may falsely be removed from the model by the penalized method, which subsequently lowers the accuracy of prediction in the selection of informative variables. Furthermore, if the pairwise correlation between the variables is low, the adaptive elastic net method may underperform. Therefore, to encourage grouping effects and reduce bias in variable selection, it is pertinent to propose an alternative initial weight which is more suitable for high dimensional data.

Most of the methods available in the literature explicitly rely on the assumption of normality. Nonetheless, most real applications present a departure from normality. This is often the case of high dimensional data due to the existence of heavy-tailed distribution or the presence of outliers in the response variable. As a result, this may affect the consistency of variable selection, encouraging grouping effect and robustness. Due to the presence of outliers in the response variable, it is known that the conventional method, such as OLS might not fully provide the desired estimation (Fan et al., 2014; Alhamzawi, 2015). Henceforth, robust penalized regression methods

have garnered the interest of some researchers due to their capability to perform robust variable selection and robust estimation. Nonetheless, a majority of previous studies do not consider the high correlation between the explanatory variables as they are only focused on solving the outlier problem. Indeed, the elastic net is better in encouraging grouping effects as compared to LASSO. However, the limitations of elastic net method should also be considered. To improve the grouping effects and robustness in the presence of outliers, an initial weight is thus proposed in the adaptive elastic net.

As of today, many stock market price researches are concentrated only on the financial aspects of the stock market price. The four main financial observations are the trading volume of the stock market prices, the behavior of the stock market, the linkage between exchange rates and the stock market, and the volatility of the stock market prices after a phenomenon such as a recession or news of a company's takeover. The in-depth investigation, on the other hand, goes into the price-earnings and earnings per share ratios and gross domestic product on specific companies or sectors in the stock market as part of the financial aspects of the stock market price. Among these studies are (Ando and Lu, 2019) which included forecasting (Nobi and Lee, 2016; Zhao and Shang, 2016; Zhong and Enke, 2017) or in machine learning studies (Inthachot et al., 2016; Qiu et al., 2016; Chatzis et al., 2018; Henrique et al., 2019; Göçken et al., 2019). Similar financial limitation aspects were also found in studies that modelled high dimensional stock market prices. As of today, the use of statistical approach is considered scarce in many studies of high dimensional stock market prices. Therefore, the application of stock market price data able to complement all the proposed methods in this study.

1.2 **Problem Statement**

It is essential to fulfill assumptions of stationarity when modelling a nonstationary time series. However, stationary transformation might contribute to the loss of data originality, notably when the series is nonstationary. Furthermore, when the data is transformed to stationary, it may have limitations when using a nonstationary approach. This may subsequently increase the complexity of performing

data analysis. On the other hand, high dimensionality is often related to selecting informative variables due to the existence of highly correlated variables. Indeed, when the number of explanatory variables exceeds the number of observations, penalized likelihood methods are often adapted. Nonetheless, the sufficiency in variable selection consistency, encouraging grouping effects and robustness to outliers, particularly in high dimension data are debatable. In addition, the selection of variables with similar homogeneity is also important as highly correlated data may possess this property. Therefore, it is pertinent to explore suitable adaptive penalized methods that able to overcome these shortcomings. Stock market prices are often nonstationary. Meanwhile, it is worth to note that when the number of stock market price is greater than that of observations, it may induce the multicollinearity problem. Thus, modelling stock market price using appropriate statistical approaches are important for a better interpretation of the data.

1.3 **Research Questions**

The study addressed the following questions related to the problem statement.

(a) How can a direct application be performed on nonstationary series without stationary transformation?

(b) How can highly correlated variables be improved through the homogenous approach?

(c) How does the proposed adaptive penalized method work on the grouping effect and robustness?

(d) What are the achievable significant advantages in using the proposed methods?

(e) Where can the proposed methods be applied to in real life?

1.4 Research Objectives

The objectives are as follows:

(a) To propose a modified algorithm for the non-transformed principal component to further improve the accuracy of nonstationary modelling.

(b) To develop a homogenous variable selection for highly correlated variables using the ordered homogeneity pursuit LASSO approach with homogenous algorithm in high dimension data.

(c) To propose a distance correlation weight for the adaptive penalized method by developing an algorithm that encourages the grouping effects in high dimensional data.

(d) To propose a robust initial weight in the penalized linear regression model for the high dimensional data.

(e) To simulate and apply the algorithms developed based on modifications on the non-transformed and proposed penalized methods for non-transformed and high dimensional stock market price data.

1.5 Significance of the Study

The analysis of stock market prices using statistical approaches is essential in this study. Thus, a better interpretation of stock market prices can be achieved through the implementation of the proposed non-transformed and penalized methods, particularly in the case of nonstationary and high dimension data. It may also assist stock companies or financial experts' decision-making process as more valuable information can be obtained through the direct application. Furthermore, the obtained results from the proposed penalized methods may be beneficial in terms of identifying which stock market price would give a significant effect to the stock index, and also

for prediction. Alternative to stationary transformation, modelling the non-transformed approach of nonstationary series by a direct application may reduce the loss of initial information, thus enabling better model interpretation. The penalized methods are adapted since they are well-recognized to achieve a better model performance and consistency in variable selection, encouraging grouping effects and improving robustness. Thus, it is crucial to enhance model performance by applying a homogenous variable selection approach. Additionally, an initial weight is proposed in the adaptive penalized methods to encourage the grouping effects and robustness of the penalized linear regression model. Notably, all approaches have their strengths and constraints. Therefore, these proposed methods are expected to improve the model performance, prediction power, variable selection, as well as robustness in comparison with the other existing methods mentioned in this study.

1.6 Scope and Limitation of the Study

The present study focuses on direct application of generally known nonstationary time series by using a non-transformed approach. Thus, only selected applications and observations are taken into consideration to ascertain the suitability and the performance of the proposed non-transformed method. Hence, Augmented Dickey-Fuller (ADF) test will be performed to the pattern of the series. Besides, the model performance will be evaluated and compared among AIC, BIC, mean square error (MSE) and percentage of explained variance.

Indeed, many high dimensions data research is more concerned about the selection of informative variables rather than the stationarity of the time series. Therefore, the proposed penalized methods are adapted to improve variable selection, grouping effects and robustness using linear regression model. The theoretical aspect of the proposed methods includes modifications and improvements made to the existing penalized methods. Also, a new initial weight is proposed in adaptive elastic net method.

Several simulation studies and real applications were tested to highlight the performance of each proposed method. The execution of both simulations and applications was carried out in R. The study has also used high performance computing (HPC) to reduce computation time in analysis. It is therefore worthy to note that comparison for this HPC was carried out for 50 times and evaluated based on root mean squared error (RMSE), coefficient of determination and number of selected variables. Meanwhile, the adaptive proposed methods were evaluated based on MSE and the number of selected variables.

Various applications of stock market prices were observed throughout the whole study by assessing their performance using different sectors, regions and observations with consideration of daily, weekly, monthly and yearly stock market prices. In general, the outliers in the stock market price are of better interest than missing values. Hence, the stock market prices that have missing values are not discussed in the study.

1.7 Thesis Organizations

This thesis is organized and structured into five main chapters. The introduction and background of the study has been discussed in Chapter 1. Meanwhile, Chapter 2 will focus on the review of past literature on non-transformed principal component and penalized linear regression methods. An insight on related methodology from the perspective of non-transformed principal components, penalized linear regression, and adaptive penalized methods will be explored in Chapter 3. The statistical properties of the existing and proposed methods which lay the fundamental ground for this study are discussed in detail in this chapter. The tuning parameter and performance assessment methods related to each study objective are also discussed. Chapter 4 focuses on the results performance detail discussion of simulations and stock market price applications. Chapter 5 ends the thesis with conclusion and recommendation as guidance for future work.

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Contr.*, *6*(19), 716–723.

Akaike, H. (1979). A Bayesian extension of the minimum AIC procedure of autoregressive model fitting. *Biometrika*, *2*(66), 237–242.

Algamal, Z. Y., & Lee, M. H. (2015a). Adjusted Adaptive LASSO in High-dimensional Poisson Regression Model. *Modern Applied Science*, *9*(4). https://doi.org/10.5539/mas.v9n4p170

Algamal, Z. Y., & Lee, M. H. (2015b). Penalized logistic regression with the adaptive LASSO for gene selection in high-dimensional cancer classification. *Expert Systems with Applications*, *42*(23), 9326–9332. https://doi.org/10.1016/j.eswa.2015.08.016

Algamal, Z. Y., & Lee, M. H. (2015c). Regularized logistic regression with adjusted adaptive elastic net for gene selection in high dimensional cancer classification. *Computers in Biology and Medicine*, *67*, 136–145.

Algamal, Z. Y., Lee, M. H., Al-Fakih, A. M., & Aziz, M. (2016). High-dimensional QSAR modelling using penalized linear regression model with L1/2-norm. *SAR and QSAR in Environmental Research*, *27*(9), 703–719. https://doi.org/10.1080/1062936X.2016.1228696

Alhamzawi, R. (2015). Model selection in quantile regression models. *Journal of Applied Statistics*, *42*(2), 445–458. https://doi.org/10.1080/02664763.2014.959905

Ando, T., & Lu, L. (2019). *Quantile co-movement in stock markets with production linkages of firms: A spatial panel quantile model with unobserved heterogeneity*.

Androulakis, E., Koukouvinos, C., & Vonta, F. (2014). Tuning parameter selection in penalized generalized linear models for discrete data. *Statistica Neerlandica*, *68*(4), 276–292.

Arashi, M., & Roozbeh, M. (2019). Some improved estimation strategies in high-dimensional semiparametric regression models with application to riboflavin production data. *Statistical Papers*, *60*(3), 317–336. https://doi.org/10.1007/s00362-016-0843-y

Bai, J., and Ng, S. (2002). Determining the Number of Factors in Approximate Factor Models. *Econometrica*, *70*, 191–221.

Basu, A., Mitra, R., Liu, H., Schreiber, S. L., & Clemons, P. A. (2018). RWEN: Response-weighted elastic net for prediction of chemosensitivity of cancer cell lines. *Bioinformatics*, *34*(19), 3332–3339. https://doi.org/10.1093/bioinformatics/bty199

Brillinger, D. R. (1974). Fourier Analysis of Stationary Processes. *Proceedings of the IEEE*, *62*(12), 1628–1643. https://doi.org/10.1109/PROC.1974.9682

Brillinger, D.R. (1981). *Time Series: Data Analysis and Theory.* San Francisco.: Holden-Day.

Brillinger, David R. (1981). *Time series: data analysis and theory*. SIAM.

Bühlmann, P., & Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications.* Springer Science & Business Media.

Caporale, G. M., Gil-Alana, L. A., & Tripathy, T. (2019). Volatility Persistence In The Russian Stock Market. *Finance Research Letters*.

Cavalcante, R.C., Brasileiro, R.C., Souza, V.L., Nobrega, J.P. and Oliveira, A.L. (2016). Computational Intelligence and Financial Markets: A Survey and Future Directions. *Expert Systems with Applications*, *55*(6), 194–211. https://doi.org/10.1016/j.eswa.2016.02.006

Chatvorawit, P., Sattayatham, P., & Premanode, B. (2016). Improving stock price prediction with SVM by simple transformation: The sample of stock exchange of Thailand (SET). *Thai Journal of Mathematics*, *14*(3), 553–563.

Chatzis, S. P., Siakoulis, V., Petropoulos, A., Stavroulakis, E., & Vlachogiannakis, N. (2018). Forecasting stock market crisis events using deep and statistical machine learning techniques. *Expert Systems with Applications*, *112*, 353–371. https://doi.org/10.1016/j.eswa.2018.06.032

Chaudhuri, A., & Hu, W. (2019). A fast algorithm for computing distance correlation. *Computational Statistics and Data Analysis*, *135*, 15–24. https://doi.org/10.1016/j.csda.2019.01.016

Chen, J., & Chen, Z. (2011). Extended BIC for small-n-large-P sparse GLM. *Statistica Sinica*, *22*(2), 555–574. https://doi.org/10.5705/ss.2010.216

Daye, Z. J., & Jeng, X. J. (2009). Shrinkage and model selection with correlated variables via weighted fusion. *Computational Statistics and Data Analysis*, *53*(4), 1284–1298. https://doi.org/10.1016/j.csda.2008.11.007

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, *74*(366a), 427–431.

Dobson, A. J., & Barnett A.G. (2008). *An Introduction to generalized linear models* (3rd Editio). Chapman & Hall/CRC, Boca Raton.

Dong, Y., Song, L., & Amin, M. (2018). SCAD-Ridge Penalized Likelihood Estimators for Ultra-high Dimensional Models. *Hacettepe Journal of Mathematics and Statistics*, *47*(2), 423–436. https://doi.org/10.15672/hjms.201612518375

Donoho, D. L., & Johnstone, J. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, *81*(3), 425–455.

Dorugade, A. V. (2016). Improved Ridge Estimator in Linear Regression with Multicollinearity, Heteroscedastic Errors and Outliers. *Journal of Modern Applied Statistical Methods*, *15*(2), 362–381. https://doi.org/10.22237/jmasm/1478002860

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, *32*(2), 407–499.

El Anbari, M., & Mkhadri, A. (2014). Penalized regression combining the L1 norm and a correlation based penalty. *Sankhya B*, *76*(1), 82–102. https://doi.org/10.1007/s13571-013-0065-4

Fan, J., & Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, *32*(3), 928–961.

Fan, J., Fan, Y., & Barut, E. (2014). Adaptive robust variable selection. *Annals of Statistics*, *42*(1), 324–351. https://doi.org/10.1214/13-AOS1191

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*(456), 1348–1360.

Fan, J., & Lv, J. (2008). Sure independence screening for ultra-high-dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *70*(5), 849–911.

Fan, J., & Lv, J. (2010). A Selective Overview of Variable Selection in High Dimensional Feature Space. *Statistica Sinica*, *20*(1), 101–148. https://doi.org/10.1038/jid.2014.371

Fan, Jianqing, & Lv, J. (2010). A Selective Overview of Variable Selection in High

Dimensional Feature Space. *Statistica Sinica*, *20*(1), 101–148. https://doi.org/10.1038/jid.2014.371

Fan, Y., & Tang, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *75*(3), 531–552. https://doi.org/10.1111/rssb.12001

Forni, M., Giannone, D., Lippi, M., & Reichlin, L. (2009). Opening the Black Box: Structural Factor Models with Large Cross Sections. *Econometric Theory*, *25*, 1319–1347.

Forni, M., Giannone, D., Lippi, M., & Reichlin, L. (2009). Opening the Black Box: Structural Factor Models with Large Cross Sections. *Econometric Theory*, *25*, 1319–1347.

Forni, M., Hallin, M., Lippi, M., & Reichlin, L. (2000). The Generalized Dynamic Factor Model: Identification and Estimation. *Review of Economics and Statistics*, *82*(4), 540–554.

Forni, M., Hallin, M., Lippi, M., & Reichlin, L. (2005). The Generalized Dynamic Factor Model: One Sided Estimation and Forecasting. *Journal of the American Statistical Association*, *100*(471), 830–840.

Forni, M., Hallin, M., Lippi, M., & Zaffaroni, P. (2015). Dynamic Factor Models with Infinite-Dimensional Factor Spaces: One-Sided Representations. *Journal of Econometrics*, *185*(2), 359–371.

Forni, M., & Lippi, M. (2011). The General Dynamic Factor Model: One-Sided Representation Results. *Journal of Econometrics*, *163*, 23– 28.

Galiaskarov, M. R., Kurkina, V. V., & Rusinov, L. A. (2017). Online diagnostics of time-varying nonlinear chemical processes using moving window kernel principal component analysis and Fisher discriminant analysis. *Journal of Chemometrics*, *31*(8), 1–9. https://doi.org/10.1002/cem.2866

Geweke, J. (1977). *The dynamic factor analysis of economic time series* (D. J. Aigner & A. S. Goldberger, Eds.). https://doi.org/10.4236/jss. 2014.211024 3,122

Ghosh, S. (2011). On the grouped selection and model complexity of the adaptive elastic net. *Statistics and Computing*, *21*(3), 451–462.

Göçken, M., Özçalıcı, M., Boru, A., & Dosdoğru, A. T. (2016). Integrating metaheuristics and artificial neural networks for improved stock price prediction. *Expert Systems with Applications*, *44*, 320–331.

Göçken, Mustafa, Özçalıcı, M., Boru, A., & Dosdoğru, A. T. (2019). Stock price

prediction using hybrid soft computing models incorporating parameter tuning and input variable selection. *Neural Computing and Applications*, *31*(2), 577–592. https://doi.org/10.1007/s00521-017-3089-2

Gregory, K. B., Wang, D., & McMahan, C. S. (2019). Adaptive elastic net for group testing. *Biometrics*, *75*(1), 13–23.

Hara, S., Kawahara, Y., Washio, T., von Bünau, P., Tokunaga, T., & Yumoto, K. (2012). Separation of stationary and non-stationary sources with a generalized eigenvalue problem. *Neural Networks*, *33*, 7–20. https://doi.org/10.1016/j.neunet.2012.04.001

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* ((2nd ed.)). NY: Springer.

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC.

Henrique, B. M., Sobreiro, V. A., & Kimura, H. (2019). Literature review: Machine learning techniques applied to financial market prediction. *Expert Systems with Applications*, *124*, 226–251. https://doi.org/10.1016/j.eswa.2019.01.012

Hoerl, E., & Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67. https://doi.org/10.2307/1271436

Hörmann, S., Kidziński, Ł., & Hallin, M. (2015). *Dynamic Functional Principal Components*. 1–36.

Hui, Y., Wong, W. K., Bai, Z., & Zhu, Z. Z. (2017). A New Nonlinearity Test to Circumvent the Limitation of Volterra Expansion with Applications. *Journal of the Korean Statistical Society*, *46*(3), 365–374. https://doi.org/10.1227/01.NEU.0000349921.14519.2A

Ijaz, M., Asghar, Z., & Gul, A. (2019). Ensemble of penalized logistic models for classification of high-dimensional data. *Communications in Statistics: Simulation and Computation*, *0*(0), 1–17. https://doi.org/10.1080/03610918.2019.1595647

Inthachot, M., Boonjing, V., & Intakosum, S. (2016). Artificial Neural Network and Genetic Algorithm Hybrid Intelligence for Predicting Thai Stock Price Index Trend. *Computational Intelligence and Neuroscience*, *2016*. https://doi.org/10.1155/2016/3045254

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. NY: Springer.

Jolliffe, I T; (2002). *Principal component analysis and factor analysis. Principal component analysis*. Springer.

Jolliffe, Ian T, & Cadima, J. (2016). Principal component analysis : a review and recent developments Subject Areas : Author for correspondence : *Philosophical Transactions of the Royal Society A*, *374*, 20150202. https://doi.org/http://dx.doi.org/10.1098/rsta.2015.0202

Kazor, K., Holloway, R. W., Cath, T. Y., & Hering, A. S. (2016). Comparison of linear and nonlinear dimension reduction techniques for automated process monitoring of a decentralized wastewater treatment facility. *Stochastic Environmental Research and Risk Assessment*, *30*(5), 1527–1544.

Ke, Y., Li, J., & Zhang, W. (2016). Structure identification in panel data analysis. *The Annals of Statistics*, *44*(3), 1193–1233.

Ke, Z. T., Fan, J., & Wu, Y. (2015). Homogeneity Pursuit. *Journal of the American Statistical Association*, *110*(509), 175–194. https://doi.org/10.1080/01621459.2014.892882

Keenan, D. M. (1985). A Tukey nonadditivity-type test for time series nonlinearity. *Biometrika*, *72*(1), 39–44.

Khurana, M., Chaubey, Y. P., & Chandra, S. (2014). Jackknifing the ridge regression estimator. *Communications in Statistics-Theory and Methods*, *43*(24), 5249–5262.

Kokoszka, P., Rice, G., & Shang, H. L. (2017). Inference for the autocovariance of a functional time series under conditional heteroscedasticity. *Journal of Multivariate Analysis*, *162*, 32–50. https://doi.org/10.1016/j.jmva.2017.08.004

Kurnaz, F. S., Hoffmann, I., & Filzmoser, P. (2018). Robust and sparse estimation methods for high-dimensional linear and logistic regression. *Chemometrics and Intelligent Laboratory Systems*, *172*, 211–222. https://doi.org/10.1016/j.chemolab.2017.11.017

Kwiatkowski, D., Phillips, P. C., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics*, *54*(13), 159–178.

Lam, C., & Yao, Q. (2012). Factor modeling for high-dimensional time series: inference for the number of factors. *The Annals of Statistics*, *40*(2), 694–726.

Lansangan, J. R. G., & Barrios, E. B. (2009). Principal components analysis of

nonstationary time series data. *Statistics and Computing*, *19*(2), 173.

Lee, K. M., Lee, M., Seok, J., & Han, S. W. (2019). Regression-Based Network Estimation for High-Dimensional Genetic Data. *Journal of Computational Biology*, *26*(4), 336–349. https://doi.org/10.1089/cmb.2018.0225

Li, J., Jia, Y., & Zhao, Z. (2013). Partly adaptive elastic net and its application to microarray classification. *Neural Computing and Applications*, *22*(6), 1193–1200.

Li, R., Zhong, W., & Zhu, L. (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association*, Vol. 107, pp. 1129–1139. https://doi.org/10.1080/01621459.2012.695654

Lin, Y. W., Xiao, N., Wang, L. L., Li, C. Q., & Xu, Q. S. (2017). Ordered homogeneity pursuit lasso for group variable selection with applications to spectroscopic data. *Chemometrics and Intelligent Laboratory Systems*, *168*(January), 62–71. https://doi.org/10.1016/j.chemolab.2017.07.004

McCarthy, D., & Jensen, S. T. (2016). Power-weighted densities for time series data. *Annals of Applied Statistics*, *10*(1), 305–334. https://doi.org/10.1214/15-AOAS893

McElroy, T., & Monsell, B. (2015). Model estimation, prediction, and signal extraction for nonstationary stock and flow time series observed at mixed frequencies. *Journal of the American Statistical Association*, *110*(511), 1284–1303.

Melkumova, L. E., & Shatskikh, S. Y. (2017). Comparing Ridge and LASSO estimators for data analysis. *Procedia Engineering*, Vol. 201, pp. 746–755. https://doi.org/10.1016/j.proeng.2017.09.615

Motta, G., & Ombao, H. (2012). Evolutionary factor analysis of replicated time series. *Biometrics*, *68*(3), 825–836.

Mu, W., & Xiong, S. (2015). Robust sparse regression with high-breakdown value. *Communications in Statistics-Theory and Methods*, *44*(5), 1033–1043.

Nelder, J. A., & Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, *3*(135), 370-384.

Nobi, A., & Lee, J. W. (2016). State and group dynamics of world stock market by principal component analysis. *Physica A: Statistical Mechanics and Its Applications*, Vol. 450, pp. 85–94. https://doi.org/10.1016/j.physa.2015.12.144

Noguchi, K., Aue, A., & Burman, P. (2016). Exploratory Analysis and Modeling of

Stock Returns. *Journal of Computational and Graphical Statistics*, *25*(2), 363–381.

Pan, J., & Yao, Q. (2008). Modelling multiple time series via common factors. *Biometrika*, *95*(2), 365–379.

Park, H. (2017). Outlier-resistant high-dimensional regression modelling based on distribution-free outlier detection and tuning parameter selection. *Journal of Statistical Computation and Simulation*, Vol. 87, pp. 1799–1812. https://doi.org/10.1080/00949655.2017.1287186

Park, J. (2017). Tolerance intervals from ridge regression in the presence of multicollinearity and high dimension. *Statistics and Probability Letters*, *121*, 128–135. https://doi.org/10.1016/j.spl.2016.10.016

Peña, D., & Poncela, P. (2006). Nonstationary dynamic factor analysis. *Journal of Statistical Planning and Inference*, *136*(4), 1237–1257. https://doi.org/10.1016/j.jspi.2004.08.020

Peña, D., & Yohai, V. J. (2016). Generalized Dynamic Principal Components. *Journal of the American Statistical Association*, *111*(515), 1121–1131. https://doi.org/10.1080/01621459.2015.1072542

Phillips, P. C., & Perron, P. (1988). Testing for a unit root in time series regression. *Biometrika*, *75*(2), 335–346.

Pourahmadi, M. (2013). *High-dimensional covariance estimation: with high-dimensional data*. John Wiley & Sons.

Qiu, M., Song, Y., & Akagi, F. (2016). Application of artificial neural network for the prediction of stock market returns: The case of the Japanese stock market. *Chaos, Solitons and Fractals*, *85*, 1–7. https://doi.org/10.1016/j.chaos.2016.01.004

Rabier, C. E., Mangin, B., & Grusea, S. (2019). On the accuracy in high-dimensional linear models and its application to genomic selection. *Scandinavian Journal of Statistics*, *46*(1), 289–313. https://doi.org/10.1111/sjos.12352

Rish, I., & Grabarnik, G. (2014). *Sparse Modeling:Theory, Algorithms, and Applications*. CRC press.

Sánchez, M. Á., Trinidad, J. E., García, J., & Fernández, M. (2015). The effect of the underlying distribution in Hurst exponent estimation. *PLoS ONE*, *10*(5).

Shen, C., Priebe, C. E., & Vogelstein, J. T. (2019). From Distance Correlation to Multiscale Graph Correlation. *Journal of the American Statistical Association*, 1–22. https://doi.org/10.1080/01621459.2018.1543125

Shen, X., & Huang, H. C. (2010). Grouping pursuit through a regularization solution surface. *Journal of the American Statistical Association*, *105*(490), 727–739.

Shin, S. J., & Artemiou, A. (2017). Penalized principal logistic regression for sparse sufficient dimension reduction. *Computational Statistics & Data Analysis*, *111*, 48–58.

Silva, J. L. D., Mexia, J. T., & Ramos, L. P. (2015). On the strong consistency of ridge estimates. *Communications in Statistics-Theory and Methods*, *44*(3), 617–626.

Souza Filho, J.B. and Diniz, P. S. (2017). A Fixed-Point Online Kernel Principal Component Extraction Algorithm. *IEEE Transactions on Signal Processing*, *23*(65), 6244–6259.

Stock, J. H., & Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, *97*(460), 1167–1179.

Stock, J. H., & Watson, M. W. (2006). Forecasting with many predictors. *Handbook of Economic Forecasting*, *1*, 515–554.

Suhail, M., Chand, S., & Kibria, B. M. G. (2019). Quantile based estimation of biasing parameters in ridge regression model. *Communications in Statistics: Simulation and Computation*, *0*(0), 1–13. https://doi.org/10.1080/03610918.2018.1530782

Székely, G. J., & Rizzo, M. L. (2009). Brownian distance covariance. *Ann. Appl. Stat*, *3*(4), 1236–1265.

Székely, G. J., & Rizzo, M. L. (2013). The distance correlation t-test of independence in high dimension. *Journal of Multivariate Analysis*, *117*, 193–213.

Székely, G. J., & Rizzo, M. L. (2017). The energy of data. *Annual Review of Statistics and Its Application*, *4*, 447–479.

Székely, G. J., Rizzo, M. L., & Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *Annals of Statistics*, *35*(6), 2769–2794. https://doi.org/10.1214/009053607000000505

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.

Wang, H., Li, B., & Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *71*(3), 671–683.

Wang, M. (2015). Nonconvex penalized ridge estimations for partially linear additive models in ultrahigh dimension. *Statistical Methodology*, *26*, 1–15.

Wang, S., Nan, B., Rosset, S., & Zhu, J. (2011). Random lasso. *The Annals of Applied Statistics*, *5*(1), 468.

Wang, X., & Wang, M. (2016). Variable selection for high-dimensional generalized linear models with the weighted elastic-net procedure. *Journal of Applied Statistics*, *43*(5), 796–809. https://doi.org/10.1080/02664763.2015.1078300

Wang, Xun, Zhang, Z., & Li, S. (2016). Set-valued and interval-valued stationary time series. *Journal of Multivariate Analysis*, *145*, 208–223. https://doi.org/10.1016/j.jmva.2015.12.010

Wang, Y., Yang, X. G., & Lu, Y. (2019a). Informative gene selection for microarray classification via adaptive elastic net with conditional mutual information. *Applied Mathematical Modelling*, *71*, 286–297. https://doi.org/10.1016/j.apm.2019.01.044

Wang, Y., Yang, X. G., & Lu, Y. (2019b). Informative gene selection for microarray classification via adaptive elastic net with conditional mutual information. *Applied Mathematical Modelling*, *71*, 286–297. https://doi.org/10.1016/j.apm.2019.01.044

Xiao, N., & Xu, Q. S. (2015). Multi-step adaptive elastic-net: reducing false positives in high-dimensional variable selection. *Journal of Statistical Computation and Simulation*, *85*(18), 3755–3765. https://doi.org/10.1080/00949655.2015.1016944

Xin, X., Hu, J., & Liu, L. (2017). On the oracle property of a generalized adaptive elastic-net for multivariate linear regression with a diverging number of parameters. *Journal of Multivariate Analysis*, *162*, 16–31.

Yu, Y., & Feng, Y. (2014). Modified cross-validation for penalized high-dimensional linear regression models. *Journal of Computational and Graphical Statistics*, *23*(4), 1009–1027.

Yue, L., Li, G., Lian, H., & Wan, X. (2019). Regression adjustment for treatment effect with multicollinearity in high dimensions. *Computational Statistics and Data Analysis*, *134*, 17–35. https://doi.org/10.1016/j.csda.2018.11.002

Zhang, R., Zhang, F., Chen, W., Xiong, Q., Chen, Z., Yao, H., … Du, Y. (2019). A variable informative criterion based on weighted voting strategy combined with LASSO for variable selection in multivariate calibration. *Chemometrics and Intelligent Laboratory Systems*, *184*(September 2018), 132–141. https://doi.org/10.1016/j.chemolab.2018.11.015

Zhao, X;, & Shang, P. (2016). Principal component analysis for non-stationary time series based on detrended cross-correlation analysis. *Nonlinear Dynamics*, *84*(2), 1033–1044.

Zhao, X, & Shang, P. (2016). Principal component analysis for non-stationary time series based on detrended cross-correlation analysis. *Nonlinear Dynamics*, *84*(2), 1033–1044.

Zhong, W., Zhu, L., Li, R., & Cui, H. (2016). Regularized quantile regression and robust feature screening for single index models. *Statistica Sinica*, *26*(1), 69.

Zhong, W., & Li, R. (2015). A selective overview of feature screening for ultrahigh-dimensional data. *Science China Mathematics*, *58*(10), 1–22. https://doi.org/110.1016/j.bbi.2017.04.008

Zhong, Wei, & Zhu, L. (2015). An iterative approach to distance correlation-based sure independence screening. *Journal of Statistical Computation and Simulation*, *85*(11), 2331–2345. https://doi.org/10.1080/00949655.2014.928820

Zhong, X., & Enke, D. (2017). Forecasting daily stock market return using dimensionality reduction. *Expert Systems with Applications*, *67*, 126–139. https://doi.org/10.1016/j.eswa.2016.09.027

Zhou, D. X. (2013). On grouping effect of elastic net. *Statistics and Probability Letters*, *83*(9), 2108–2112. https://doi.org/10.1016/j.spl.2013.05.014

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418–1429.

Zou, H., & Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

Zou, H., & Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *Annals of Statistics*, *37*(4), 1733–1751. https://doi.org/10.1214/08-AOS625

Zou, Hui, & Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *Annals of Statistics*, *37*(4), 1733–1751. https://doi.org/10.1214/08-AOS625

LIST OF PUBLICATIONS

Indexed Journal

1. Andu, Y., Lee, M. H., & Algamal, Z. Y. (2019). Non-transformed principal component technique on weekly construction stock market price. *MATEMATIKA: Malaysian Journal of Industrial and Applied* Mathematics, 35(2), 139-147. (**WoS**)
2. Andu, Y., Lee, M. H., & Algamal, Z. Y. (2020). Penalized linear regression on the grouping effect and robust of high dimensional stock market price. *Will be submitted for publication* in Communications for Statistical Applications and Methods . (**Q3**)

Indexed Conference Proceedings

1. Andu, Y., Lee, M. H., & Algamal, Z. Y. (2020). Variable selection of yearly high dimension stock market price using ordered homogenous pursuit lasso. *Accepted for publication in AIP Conference Proceedings.* (**SCOPUS)**

2. Andu, Y., Lee, M. H., & Algamal, Z. Y. (2018). Generalized dynamic principal component for monthly nonstationary stock market price in technology sector. In *Journal of Physics: Conference Series* (Vol. 1132, No. 1, p. 012076). IOP Publishing. (**SCOPUS**)

Non-Indexed Conference Proceedings

1. **Yusrina Andu**, Muhammad Hisyam Lee and Zakariya Yahya Algamal. 26-27 November 2019. Variable selection of yearly high dimension stock market price using ordered homogenous pursuit lasso. Symposium Kebangsaan Sains Matematik ke 27 (SKSM27), Universiti Putra Malaysia, Hotel Tenera, Bangi. (**Indexed by SCOPUS)**

2. **Yusrina Andu**, Muhammad Hisyam Lee and Zakariya Yahya Algamal. 18-23th August 2019. Non-transformed dimensionality reduction on monthly agriculture stock market price. 62nd ISI World Statistics Congress 2019. (**Speaker CPS-session**)

3. **Yusrina Andu**, Muhammad Hisyam Lee and Zakariya Yahya Algamal. 13-15th August 2018. Nonstationary weekly construction stock market price using non-transformed principal component technique. 7th International Graduate Conference on Engineering, Science and Humanities 2018 (ICGESH 2018), Universiti Teknologi Malaysia. Faculty of Built Environment, UTM. (**Presenter**)

4. **Yusrina Andu**, Muhammad Hisyam Lee and Zakariya Yahya Algamal. 2-3rd May 2018. Non-transformed dimensionality reduction technique on daily nonstationary stock market price in healthcare sector. Emerging Scientist Conference 2018 (ESCon 2018), Universiti Teknologi Malaysia. Pulai Springs Resort. (**Presenter**)

5. **Yusrina Andu**, Muhammad Hisyam Lee and Zakariya Yahya Algamal. 6-8th February 2018. Generalized dynamic principal component for monthly

nonstationary stock market price in technology sector. 3rd International Conference on Mathematical Sciences and Statistics (ICMSS) 2018, Universiti Putra Malaysia and Banasthali University. Le Meridian Hotel Putrajaya. (**Presenter**)