MINISTRY OF EDUCATION
AND TRAINING

VIETNAM ACADEMY OF
SIENCE AND TECHNOLOGY

VIETNAM ACADEMY OF SCIENCE AND TECHNOLOGY

NGUYEN MINH HAI

KERNEL-BASED VARIABLE DIMENSION REDUCTION METHOD AND ITS APPLICATION FOR FORECASTING EXPORT TURNOVER

SUMMARY OF DISSERTATION ON INFORMATION SYSTEM
Code: 9 48 01 04

Ha Noi - 2024

The dissertation is completed at: Vietnam Academy of Science and Technology, Vietnamese Academy of Science and Technology.

Supervisors:

Supervisor 1: Assoc. Prof. Do Van Thanh, Duy Tan University

Supervisor 2: Assoc. Prof. Nguyen Duc Dung, Institute of Information Technology

Referee 1: …

Referee 2: …

Referee 3: ….

The dissertation will be examined by Examination Board of Graduate University of Science and Technology, Vietnam Academy of Science and Technology at…….. (time, date, year…)

This dissertation can be found at:

1) Graduate University of Science and Technology Library
2) National Library of Vietnam

LIST OF THE PUBLICATIONS RELATED TO THE DISSERTATION

1. Thanh, D. Van, Hai, N. M., & Hieu, D. D. Building unconditional forecast model of Stock Market Indexes using combined leading indicators and principal components: application to Vietnamese Stock Market. *Indian Journal of Science & Technology*, *11*(2), 2018. https://doi.org/10.17485/ijst/2018/v11i2/104908.
2. Hai, N. M., Thanh, D. Van, & Dung, N. D. Building Export Forecast Model Using a Kernel-based Dimension Reduction Method. *Economic Computation and Economic Cybernetics Studies and Research*, *56*(1), pp.91–106, 2022. https://doi.org/10.24818/18423264/56.1.22.06.
3. Thanh, D. Van, & Hai, N. M. The performance of a kernel-based variable dimension reduction method. *In Nature of Computation and Communication: 8th EAI International Conference, ICTCC 2022, Cham: Springer Nature Switzerland*, 2023. https://doi.org/10.1007/978-3-031-28790-9_4.
4. Nguyễn Minh Hải, Đỗ Văn Thành và Nguyễn Đức Dũng. Xây Dựng Mô Hình Dự Báo Không Điều Kiện Sử Dụng Phương Pháp Giảm Chiều Biến Dựa Vào Thủ Thuật Kernel, *Proceedings of the 15th National Conference on Fundamental and Applied Information Technology*, pp. 211-218, 2022. https://doi.org/ *10.15625/vap.2022.0226*
5. Thanh, D. Van, & Hai, N. M. Forecast of the VN30 Index by Day Using a Variable Dimension Reduction Method Based on Kernel Tricks. *In Nature of Computation and Communication: 7th EAI International Conference, ICTCC 2021, Virtual Event, October 28–29, 2021*, Proceedings 7, pp. 83-94. Springer International Publishing, 2021. https://doi.org/10.1007/978-3-030-92942-8_8
6. Đỗ Văn Thành và Nguyễn Minh Hải. Dự báo trên tập dữ liệu chuỗi thời gian lớn sử dụng phương pháp giảm chiều dựa vào hàm kernel và ứng dụng. *Hội thảo quốc gia lần thứ 25: Một số vấn đề chọn lọc của công nghệ thông tin và truyền thông*, pp. 48-54, 2022.

INTRODUCTION

1. Research basis and motivation

Real-world datasets in the field of economics and finance often consist of time series data, where the number of variables is generally large, even much larger than the number of observations. Building forecasting models and performing forecasting on such datasets using statistical techniques is impossible. There are currently two main approaches to overcome this challenge: deep learning and dimensionality reduction.

The deep learning approach considered most suitable for time series datasets is the use of Long Short-Term Memory (LSTM) neural network models (C. Zhang et al., 2024), (Sako et al., 2022), (Zaheer et al., 2023), (Hopp, 2022), Gated Recurrent Unit (GRU) models (Torres et al., 2021), and Transformer models for time series data (Ahmed et al., 2023), (Wen et al., 2022). However, LSTM and GRU deep learning models have limitations in handling input sequential data with long-term dependencies, capturing temporal backpropagation relationships, handling seasonal patterns, and dealing with a large number of variables and gradient issues (Vaswani et al., 2017). In the research (Kapetanios et al., 2018), LSTM and GRU models are suitable for forecasting problems on datasets where the number of observations is large, but the number of variables is not too large. The Transformers deep learning model has the advantage of capturing dependencies and interactions over long ranges between variables, which is why it is attracting research attention for time series forecasting. The results achieved by the Transformer model for time series data are still at an initial level (Wen et al., 2022). Through experimental research, (Zeng et al., 2023) showed that the model based on a simple multilayer neural network can still achieve better forecasting results than the time series Transformer model. It can be said that the application of the aforementioned deep learning methods in forecasting problems on large time series datasets (or datasets with a large number of time series variables) in the fields of economics and finance is still limited (Hopp, 2022), (Sezer et al., 2020; Torres et al., 2021). Research (Hopp, 2022) showed that the application of deep learning methods in socio-economic forecasting is still in its early stages, partially due to the limitations involved in their implementation.

The research (Kim & Swanson, 2018b) found ample evidence that combining dimensionality reduction techniques and machine learning techniques to build forecasting models is a dominant approach in building forecasting models on large time series datasets. Studies (Chikamatsu et al., 2021), (Bragoli, 2017), (Urasawa, 2014), (Jardet & Meunier, 2022), (Chinn et al., 2023) have demonstrated that the forecasting accuracy of models built based on factor models, where factors are extracted from the original dataset using PCA or SPCA dimensionality reduction methods, is always equal to or higher than that of other benchmark forecasting models. A recent study (Chinn et al., 2023) also evaluates that the forecasting accuracy of models built on large time series datasets using the three-step approach of variable selection, PCA dimensionality reduction, and economic random forest regression is the highest compared to models built using various other approaches, including deep learning, Markov chains, quantile regression, linear least squares estimation, etc.

PCA is a typical linear dimensionality reduction method. Research (Shlens, 2014) demonstrated that PCA is the best linear dimensionality reduction method as it preserves the covariance structure and maximizes the variance of the original dataset. Experimental studies (Van

Der Maaten et al., 2009), (Zhong & Enke, 2017) have shown that none of the top 12 non-linear dimensionality reduction methods performed better than PCA on real-world datasets, although all 12 methods showed good dimensionality reduction results on artificial datasets. Research (Koren & Carmel, 2004) has shown that PCA is ineffective for datasets that do not approximate a hyperplane. Thus, the research results of (Van Der Maaten et al., 2009), (Zhong & Enke, 2017) revealed that the real-world datasets used in those studies seem to approximate a hyperplane. However, reality shows that real-world time series datasets are not always like that.

The above presentations are the motivation for the Thesis to propose a new variable dimensionality reduction method for large time series data sets. The studies (Chikamatsu et al., 2021), (Bragoli, 2017), (Urasawa, 2014), (Jardet & Meunier, 2022), and especially (Van Der Maaten et al., 2009), (Zhong & Enke, 2017), and (Chinn et al., 2023) suggest that this method is a natural extension of the PCA method (i.e., in exceptional cases, the proposed method is the PCA method), overcoming the limitations of the PCA method indicated in the study (Koren & Carmel, 2004) that it can be used to reduce the dimensionality of large time series datasets that are not approximately a hyperplane, and the dimensionality reduction performance of the proposed method should be equal to or higher than the dimensionality reduction performance of the PCA method. Here, the performance of a dimensionality reduction method is measured by the root mean squared error (RMSE) as the LOSS function.

The purpose of dimensionality reduction is to increase efficiency (less time and memory) and ease of interpretation for forecasting models built on large data sets using dimensionality reduction methods. Proposing a forecasting process or algorithm on a large time series data set using the proposed variable dimensionality reduction method and applying that process or algorithm to forecast important economic-financial indicators also needs to be researched. For every country, forecasting the export turnover of the entire economy and each economic sector is always one of the most important macroeconomic forecasts. Vietnam has an open economy, where export and import turnover account for a high proportion of gross domestic product (GDP), so forecasting export turnover is even more important and necessary. Along with the deepening international integration process, the factors affecting Vietnam's export turnover are increasingly greater. The problem of forecasting export turnover on large data sets has been raised. Therefore, proposing a forecasting process/algorithm using the proposed dimensionality reduction method and applying it to forecast Vietnam's monthly export turnover is also one of the main research motivations for the Ph.D. student to carry out their thesis " KERNEL-BASED VARIABLE DIMENSION REDUCTION METHOD AND ITS APPLICATION FOR FORECASTING EXPORT TURNOVER."

Specifically, the Thesis focuses on researching and proposing a dimensionality reduction method on large time series data sets to overcome limitations and have superior dimensionality reduction performance compared to some commonly used and considered the most effective dimensionality reduction methods in economics and finance. Propose a forecasting process/algorithm on a large time series data set using the proposed dimensionality reduction method and its application in the economic and financial fields, first of all, the area of exportation.

2. Research objectives of the Thesis

The general objective of this Thesis is to research and propose a method for effectively variable dimensionality reduction on large time-series datasets and their applications in economic and financial forecasting. The specific objectives of the Thesis are as follows:

- Propose a novel dimensionality reduction method to overcome the disadvantages of prevalent dimensionality reduction methods effectively utilized in the field of economics and finance. The proposed dimensionality reduction method not only mitigates these limitations but also achieves dimensionality reduction performance comparable to or better than widely adopted methods in the economics and finance domain.

- Propose a forecasting process/algorithm (conditional and unconditional) for large time series datasets using the proposed dimension reduction method and apply it to forecast Vietnam's export turnover on a dataset of a large number of economic and financial indicators.

3. Layout of the Thesis

The thesis layout consists of the following sections:-

- **Introduction**: This section presents the theoretical foundation and research motivation of the Thesis, along with its research objectives, subjects, scope, research methodology, primary contributions, and thesis structure.

- **Chapter 1:** Provides an overview of methods for building forecasting and nowcasting models on large time series datasets. It identifies the problem and scope of research, offers some related knowledge, and draws some conclusions.

- **Chapter 2**: Proposes a variable dimensionality reduction method for large time series datasets based on the kernel trick, called KTPCA, and comparison of the variable dimensionality reduction performance of the KTPCA method based on the RMSE-best model with that of PCA and SPCA methods on datasets with and without the same sampling frequencies, and concludes with some findings.

Chapter 3: Proposes conditional and unconditional forecasting algorithms on large time series datasets using the proposed dimensionality reduction method and applies this algorithm to forecast the monthly export turnover of Vietnam with and without conditions.

The conclusion section presents the main research contributions of the Thesis and discusses its limitations.

CHAPTER 1. OVERVIEW OF THE METHOD FOR BUILDING FORECASTING MODELS ON LARGE TIME SERIES DATA SETS

1.1. Overview of domestic and foreign research

The overview of domestic and foreign research is presented in 17 pages. For details, please refer to pages 9 through 24 in the Thesis.

1.2 Remaining problems

Based on the above analysis and evaluation of related domestic and foreign research, the Thesis focuses on researching solutions to overcome the above problems. Specifically, the Thesis focuses on researching:

1) Propose a new dimensionality reduction method that is considered a natural extension of the PCA method while overcoming the disadvantages of the PCA method on datasets that do not approximate a hyperplane and achieve dimensionality reduction performance equal to or better than the dimensionality reduction performance of PCA and SPCA methods in forecasting and nowcasting problems, respectively, on datasets with and without the same sampling frequencies.

2) Propose a forecasting process or algorithm using the proposed dimensionality reduction method and apply it to forecast important macroeconomic indicators on a large dataset.

1.3 Some basic knowledge

The content of this section presents the basic knowledge for the Thesis, spanning 20 pages. For details, please refer to pages 28 through 48 in the Thesis.

1.4 Conclusion

In this chapter, the Thesis has presented several English terms that, when translated into Vietnamese, closely align with the concept of "forecasting". This chapter has provided an overview of relevant domestic and foreign studies to identify research gaps, thereby defining the Thesis's problem and research scope. It has also presented some fundamental knowledge essential for subsequent research chapters.

CHAPTER 2. THE KERNEL-BASED VARIABLE DIMENSIONALITY REDUCTION METHOD

This chapter proposes a new dimensionality reduction method based on the kernel trick, serving as another natural extension of the PCA method, named the KTPCA method. The experimental evaluation of the KTPCA method's dimensionality reduction performance is based on the RMSE-best model (referred to as KTPCA#) on datasets with the same sampling frequency, as well as mixed sampling frequency, in comparison to the dimensionality reduction performance of PCA, SPCA, RSPCA, and ROBSPCA methods which are also presented in this chapter.

2.1. The method of variable dimensionality reduction based on the kernel trick

Suppose $\mathbf{X} = [X_1, X_2, \dots, X_m]_{N\times m}$ is a dataset of time series explanatory variables where $X_i \in \mathbb{R}^{\mathbf{N}}, i = 1, \dots, m$; m is very large. Without loss of generality, $\mathbf{X}$ is a centered matrix, meaning $\sum_{j=1}^{N} x_{i_j} = 0\,, \forall \mathrm{i} = 1, \dots, \mathrm{m}$.

2.1.1. The kernel-based dimensionality reduction method

Chapter 1 clearly stated that although KPCA is a natural extension of PCA for linear datasets, PCA is still the best dimensionality reduction method for such datasets. The performance of KPCA in reducing dimensionality is not as good as that of PCA for approximately linear datasets. Determining the level of linearity approximation of a dataset to ensure that the dimensionality reduction performance of PCA is better than KPCA remains an open issue. This Thesis has not addressed this problem. However, the idea of KPCA suggests a new dimensionality reduction method based on the kernel trick, called KTPCA, to differentiate it from KPCA. This method differs from KPCA. Please refer to pages 49-50 of the Thesis.

The dimensionality reduction algorithm using the KTPCA method can be written as pseudocode as follows:

KTPCA Algorithm

Input: $X \in \mathbb{R}^{N\times m}$

Output: $Y \in \mathbb{R}^{N\times p}$

1. Construct the kernel matrix $K=[\kappa(X_i,X_j)] \equiv [\Phi(X_i).\boldsymbol{\Phi}(X_j)]$
2. Find the eigenvalues and eigenvectors of the kernel matrix
3. Sort the eigenvectors according to the eigenvalues in decreasing order
4. Construct the matrix $\tilde{\mathbf{E}}_{m\times p}$ with the first p eigenvectors
5. Transform X using $\tilde{\mathbf{E}}_{m\times p}$ to obtain a new subspace $Y = X.\ \tilde{\mathbf{E}}_{m\times p}$

Thus, it can be seen that the KTPCA method is a combination of dimensionality reduction ideas from both KPCA and PCA methods. When the kernel function $\boldsymbol{\kappa}$ is the inner product of two input vectors, i.e., $\boldsymbol{\kappa}(X_i,X_j) = \langle X_i,X_j\rangle$, the kernel matrix **K** becomes the covariance matrix, and the KTPCA method becomes the PCA method. This is what the Thesis aims for.

When employing the KTPCA method for dimensionality reduction, the selection of a suitable kernel function is crucial so that the RMSE of the dependent variable forecast model based on the extracted factors corresponding to this kernel function is minimal. Similar to the KPCA method, there is currently no standard for choosing the optimal kernel function for the KTPCA method. Consequently, the most appropriate kernel function for reducing data dimensionality using the KTPCA method can only be determined through trial and error, based on the RMSE-best model. The KTPCA method based on the RMSE-best model is referred to as KTPCA#.

Table 2.1 below summarizes the PCA, KPCA, and KTPCA methods. This table shows the main differences between these methods; refer to pages 49 and 53 in the Thesis.

Table 2.1: Differences between PCA, KPCA, and KTPCA methods

PCA (Shlens, 2014)	**KPCA** (Schölkopf et. al. 1998)	**KTPCA**
- The data set $\mathbf{X} \in \mathbb{R}^{N\times m}$ is mean-centered. - Find the eigenvalues and eigenvectors of the covariance matrix of **X** - Arrange eigenvectors according to eigenvalues - The first p factors are determined by: $\boldsymbol{PC}_{N\times p} = \boldsymbol{X}_{N\times m}.\boldsymbol{E}_{m\times p}$	- The data set $\mathbf{X} \in \mathbb{R}^{N\times m}$ - Determine the kernel matrix $\mathbf{K} = [\boldsymbol{\kappa}(\chi_i,\chi_j)]$, χ_i is the data point vector of **X** and the Gramm matrix of level $N \times N$: - $\boldsymbol{K_c} = \boldsymbol{K} - \boldsymbol{K}.\mathbf{1}_N - \mathbf{1}_N.\boldsymbol{K} + \mathbf{1}_N.\boldsymbol{K}.\mathbf{1}_N$ - Find the eigenvalues and eigenvectors of $\mathbf{K_c}$ - The principal components of the kernel are determined through the point function: $f_v(\boldsymbol{\Phi}(Z)) = v.\boldsymbol{\Phi}(Z) = \sum_{i=1}^{m}\alpha_i\ \boldsymbol{\Phi}(\chi_i).\boldsymbol{\Phi}(Z) = \sum_{i=1}^{m}\alpha_i\,\boldsymbol{\kappa}(\chi_i, Z)$, here Z is the data point of **X**.	- The data set $\mathbf{X} \in \mathbb{R}^{N\times m}$ is mean-centered. - Determine the kernel matrix $\boldsymbol{K}_{m\times m} = [\boldsymbol{\kappa}(X_i,X_j)]$, X_i is the data vector of **X.** - Find the eigenvalues and eigenvectors of **K** orresponding to the kernel function $\boldsymbol{\kappa}$; - The first p factors are determined by: $\boldsymbol{PC}_{N\times p} = \boldsymbol{X}_{N\times m}.\widetilde{\boldsymbol{E}}_{m\times p}$

2.1.2. Variable dimensionality reduction using KTPCA# method

The variable dimensionality reduction using the KTPCA# method is presented in Figure 2.1 below. Figure 2.1 illustrates that the forecasting or nowcasting model constructed using the KTPCA# dimension reduction method always has equal or higher forecast accuracy compared to the model constructed using the PCA dimension reduction method.

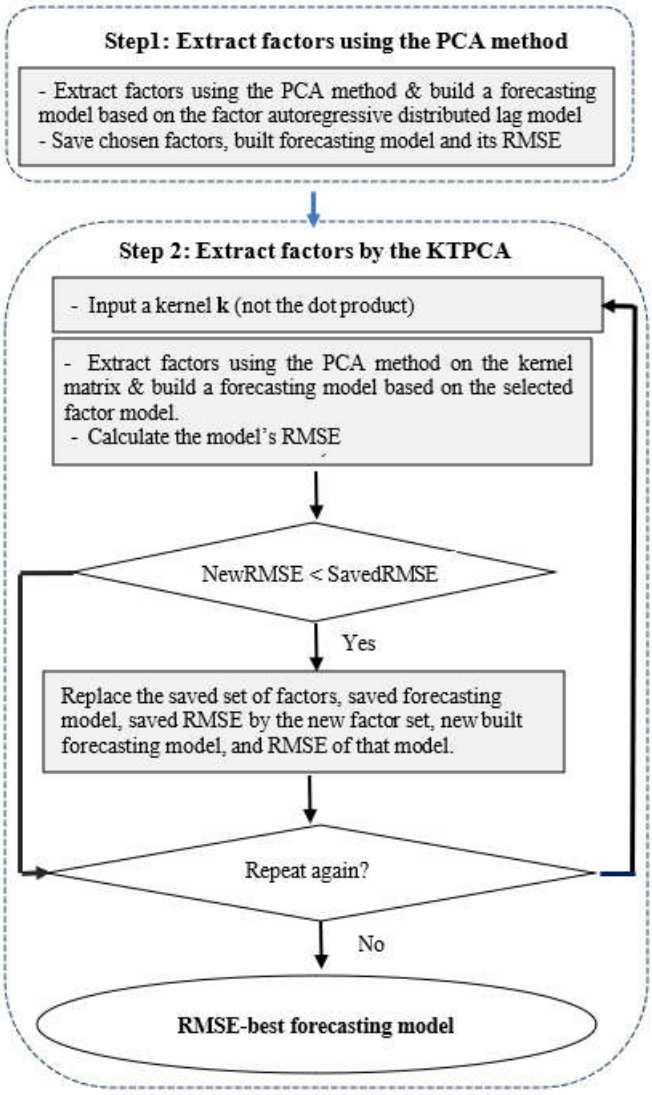

Figure 2.1: Flow chart of the KTPCA method based on an RMSE- best model

2.3. Dimensionality reduction performance of the KTPCA# method

The performance of a dimensionality reduction method is measured by the RMSE of a nowcast or forecasting model constructed based on the DFM model or the factor ARDL model, in which the factors are extracted from the large dataset of predictors at higher frequency, along with predictors at the same frequency as the dependent variable, utilizing the KTPCA# method. A smaller RMSE indicates higher performance of the dimensionality reduction method; refer to details on pages 55-56 in the Thesis.

2.2.1. For datasets with the same sampling frequency

2.2.1.1 Experimental data

The datasets utilized for the experiments comprise 04 real-world datasets of the Vietnamese economy and 07 datasets from the UCI-Machine Learning Repository, as outlined in Table 2.2 below; refer to pages 56-57 in the Thesis.

Table 2.2: Statistical characteristics of experimental data sets

Data Sets	Type of Data set	Type of Attribute	Num. of Obser.	Num. of Variables	Missing Data	Dependent Variable	Freq.
EXP	Time series	Real	60	63	No	Total Export	Monthly
VN30	Time series	Real	366	34	No	VN30 Index	Daily
CPI	Time series	Real	72	102	No	CPI Index	Monthly
VIP	Time Series	Real	60	265	No	Value of Industries Production	Monthly
Residential Building	Multivariate	Real	371	27	No	Sales Price	
S&P500	Time series	Real	1760	52	Yes	S&P500 Index	Daily
DJI	Time series	Real	1760	81	Yes	Dow Jones Index	Daily
NASDAQ	Time series	Real	1760	81	Yes	Nasdaq Index	Daily
Air Quality	Time series	Real	9348	12	Yes	CO of Air	Hour
Appliances Energy	Time series	Real	19704	23	No	The energy use of Appliances (wh)	Every 10 min
SuperConduct.	Multivariate	Real	21263	81	No	Critical temperature	

2.2.1.2. Experimental method

To compare the dimensionality reduction performance of the KTPCA# method with the PCA, SPCA, RSPCA, and ROBSPCA methods across 11 experimental datasets, the Thesis selected only 06 different kernel functions for experimentation with the KTPCA method, including 03 polynomial kernel functions and 03 Gaussian kernel functions. Specifically, the selected kernel functions are as follows: for the polynomial kernel functions, the special polynomial kernel function $\boldsymbol{\kappa}(X_i, X_j) = \boldsymbol{PL}(1,1,0)$ is consistently included, rendering the KTPCA and PCA methods equivalent; for the EXP, VN30, CPI, Air Quality, and Appliances Energy datasets, the other two polynomial kernel functions are of the form $\boldsymbol{\kappa}(X_i, X_j) = \boldsymbol{PL}(1,2,0.5)$ and $\boldsymbol{\kappa}(X_i, X_j) = \boldsymbol{PL}(1,3,0.5)$, while for the remaining datasets, the two polynomial kernel functions are $\boldsymbol{\kappa}(X_i, X_j) = \boldsymbol{PL}(0.5,2,0.5)$ and $\boldsymbol{\kappa}(X_i, X_j) = \boldsymbol{PL}(0.5,3,0.5)$. For the Gaussian kernel function with parameter ρ^2, the parameter values for the three selected functions are equal to, less than, and greater than the value ρ_0^2, denoted as $\boldsymbol{GA_4}, \boldsymbol{GA_5}$, and $\boldsymbol{GA_6}$, respectively. The ARDL model, according to equation (1.34), is utilized to construct the forecasting model using the dataset of the explanatory variables with the same sampling frequency.

2.2.1.3 Result

The minimum average distance between two column vectors of the 11 datasets used in the experiment is calculated according to formula (2.2) and presented in Table 2.3 for the corresponding datasets. This value serves as an important hint to choose appropriate Gaussian kernel functions $\boldsymbol{\kappa}(X_i, X_j) = \boldsymbol{GA}(\rho^2)$ when implementing the KTPCA method on a certain corresponding dataset.

a. The performance of KTPCA# compared to PCA, SPCA, RSPCA, and ROBSPCA methods

Extracted from Table A1 in the Appendix, Table 2.4 summarizes the dimensionality reduction results of KTPCA#, PCA, SPCA, RSPCA, and ROBSPCA methods across 11 experimental datasets with the same sampling frequency of explanatory variables.

Table 2.4: Dimensionality reduction performance of the KTPCA# method

Datasets	Method	KTPCA#	PCA	SPCA	RSPCA	ROBSPCA
	Number of factors	$GA_6, 6$	14	10	10	10
EXP	RMSE	**0.0104**	NA	NA	NA	NA
	Number of factors	$GA_4, 14$	14	14	14	15
VN30	RMSE	**0.1819**	0.1895	0.1968	0.1968	0.2054
	Number of factors	$GA_5, 6$	4	4	4	4
CPI	RMSE	**0.4452**	1.4836	1.0659	1.0673	1.0659
	Number of factors	$PL_2, 4$	4	4	4	4
VIP	RMSE	**672.66**	715.96	826.28	1373.57	2642.83
	Number of factors	$GA_5, 2$	1	1	1	1
Res. Building	RMSE	**919.9**	1152.4	1152.5	1152.5	1151.2
	Number of factors	$GA_5, 2$	1	1	1	1
S&P500	RMSE	**61.60**	161.415	161.441	161.441	161.441
	Number of factors	$PL_1, 1$	1	1	1	1
DJI	RMSE	**91.82**	**91.82**	309.24	309.24	309.23
	Number of factors	$PL_2, 1$	1	1	1	1
NASDAQ	RMSE	**81.05**	365.97	85.47	85.47	85.46
	Number of factors	$GA_5, 5$	1	1	1	1
Air Quality	RMSE	**50.297**	71.459	71.499	71.499	71.427
	Number of factors	$GA_5, 6$	3	3	3	3
App. Energy	RMSE	**98.81**	101.74	101.76	101.76	101.75
	Number of factors	$GA_5, 2$	2	2	2	2
SuperCon.	RMSE	**26.094**	27.314	27.332	27.332	27.319

where, the NA symbol is "No Available".

For the EXP data set (line 1), when using PCA as the dimensionality reduction method, the number of principal component factors chosen is 10. Consequently, we cannot regress the dependent variable on the data set of 60 observations and 76 explanatory variables, which includes 10 selected factors + (10 factors + 01 dependent variable) lagged from 1 to 6. However, if the variable dimensionality reduction method is KTPCA, the above challenge can be easily overcome.

From the analysis in Table 2.4, it can be concluded that the dimensionality reduction performance of the KTPCA# method is equal to or higher than that of the PCA and SPCA family methods.

b. The performance of the PCA method compared to the SPCA method family

Table 2.5 (excluding data related to the KTPCA# method) below and Figure 2.2 both suggest that the dimension reduction performance of PCA and SPCA methods is competitive. This finding contradicts the traditional belief that the SPCA family method seems to outperform the PCA method in dimension reduction performance (refer to pages 62-63 in the Thesis).

Table 2.5: Dimensionality reduction performance of methods (RMSE)

Methods	*DS2*	*DS3*	*DS4*	*DS5*	*DS6*
KTPCA#	**0.1819**	**0.4452**	**672.6600**	**919.9000**	**61.6000**
PCA	0.1895	1.4836	715.9608	1152.3950	161.4154
SPCA	0.1968	1.0660	826.2757	1152.5310	161.4407
RSPCA	0.1968	1.0673	1373.5670	1152.5310	161.4407
ROBSPCA	0.2054	1.0659	2642.8340	1151.2470	161.4410

Methods	*DS7*	*DS8*	*DS9*	*DS10*	*DS11*
KTPCA[#]	**91.8236**	**81.0500**	**50.2970**	**98.8100**	**26.0940**
PCA	**91.8236**	365.9698	71.45873	101.7423	27.3143
SPCA	309.2405	85.4666	71.4989	101.7635	27.3318
RSPCA	309.2405	85.4666	71.4989	101.7635	27.3318
ROBSPCA	309.2349	85.4621	71.4266	101.7468	27.3193

2.2.2 For the mixed frequency data set

In this section, regression models, such as the BE factor model, U-MIDAS factor model, and several other factor models with limitations, including the STEP-MIDAS factor model, PAW-MIDAS factor model, and EAW-MIDAS factor model, are utilized to construct the nowcast model.

2.2.2.1 Experimental datasets

Table 2.6 presents the experimental data sets used. Specifically, these include 07 datasets from the UCI—Machine Learning repository, as listed in Table 2.2, and 03 real-world datasets related to the Vietnamese economy, including the CPI dataset from Table 2.2 and the newly added RGDP and IIP datasets (refer to pages 64 - 65 in the Thesis).

Table 2.6: Statistical characteristics of experimental data sets

Stat. Characteristics	RGDP	CPI	IIP	Air Quality	App. Energy
Characteristics of dataset	Time-series	Time-series	Time-series	Time-series	Time-series
Variable characteristic	Real	Real	Real	Real	Real
No. of low-freq. variables	3	3	1	1	1
No. of high-freq. variables	87	102	42	12	27
Total number of Obser.	72	72	1840	9348	19704
No. of low-freq. Obser.	24	24	92	779	3284
S -No. of high-freq. values for a low-freq. value [1]	3	3	20	12	6
Missing data	No	No	Yes	Yes	No
The dependent variable	The growth rate of GDP	Consumer Price Inflation	Index of Industrial Production	The Air CO	Energy use of Appliances
Các đặc điểm thống kê	Res. Build.	S&P 500	DJI	NASDAQ	SuperCond.
Characteristics of dataset	cross data	Time-series	Time-series	Time-series	cross data
Variable characteristic	Real	Real	Real	Real	Real
No. of low-freq. variables	1	1	1	1	1
No. of high-freq. variables	27	52	81	81	81
Total number of Obser.	366	1760	1760	1760	21260
No. of low-freq. Obser.	122	88	88	88	1063
S -No. of high-freq. values for a low-freq. value	3	20	20	20	20
Missing data	No	Yes	Yes	Yes	No
The dependent variable	Sale Price	S&P 500 Index	DJI Index	Nasdaq Index	Critical Temperature

[1] : Total number of observations (or number of high-frequency observations) = *S* * number of low-frequency observations.

2.2.2.2 Experimental method

To construct nowcast models, the low-frequency dependent variable, explanatory variables at the same frequency as the dependent variable, and factors extracted from the higher-frequency explanatory variables are first transformed into stationary time series. The criterion for selecting the number of factors at high frequency is the percentage of their cumulative eigenvalues (Zhang et al., 2012). The nowcast models are estimated under ideal conditions, where the lags of the high-frequency explanatory variables are precisely determined. For further details, refer to pages 66-67 in the Thesis.

Additionally, the dimensionality reduction performance of the KTPCA# method is compared with the PCA, SPCA, RSPCA, and ROBSPCA methods on the 06 kernel functions mentioned in Section 2.2.1.2.

2.2.2.3 Result

Excluding the RGDP and IIP datasets, the remaining 08 datasets in Table 2.6 are sourced from the respective datasets with the same names in Table 2.2. Furthermore, the number of high-frequency explanatory variables and the number of observations in these 08 datasets remain unchanged compared to the corresponding datasets in Table 2.2. Therefore, the minimum average distance between two column vectors in these eight datasets is determined, as shown in Table 2.3. For the RGDP and IIP datasets, the distances are $\rho_0^2 = \exp(1.464)$ and $\rho_0^2 = \exp(8.978)$, respectively.

With the same cumulative eigenvalue percentage threshold of 75% for all the variable dimensionality reduction methods mentioned above, for all experimental data sets and 05 regression models (BE, PAW- MIDAS, STEP-MIDAS, U-MIDAS and EAW-MIDAS), dimensionality reduction results, RMSE of the forecasting models according to the factors extracted by the dimensionality reduction methods and the most appropriate kernel functions among 06 experimental kernel functions are presented in Table B (Appendix).

a. *Performance of KTPCA# compared to PCA, SPCA, RSPCA, and ROBSPCA methods*

Table 2.7 below is extracted from Table B in the Appendix section. This table includes five sub-tables 3a, 3b, 3c, 3d, and 3e containing RMSE of nowcast models built based on BE factor models, U-MIDAS models, STEP-MIDAS models, PAW-MIDAS models, and EAW-MIDAS models. Here, the factors are extracted from the aforementioned experimental datasets using PCA, SPCA, RSPCA, ROBSPCA, and KTPCA# methods.

Table 2.7 also demonstrates that for all ten experimental datasets and the 05 regression models mentioned above, the dimension reduction performance using the KTPCA# method is consistently superior. Specifically, across all 05 regression models, it is always possible to select a kernel function such that the RMSE of the nowcast model built on factors extracted by the KTPCA method corresponding to this kernel function is less than or equal to the RMSE of nowcast models built on factors extracted by one of the PCA, SPCA, RSPCA, and ROBSPCA methods.

Table 2.7: The variable dimensionality reduction performance of the proposed methods

3a. BE	PCA	SPCA	RSPCA	ROBSPCA	KTPCA#	**3b.STEP**	PCA	SPCA	RSPCA	ROBSPCA	KTPCA#
SET1	**0.000493**	0.000788	0.00079	0.000788	**0.000493**	SET1	**0.00744**	0.009727	0.009722	0.009727	**0.00744**
SET2	**0.000183**	0.000485	0.00051	0.000485	**0.000183**	SET2	0.008236	0.00439	0.004387	0.00439	**0.003948**
SET3	1.348981	1.203836	1.04437	1.545299	**0.56932**	SET3	26.52232	21.39361	28.86856	28.13315	**8.78805**
SET4	0.615228	0.611051	0.6104	0.61106	**0.592861**	SET4	**0.630038**	0.63004	0.63004	**0.630038**	**0.630038**
SET5	377.6252	377.2618	377.262	377.0618	**360.131**	SET5	**385.1972**	385.68	385.68	385.3454	**385.1972**
SET6	565.5147	565.523	565.523	565.516	**513.6189**	SET6	430.8412	430.8373	430.8373	430.8397	**421.709**
SET7	**4.3074**	4.3076	4.3076	4.3076	**4.3074**	SET7	259.8844	259.8083	257.6644	259.8065	**72.7871**
SET8	57.1033	56.4321	56.4321	56.4321	**56.2975**	SET8	4101.593	4101.958	4101.958	4102.275	**1024.708**
SET9	18.5945	18.5941	18.5941	18.5489	**18.3479**	SET9	1419.767	1419.807	1419.807	1419.756	**687.2987**
SET10	13.5381	13.5397	13.5425	13.5429	**13.3662**	SET10	14.3425	14.3462	14.3462	14.3431	**13.9649**
3c.PAW	PCA	SPCA	RSPCA	ROBSPCA	**KTPCA#**	**3d.EAW**	PCA	SPCA	RSPCA	ROBSPCA	**KTPCA#**
SET1	**0.000026**	0.000208	0.000197	0.000208	**0.000026**	SET1	0.005232	0.005274	0.005277	0.005274	**0.004544**
SET2	**0.001473**	0.001833	0.001819	0.001833	**0.001473**	SET2	0.006911	0.005465	0.007418	0.005465	**0.00509**
SET3	1.1268	0.7342	0.7508	0.6208	**0.0433**	SET3	4.4983	4.7174	4.3561	4.3146	**4.1810**
SET4	0.6298	0.6293	0.6402	0.6298	**0.6174**	SET4	0.4762	0.4765	0.4765	0.4761	**0.4392**
SET5	384.4007	384.4115	384.3218	384.3270	**384.0171**	SET5	385.4549	385.4515	385.4515	385.4597	**385.000**
SET6	404.3389	399.4798	399.4798	399.4800	**399.3498**	SET6	504.9074	504.9076	504.9076	504.9069	**379.0157**
SET7	40.7019	42.8444	42.8444	42.8444	**33.6159**	SET7	**2.806**	2.953	2.953	2.953	**2.8060**
SET8	337.8048	337.8025	337.8025	337.8026	**311.3913**	SET8	240.0	239.7	239.7	239.5	**118.900**
SET9	107.9667	107.9666	107.9666	107.9666	**107.0302**	SET9	82.2279	82.1254	82.1254	82.0357	**36.3656**
SET10	13.9580	13.9580	13.9580	13.9580	**13.9485**	SET10	13.9322	13.931	13.931	13.9322	**13.9302**
3e.U	PCA	SPCA	RSPCA	ROBSPCA	KTPCA#	**3e.U**	PCA	SPCA	RSPCA	ROBSPCA	KTPCA#
SET1	0.00204	0.000951	0.000919	0.000951	**0.000699**	SET6	430.1182	430.1732	430.1732	430.1286	**389.1229**
SET2	**0.000109**	0.002515	0.002955	0.002512	**0.000109**	SET7	0.000701	5.58E-05	0.0000558	0.0000587	**0.0000546**
SET3	**0.0283**	0.9860	0.3109	0.6632	**0.0283**	SET8	2.932	2.931	2.931	2.931	**2.9300**
SET4	0.4054	0.4058	0.4058	0.4055	**0.3330**	SET9	0.8993	0.8992	0.8992	0.8992	**0.8841**
SET5	376.9851	377.4016	377.4016	376.8008	**351.2000**	SET10	14.0231	14.0219	14.0219	14.0231	**13.9115**

Note: The symbols SET1 through SET10 in Table 2.7 correspond to the ten experimental datasets in Table 2.6.

b. Performance of PCA compared to SPCA, RSPCA, and ROBSPCA methods

Figures 2.3, 2.4, 2.5, 2.6, and 2.7 below are drawn from the respective sub-tables 3a, 3b, 3c, 3d, and 3e in Table 2.7 above and Table 2.8 below. The dimension reduction performance of the SPCA method is not superior to that of the PCA method; it is competitive. Refer to the detailed information on pages 70-72 in the Thesis.

Table 2.8: Dimension reduction performance of PCA compared to the SPCA family

DFM model	*Bằng*	*Cao hơn*	*Thấp hơn*
BE	SET4, SET5, SET6, SET8, SET9, SET10	SET1, SET2, SET3	SET7
STEP3-MIDAS	SET5, SET6, SET7, SET8, SET9, SET10	SET1, SET4	SET2, SET3
PAW2-MIDAS	SET4, SET5, SET6, SET8, SET9, SET10	SET1, SET2, SET7	SET3
EAW-MIDAS	SET1, SET5, SET6, SET8, SET9, SET10	SET3, SET4, SET7	SET2
U-MIDAS	SET4, SET5, SET6, SET8, SET9, SET10	SET2, SET3	SET1, SET7

2.4 Conclusion

This chapter proposes a dimension reduction method based on kernel trick (KTPCA for short). It also highlights the differences between it, KPCA and PCA methods. The KTPCA method is considered a natural extension of the PCA method, since it becomes the PCA method when the kernel function is the dot product of two vectors. The KTPCA method has overcome the limitation

of the PCA method in that it can reduce the dimensionality of data sets that do not approximate a hyperplane. The dimension reduction performance of the KTPCA method, based on the RMSE-best model, is equal to or higher than that of the PCA, SPCA, RSPCA, and ROBSPCA methods on datasets with the same and mixed sampling frequency.

This chapter also demonstrates that the dimension reduction performance for both datasets with the same sampling frequency and mixed datasets of the PCA method and the SPCA variants is competitive. This contradicts the long-held belief that SPCA methods outperform the PCA method in dimension reduction performance.

The research findings related to this chapter are published in studies [CT3] and [CT6] in the list of author's publications.

CHAPTER 3. THE FORECASTING ON LARGE TIME SERIES DATASET USING THE KERNEL-BASED DIMENSION REDUCTION METHOD

Chapter 3 proposes unconditional and conditional forecasting algorithms on large datasets, utilizing the KTPCA# dimensionality reduction method introduced in Chapter 2. The forecasting models are constructed based on the factor ARDL model, employing equation (1.34) for the conditional forecast model and equation (1.16) for the unconditional forecast model, in which factors are extracted using the KTPCA# method. Additionally, the chapter presents the modeling the forecast of Vietnam's export turnover by monthly frequency using the proposed algorithm.

3.1 Unconditional and conditional forecasting process using KTPCA# method

The forecasting process on large time series datasets using the KTPCA# dimensionality reduction method is developed based on the economic and financial forecasting modeling process outlined in section 1.3.6 of Chapter 1.

Figure 3.1 includes two figures, 3.1a and 3.1b, respectively, depicting the conditional and unconditional forecasting process on a large time series dataset using the KTPCA# dimensionality reduction method. Both of these processes can be divided into four stages. While the main content that needs to be done in the basic stages remains consistent, there are some differences. Specifically, the main content of the stages in these two forecasting processes is presented in detail on pages 73 - 79 in the Thesis.

.

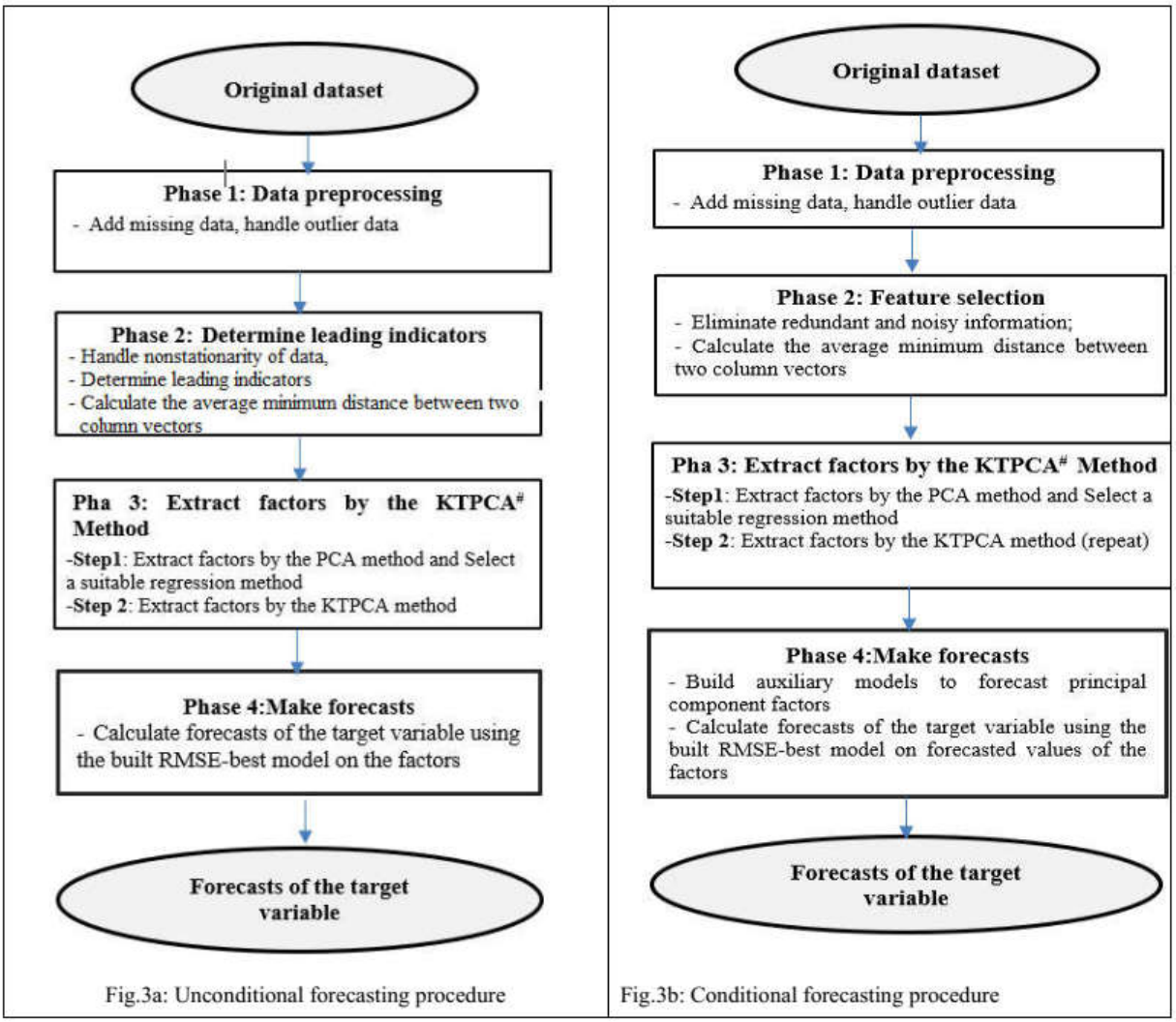

Figure 3.1: Flow chart of unconditional forecasting and conditional forecasting process

Table 3.1 summarizes the results comparing the approach to building a conditional forecast model in this Thesis with the 3-step approach to building a forecast model in the study (Chinn et al., 2023); refer to pages 78-79 in the Thesis.

Table 3.2: Comparison of two approaches to building conditional forecast models

Thesis vs. Research (Chinn et al., 2023)	Stage 2- Step 1: Feature selection	Stage 3- Step 2: Feature Learning	Stage 4- Step 3: Regression method
Thesis	Use Least Angle Regression (LARs) but handle redundant data. Rating: better	The dimensionality reduction method was used on both the data sets that approximated or did not approximate a hyperplane. Rating: better	The autoregressive distributed lag (ARDL) model is based on factors extracted from the data set of all input variables. Rating: Poorer
Research	Use Least Angle	Use the PCA	Economic random forest regression.

(Chinn et al., 2023)	Regression, but do not handle redundant data. Rating: poorer	dimensionality reduction method (a particular case of the dimensionality reduction method in the Thesis) for data sets that do not approximate the hyperplane. Rating: poorer	Its essence is to divide explanatory variables into subgroups, build a forecast model for the dependent variable on subgroups using the autoregressive distributed lag (ARDL) model, and then combine the dependent variable forecast results of the component models. Rating: better

3.2 Forecasting algorithms on large time series data sets

These algorithms are constructed following the process outlined in Figure 3.1. Suppose $\mathbf{X_t} = [X_{1,t}, X_{2,t}, \dots, X_{m,t}] \in \mathbb{R}^{\mathbf{t\times m}}$ is a data set of time series variables, $X_{i,t} \in \mathbb{R}^{\mathbf{t}}, i = 1, \dots, m$; $Y_t \in \mathbb{R}^{\mathbf{t}}$ is the dependent variable, where *m* and *t* are the number of variables and observations, respectively; m is very large.

The problem is to develop an algorithm capable of automatically generating unconditional or conditional forecasts of the dependent variable Y_t based on the set of explanatory variables $\mathbf{X_t}$.

The forecasting algorithms proposed in the subsequent section for large time series datasets are constructed based on the aforementioned forecasting process.

3.2.1 Conditional and unconditional forecasting algorithm

Without loss of generality, assuming the dataset of explanatory variables, denoted as x, is mean-centered. This dataset is utilized to extract factors using the KTPCA method corresponding to each kernel function tested.

Conditional and unconditional forecasting algorithms for large time series datasets are presented in pseudocode as follows:

Algorithm 1a: **CONF algorithm**	Algorithm 1b: **UNCONF algorithm**
Input: $\mathbf{X_t} \in \mathbb{R}^{\mathbf{t\times m}}$, $Y_t \in \mathbb{R}^{\mathbf{t}}$, α and β are user-defined relevant and redundant thresholds, q(%) is the user-defined threshold of the cumulative eigenvalue percentage. **Output**: $\hat{Y}_{t+h}$: the h-step-ahead forecast made at the time *t* of the variable Y_t on $\mathbf{X_t}$. **Begin** 1. Determining h - the farthest time of forecast; 2. *Repetition* ← "Yes"; 3. *FeatureSelection* ($\mathbf{X_t}$, Y_t); 4. Center $\mathbf{X_t}$; 5. Calculate the average minimum distance between two data vectors of predictors; 6. Tính ma trận hiệp phương sai **K** của $\mathbf{X_t}$; 7. *FeatureLearning(***K***)*; 8. Save the retained factors, the forecast model on the set of retained factors, and the RMSE of this model. 9. **Repeat** 10. Input a kernel $\boldsymbol{\kappa}$**:** $\mathbb{R}^{\mathbf{t}}$ x $\mathbb{R}^{\mathbf{t}} \longrightarrow \mathbb{R}$**;** 11. Calculate the kernel matrix $\boldsymbol{K}$; 12. *FeatureLearning* ($\boldsymbol{K}$);	**Input:** $\mathbf{X_t} \in \mathbb{R}^{\mathbf{t\times m}}$, $Y_t \in \mathbb{R}^{\mathbf{t}}$, q(%) is the user-defined threshold of the cumulative eigenvalue percentage. **Output**: $\hat{Y}_{t+h}$: the out-of-sample the h-step-ahead forecast made at the time *t* of the variable Y_t *// h is at least 1 but not predefined.* **Begin** 1. Determine the common lag p for all variables; …… . 3. *LeadingIndicatorSelection* ($\mathbf{X_t}$, Y_t); …… . ……

13. **if** *RMSE* of the newly built model < RMSE currently saved **then** *Replace the saved factors, forecast model, and corresponding saved RMSE with the new factors, newly built forecast model, and RMSE of this model.* 14. **end** 15. **Until (***Repetition* = "No") 16. *Forecast(*$\widehat{Y}_{t+h}$, The forecasting model of variable Y_t*);* **End.**	…… …… 16. *Calculate(*$\widehat{Y}_{t+h}$,The forecasting model of variable Y_t*);* **End.**

The functions *FeatureSelection*, *LeadingIndicatorSelection*, the procedures *FeatureLearning*, *Forecast*, and *Calculate* are introduced in more detail below.

Algorithm 2a: *FeatureSelection* Algorithm
 	
Input: $\mathbf{X_t} \in \mathbb{R}^{\mathbf{t\times m}}$, $Y_t \in \mathbb{R}^{\mathbf{t}}$.
Output: Subset of non-redundant and relevant variables for Y_t in $\mathbf{X_t}$.
begin
1. Remove little relevance or irrelevant variables to Y_t.
2. Order **($\mathbf{X_t}$)** // *Arrange variables in descending order of Pearson measure*.
3. Remove redundant variables of $\mathbf{X_t}$
4. return $\mathbf{X_t}$
end;
	Algorithm 2b: *LeadingIndicatorSelection* Algorithm
	Input: $\mathbf{X_t} \in \mathbb{R}^{\mathbf{t\times m}}$, $Y_t \in \mathbb{R}^{\mathbf{t}}$, p is the general lag.
	Output: Subset of leading indicators of Y_t with the lag of p in $\mathbf{X_t}$; α (%) - statistical significance level;
	begin
	1. Transform Y_t and variables in $\mathbf{X_t}$ into stationary time series variables;
	2. **for** each variable in $\mathbf{X_t}$ perform
	- Build a forecasting model to variable $Y_t$according to this variable based on model (2.2)
	- Calculate the probability of the F statistic in the forecast model;
	3. **if** that probability < α **then** the explanatory variable is the leading indicator;
	4. **end for**
	end;

Algorithm 3a: *FeatureLearning* Procedure **Input**: The kernel matrix $\mathbf{K_{m\times m}}$. **Output:** The set of factors is retained; forecast model Y_t according to the retained factors, and the *RMSE* of this model **begin** 1. Calculate eigenvalues and eigenvectors of	Algorithm 3b: *FeatureLearning* Procedure **Input**: Matrix $\mathbf{K_{N\times g}}$ is the kernel matrix of the dataset including *g* leading indicator; **Output:** The set of chosen factors; The forecast model of variable $Y_t$on the selected factors, and the RMSE of this model. **begin**

the matrix **K** 2. Sort the eigenvectors in descending order of their respective eigenvalues; 3. Extract factors by projecting the mean-centered input dataset $\mathbf{X_t}$ onto eigenvectors; 4. Create a set of the first p factors such that their cumulative eigenvalue percentage is not less than the given q(%); 5. Building the forecasting model of Y_t on retained factors based on the ARDL model; 6. Calculate the RMSE of the model Y_t **end**;	…… …… 5. Build the forecasting model of Y_t on selected factors of leading indicators based on the ARDL model, where the lags of the dependent variable and explanatory variables have been predetermined. …… . **end**;

Algorithm 4a: *Forecast* Algorithm
Input: The set of factors is chosen for the last time, the forecasting model of Y_t is built based on the chosen factors;
Output: $\hat{Y}_{t+h}$: the h-step-ahead forecasts made at the time *t* for the variable Y_t;
begin
1. Build an auxiliary forecasting model for the factors in the Y_t variable forecasting model based on the autoregressive model with quadratic trend AR(p);
2. Perform out-of-sample the h-step-ahead forecasting for the factors using corresponding auxiliary forecasting models;
3. Calculate $\hat{Y}_{t+h}$ using the forecasting model of Y_t
end;
	Algorithm 4b: *Calculate* algorithm
	Input: The set of factors is chosen for the last time, the forecasting model of Y_t is built based on the chosen factors.
	Output: $\hat{Y}_{t+h}$: the h-step-ahead forecasts made at the time *t* for the variable Y_t, $(1 \leq h \leq \text{p})$;
	begin
	1. Calculate $\hat{Y}_{t+h}$ using the forecasting model of Y_t at the time *t*.
	end;

Specifically, the meanings of the command lines of algorithms, functions, and procedures are elaborated on pages 80 - 86 of the Thesis.

The estimation of the computational complexity of the unconditional and conditional forecasting algorithms will be presented in the next section below.

3.2.3 Computational complexity

3.2.3.1 Computational complexity of the CONF algorithm

Let m and N denote the number of variables and observations in the input dataset $\mathbf{X_t}$, respectively; q is the number of iterations of the KTPCA dimension reduction method, and the model builds on the extracted factors.

The computational complexity of the conditional forecasting algorithm is dependent on the computational complexity of: (1) the *FeatureSelection* algorithm (line 3) in the CONF algorithm, (2) the computation of the kernel matrix (with the kernel function being either dot product or not dot product) (line 6 or line 11), (3) the *FeatureLearning* procedure (line 7 or line 12), and (4) the *Forecast* algorithm at line 16, refer to details on pages 86-88 in the Thesis.

- The computational complexity of the *FeatureSelection* algorithm is: $O(m^2)$ (3.2)
- The computational complexity of command lines 7 and 8 is $O(N.m^2 + N^3)$ (3.3)
- The computational complexity of command lines 12 and 13 is $O(N.m^2 + N^3 + m^3)$.
- Since there are q such loops, the computational complexity of lines 10 to 16 is:

$$q.O(N.m^2 + N^3 + m^3). \quad (3.4)$$

- The computational complexity of the *Forecast* algorithm in command line 17 of the CONF algorithm (P.M. Tan al. et., 2018), the computational cost to build such a model is $O((s+2)^2.N + (s+2)^3) = O(N)$, here s is the optimal lag length of exogenous variables, and there are 02 trend variables: *tr* and tr^2. Moreover, the computational complexity of the *Forecast* algorithm is $p.O(N) = O(N)$ (since p is minimal) (3.5)

From (3.2), (3.3), (3.4), and (3.5), the computational complexity of the CONF conditional forecasting algorithm is derived as $q.O(N.m^2 + N^3 + m^3)$. (3.6)

3.2.3.2 Computational complexity of the UNCONF algorithm

The unconditional forecasting algorithm differs from the conditional algorithm mainly in the *LeadingIndicatorSelection* and *Calculate* algorithms. As the computational cost of *Calculate* is negligible compared to *FeatureLearning* algorithms, it can be disregarded.

For each explanatory variable, the computational cost determining whether this variable is a Granger cause with s lag of the dependent variable is $O((2s+1)^2.N + (2s+1)^3 = O(N)$, since s is fixed and small (P.M.Tan al. et., 2018). Therefore, the computational complexity of the *LeadingIndicatorSelection* algorithm is:

$$O\big(m.O(N)\big) = O(m.N) \quad (3.7)$$

Following a similar argument as with the CONF algorithm, we determine that the complexity of the UNCONF algorithm is $q.O(N.m^2 + N^3 + m^3)$. Consequently, the complexity of the forecasting algorithm, encompassing both unconditional and conditional forecasting, is $q.O(N.m^2 + N^3 + m^3)$ (3.8)

3.3 Forecasting export turnover using the KTPCA dimensionality reduction method

3.3.1 Identifying the forecasting problem

With increasing international integration, the factors influencing Vietnam's export turnover are becoming more numerous and diverse. The collection of such data is becoming easier and more complete, thanks to advancements in information technology. How to forecast Vietnam's export turnover amidst the multitude of influencing factors is the motivation for this Thesis, to study the application of the conditional and unconditional forecasting model utilizing the dimensionality

reduction method based on the kernel trick proposed in Chapter 2 to forecast Vietnam's monthly export turnover.

The problem that needs to be solved in this section is forecasting Vietnam's monthly export turnover by considering all potential domestic and foreign factors (variables) that affect Vietnam's export activities.

3.3.2 Factors impacting export turnover and data collection

3.3.2.1 Factors impacting export turnover

One commonly utilized model for forecasting export turnover is the export demand model. This model operates under the assumption of infinitely elastic supply, implying that any amount of supply can be generated in response to demand. Within export demand models, most variables such as exchange rates, price indices, and relative prices of exported goods are incorporated, among which relative price is one of the factors holding significant importance, determining the competitiveness of export activities; comparative advantage is presented in detail according to the theoretical framework in the study (Siggel, 2006). Specifically, research (Siggel, 2006) proposes a general form of the forecast model for total export turnover as follows:

$$X_t = \boldsymbol{f}\left(X_{t-i,} ED_{t-i+1,} ER_{t-i+1,} \boldsymbol{P_{t-i+1}}\right), i \geq 1 \qquad (3.2)$$

where X_t is the value of merchandise and service exports (expressed in nominal or real terms), ED_t is a composite measure of external demand, ER_t is the exchange rate (nominal or real), and $\boldsymbol{P_t}$ is the price vector, creating price dynamics for the goods group in the international market.

3.3.2.2 Data

Researchers (Siggel, 2006), (Stoevsky, 2009), (Lehmann, 2015) have suggested the types of data required for forecasting Vietnam's monthly export turnover. In this Thesis, the actual dataset employed for forecasting Vietnam's export turnover by month is a dataset of 161 explanatory variables, including hard and soft variables, called EXP, encompassing factors influencing export turnover as per the export demand model (Siggel, 2006), (Stoevsky, 2009) are presented in Table 3.1, pages 92 - 93.

3.3.3 Unconditional forecast of export turnover

The EXP dataset is extensive. To make forecasts on such a set, economic forecasters often select only a handful of leading indicators with high statistical significance to include in the unconditional forecast model of export turnover. It is evident that forecasting accuracy using such an approach will be limited since numerous variables influence changes in export turnover but have not been included in its forecasting model. This limitation can be easily overcome by employing the unconditional forecast algorithm on large datasets using the proposed kernel-based dimensionality reduction method. Presented below are presented the intermediate results obtained from applying the forecasting algorithm to unconditionally forecast Vietnam's monthly export turnover.

3.3.3.1 Stage 1: Data preprocessing

Addressing missing values, transforming data - handling non-stationarity according to formula (3.10) (Eskin & Gusev, 2009). Using the formula (3.10) to transform data also aids in handling the non-stationarity of time series variables.

To conduct the forecast and validate the constructed model, the dissertation divides the input dataset of 65 observations into 02 sets: a training set comprising 62 observations from February

2014 to March 2019, and a testing set containing 03 observations from April 2019 to June 2019. Initially, the unconditional forecast model is constructed using the training dataset.

3.3.3.2 Stage 2: Determining leading indicators

Test the stationarity of the export variable (denoted as EX) and 161 explanatory variables using the training dataset.

Execute lines 3 to 5 in the *LeadingIndicatorSelection* algorithm to test the Granger causality between the dependent variable EX and each explanatory variable with an optimal lag of 06, determined based on economic theories as suggested (Wooldridge, 2016). Utilizing a threshold $\alpha < 0.1$, implying a probability of rejection is < 0.1, the Thesis identifies 37 variables as leading indicators of the dependent variable. The statistically significant leading indicators for the export variable EX are listed in Table 3.2 below; refer to page 98 in the Thesis.

3.3.3.3 Stage 3: Extract factors and build models

First, the dataset of 37 leading indicators is mean-centered, and the minimum average distance between 02 column vectors on this data set is calculated as $\rho_0^2 = 0.3273 \approx e^{-1,12}$. Factors are extracted using the KTPCA# method based on the 06 corresponding kernel functions listed in the first column of Table 3.4 below, and the optimal lag is chosen to be 06 as suggested in (Wooldridge, 2016) for economic and financial datasets at monthly frequency. With a chosen cumulative eigenvalue percentage threshold of ≥ 75%, Table 3.4 below presents the number of selected factors, the cumulative eigenvalue percentage, and the RMSE of the unconditional forecasting model for the export turnover variable, EX.

The first row in Table 3.4 shows the result of factor extraction using the KTPCA method, with the polynomial kernel function being the dot product of two vectors $\boldsymbol{\kappa_1}(x_i, x_j) = \langle x_i, x_j \rangle$, indicates the necessity of selecting 12 factors to achieve the cumulative percentage of variance above threshold of 77.01%. We cannot build a forecasting model for EX export turnover according to model (1.16) on the 12 selected factors, because with a general optimal lag of 6, the number of variables in the EX forecasting model = 12* 6 (number of lagged variables) + 6 (lagged variables of EX) = 78 variables. In contrast, the number of observations of the EXP dataset is only 63.

Table 3.4: Results of factor extraction using the KTPCA# method

Kernel	**Kernel function**	**The number of factors**	**% Cumulative eigenvalue**	***RMSE***
Polynomial	$\boldsymbol{\kappa_1}(x_i, x_j) = \langle x_i, x_j \rangle$ *hay PCA*	12	77,01	Not continue
	$\boldsymbol{\kappa_2}(x_i, x_j) = \langle x_i, x_j \rangle^{\mathbf{2}}$	2	83,34	0.0228
	$\boldsymbol{\kappa_3}(x_i, x_j) = \langle x_i, x_j \rangle^{\mathbf{3}}$	1	74,83	0.0270
Gaussian	$\boldsymbol{GA_4}$: $\rho^2 = e^{-0.4}$	5	76,03	0.0202
	$\boldsymbol{GA_5}$: $\rho^2 = e^{-1,12}$	9	75,20	Not continue
	$\boldsymbol{GA_6}$: $\rho^2 = e^{-1,2}$	10	76,41	Not continue

Table 3.4 shows that among the six experimental kernel functions, the most suitable one is the Gauss kernel function GA_4 with parameter $\rho^2 = e^{-0,4}$, with the model's RMSE of 0.0202 being the lowest, and 05 factors are selected to replace the dataset of 37 leading indicators. Testing the stationarity of the 05 factors shows that all factors are stationary. The model unconditionally

forecasts Vietnam's monthly export turnover based on 05 factors is in the form of equation (3.11); refer to page 100 in the Thesis.

3.3.3.4 Stage 4: Implement forecasts

Forecasting model acceptance testing is conducted on the testing dataset comprising 03 observations in the months of April 2019, May 2019, and June 2019 using the Calculate algorithm with model (3.11). It can be observed that according to model (3.11) and solely based on the leading indicators in the current month, we can only forecast EX export turnover for the following month, specifically April 2019.

Unconditional forecasts of Vietnam's export turnover for April 2019, May 2019, and June 2019, generated in this manner, are compared with the actual statistical values of export turnover for these months, and compared with the forecast results obtained from some other typical univariate models, including the AR(6) model, the ARIMA model, and the Holt-Winter model. The results are presented in Table 3.5 below, where the symbol EXF denotes the forecast value of EX by model (3.11), and univariate models AR, ARIMA, and Holt–Winter.

Table 3.5: Comparison of forecasted export turnover results from various models with actual figures.

Model		**Proposed model**		**AR(6)**	
Month	*EX*	*EXF*	% forecast error	*EXF*	% forecast error
04/2019	20439.83	20299.12	0.69	18891.92	7.57
05/2019	21904.59	21173.66	3.34	20724.46	5.39
06/2019	21427.77	21418.12	0.05	20211.47	5.68
		$RMSE_{OUT}$ = **429.79**	Abs(% forecast error) TB = **1.36**	$RMSE_{OUT}$ = 1325.16	Abs(% forecast error) TB = 6.21
Model		**ARIMA(2, 1, 2)**		**Holt-Winter Add**	
Month	*EX*	*EXF*	% forecast error	*EXF*	% forecast error
04/2019	20439.83	19238.68	5.88	19389.46	5.14
05/2019	21904.59	21213.68	3.15	20644.72	5.75
06/2019	21427.77	20958.26	2.19	20349.69	5.03
		$RMSE_{OUT}$= ***844.70***	Abs(% forecast error) TB = ***3.74***	$RMSE_{OUT}$= 1133.26	Abs(% forecast error) TB = 5.31

In this context, % forecast error is calculated as 100*(actual statistical value - forecast value)/actual statistical value. Abs(% error db)TB is the average of the absolute value of the percentage of forecast error across 3 months of April, May, and June 2019; refer to pages 101 - 102 in the Thesis.

Table 3.5 illustrates that the forecasting accuracy of the unconditional export turnover forecast model, constructed based on the factorial ARDL model, according to the unconditional algorithm using the proposed dimensionality reduction method, significantly surpasses the forecast accuracy of the univariate unconditional forecasting models AR(p), ARIMA, and Holt-Winter.

The average forecast error % for the 03 months of April 2019, May 2019, and June 2019, obtained from the export turnover forecast model using the unconditional algorithm, exceeds the % forecast error of the unconditional forecast model built based on the best univariate forecast model ARIMA(2,1,2) by 2.38 percentage points, enhancing forecast accuracy to 63.6%. Therefore, we can

accept the constructed forecast model and use this model for forecasting export turnover for subsequent out-of-sample months, such as July 2019.

3.3.3.5 Out-of-sample forecast of export turnover

The data for export turnover forecasting includes observations up to June 2019. To forecast export turnover for the following month, follow these steps:

- Updated additional observations up to June 2019 for 05 factors extracted from the 37 leading indicators using the KTPCA method with Gauss kernel function with parameter $\rho^2 = e^{-0.4}$.

- Re-estimate model (3.11) using the leading factors with observations up to June 2019;

- Use the newly re-estimated model (3.11) to forecast Vietnam's export turnover for July 2019.

To execute these tasks using the unconditional forecasting process, simply repeat the Stages 3 and 4 with the Gauss kernel function with parameter $\rho^2 = e^{-0.4}$ in the same way as described above.

3.3.4 Conditional forecast of export turnover

The EXP dataset mentioned above is also used to forecast export turnover for the upcoming months, such as the next three months.

3.3.4.1 Stage 1: Data preprocessing

Similar to the case of unconditional forecasting.

3.3.4.2 Eliminate noisy and redundant variables

On the training dataset, with relevance and redundancy thresholds of 0.2 and 0.9, respectively, using the FeatureSelection algorithm, only 63 variables are identified as relevant and non-redundant for the purpose of forecasting export turnover, EX. These variables are presented in Table 3.6 on page 104.

3.3.4.3 Stage 3: Factor extraction using KTPCA# method

With the cumulative eigenvalue percentage threshold ≥ 75% and an overall optimal lag of the factors in the estimation model, determined as suggested in (Wooldridge, 2016), and is 6. The results of factor extraction using the KTPCA# method are presented in Table 3.7.

Table 3.7: Factor extraction using KTPCA# method

Kernel $\boldsymbol{\kappa}$	*Parameters*	*The number of Factors*	*%Cumulative eigenvalue*	*RMSE*
(PCA)	$\kappa_0(,): \mathrm{c} = 0, \mathrm{d} = 1$	14	76.72	Not continue
	$\kappa_1(,): \mathrm{c} = 0, \mathrm{d} = 2$	5	76.02	0.0153
Polynomial	$\kappa_2(,): \mathrm{c} = 0, \mathrm{d} = 3$	2	81.97	0.0270
	$\kappa_3(,);\ \rho^2 = 0.569$	10	75.56	Not continue
Gauss	$\kappa_4(,): \rho^2 = 0.833$	6	76.16	0.0104
	$\kappa_5(,): \rho^2 = 0.500$	12	76.09	Not continue

Table 3.7 also shows that the kernel function $\kappa_4(X_i, X_j)$ is the most suitable among the experimented kernel functions because the RMSE of the predictive model for the variable EX, using the factors selected by KTPCA with this kernel function, attains the smallest, equal to 0.0104, and the parameter ρ^2 in this kernel function is not the minimum average distance between two

column vectors in the input dataset. Following the iteration process, we obtain the optimal forecast model for export turnover in the form of equation (3.13) on page 106.

3.3.4.4 Stage 4: Building a forecast model for exogenous variables and conducting forecasts

a. Forecasting the factors in the built forecast model

The auxiliary forecast model for the factors in the model (3.13) is constructed based on the AR(p) model with trends, as per equation (3.1). Table 3.8 below presents the forecast results for 06 factors in the months of April, May, and June 2019; refer to page 107 in the Thesis.

b. Building an export demand model and forecast the exogenous variables in the model

To compare and evaluate forecast accuracy for export turnover (EX) using the proposed conditional forecast model, the Thesis also forecasts EX using the forecast model constructed based on the export demand model introduced in section 3.3.2.1.

Testing the stationarity of the variables ER, ED, POIL, PRICE_VN, and PEX/PWEX reveals that they are all stationary time series. The forecast model for export turnover is derived from the export demand model using equation (3.9) with a common optimal lag of 6 (Wooldridge, 2016). The forecast results of export turnover using the export demand model in April, May, and June 2019 are presented in Table 3.11.

c. Implementing forecasting testing, comparison, and evaluation

The symbols EXF and DEXF represent the forecast values of EX using the factor model and the export-demand model, respectively. The forecast results of EX for April, May, and June 2019 using these two approaches, along with the actual statistical values, are presented in Table 3.10 below.

Table 3.11 Comparison of forecasted export values with actual values

	Proposed model			***Export demand model***	
Month	*EX*	*EXF*	% forecast error	*DEXF*	% forecast error
04/2019	20439.83	20051.57	1.90	19757.77	3.34
05/2019	21904.59	21603.89	1.37	21464.56	2.01
06/2019	21427.77	21203.48	1.05	22246.80	-3.82
	Abs(% forecast error) TB = **1.44**			Abs(% forecast error) TB = **3.06**	
RMSE	0.0104			0.0261	
$RMSE_{OUT}$	0.0038			0.0296	

Calculate the average of the absolute value of the percentage error of export turnover forecast for the 03 months of April, May, and June 2019 using the proposed conditional forecasting model and the export demand model under the same assumed conditions that the factors affecting exports in April, May, and June 2019 do not have unusual fluctuations. The forecast accuracy of the model constructed based on a conditional algorithm surpasses that of the export demand model by 1.62 percentage points, and enhancing forecasting accuracy up to 52.9%.

3.3.4.5 Forecasting export turnover and building forecast scenarios

a. Out-of-sample forecast of export turnover

Similar to the out-of-sample forecasting of export turnover using the unconditional forecasting model approach, conditionally forecasting this variable entails the following steps:

- Update and supplement the observations until June 2019 for the 06 factors extracted from the 63 explanatory variables using the KTPCA method with a Gaussian kernel function parameter $\rho^2 = e^{0.833}$.

- Re-estimate the model (3.14) with the factors that have observations until June 2019;

- Forecast the factors in the model (3.14) for the next 03 months.

- Utilize the updated model (3.14) and the forecast results of the factors in that model to forecast Vietnam's export turnover for the next 03 months.

It is understood that forecasting using quantitative models acknowledges that the future will resemble the present and the past. However, in reality, this may not always hold true, especially in the context of today's economic globalization. There are numerous unpredictable fluctuations that can impact Vietnam's export activities. To address this reality when conducting conditional forecasting, three commonly used approaches are employed; refer to pages 111-113 in the Thesis.

3.4 Conclusion

Based on the time series modeling process presented in Chapter 1, this chapter proposed a process and forecasting algorithm (unconditional and conditional) for large time series datasets using the dimensionality reduction method proposed in Chapter 2. The computational complexity of this algorithm was also estimated to be polynomial.

The dimension reduction in the algorithm is proposed to be done using both feature selection and feature learning methods. The feature selection method is built based on the Granger causality relationship for the unconditional forecasting algorithm or the Pearson correlation coefficient measure for the conditional forecasting algorithm. The feature learning method is KTPCA#..

Chapter 3 also presents the application of unconditional and conditional forecasting algorithms on large time series data sets to forecast Vietnam's monthly export turnover. The forecast accuracy (unconditional and conditional) of Vietnam's export turnover is quite high, showing that the proposed dimensionality reduction forecasting algorithm can be applied to forecast export turnover as well as forecast other economic-social indicators on large time series datasets..

The research findings related to this chapter are published in studies [CT1], [CT2], [CT4], and [CT5] in the list of author's publications.

CONCLUSION

1. Research results of the Thesis

The Thesis focuses on solving the limitations of PCA and SPCA methods on large time series datasets. The Thesis makes the following main research contributions:

1. *Theoretical contributions*

- Proposing a dimensionality reduction method based on the kernel trick, abbreviated as KTPCA. It is a natural extension of the PCA method and overcomes its limitations in reducing the dimensionality of datasets that do not approximate a hyperplane. The dimensionality reduction performance of the KTPCA method, based on the RMSE-best model (referred to as KTPCA#), equals or surpasses that of PCA, SPCA, RSPCA, and ROBSPCA methods on datasets with similar or mixed sampling frequencies. Additionally, the Thesis demonstrates that the dimensionality reduction performance of PCA and SPCA methods is competitive. This result is different from the

long-standing belief that the dimensionality reduction performance of the SPCA method and its developed versions is equal to or superior to the PCA method.

- Proposing a procedure and algorithm for conditional and unconditional forecasting on large time series datasets using the proposed dimensionality reduction method. The computational complexity of this algorithm is a third-degree polynomial of the number of observations and variables in the input dataset. Comparing this forecasting procedure with the three-step forecasting approach in (Chinn et al., 2023) (considered the most superior forecasting method currently), the initial two steps of the proposed forecasting procedure using the dimensionality reduction method outperform the corresponding two steps in the three-step forecasting approach, while the remaining third step has not yet been compared.

2. *Practical application contributions*

The application of the forecasting algorithm employing the proposed dimensionality reduction method to forecast Vietnam's monthly export turnover, utilizing a dataset of 161 time-series explanatory variables, reveals that:

- The forecast error percentage of the conditional export turnover forecasting model using the proposed algorithm is lower than that of the export demand forecasting model by 1.62 percentage points, enhancing the forecasting accuracy by up to 52.9% compared to the export demand model. In the fields of economics and finance, the export demand model is widely adopted and regarded as superior for forecasting export turnover in countries, including Vietnam.

- The forecasting error percentage of the unconditional export turnover forecasting model using the proposed algorithm is lower than the forecasting error percentage of the ARIMA (2,1,2) model (the best forecasting model among univariate models constructed based on ARIMA, AR(p), and Holt-Winter models) by 2.38 percentage points, improving the forecasting accuracy by up to 63.6% compared to the ARIMA (2,1,2) model.

These outcomes, along with the computational complexity of the forecasting algorithms being a third-degree polynomial, indicate the prospects of applying the proposed forecasting procedure and algorithm employing the dimensionality reduction method in forecasting not only export turnover but also various other economic and financial indicators on large time series datasets.

The thesis results have been published in domestic and international journals and conferences with reviews.

2. Limitations of the Thesis

The Thesis has the following primary limitations:

1. Firstly, the forecasting algorithms (conditional and unconditional forecasting) using the proposed dimensionality reduction method and their applications have only been proposed for datasets with uniform sampling frequency, not for datasets with mixed sampling frequency.

2. Secondly, the forecasting algorithm based on the proposed procedure has only been partially computerized, lacking fully computerization. This limits the proposed forecasting procedure using the dimensionality reduction method to forecast economic and financial indicators on large time series datasets.