Hai Nguyen MINH, PhD Candidate
E-mail: nguyenminhhaidhcn@iuh.edu.vn
Faculty of Fundamental Science
Industrial University of Ho Chi Minh City
Associate Professor Thanh Do VAN, PhD
E-mail: dvthanh@ntt.edu.vn
Faculty of Information Technology
Nguyen Tat Thanh University
Associate Professor Dung Nguyen DUC, PhD
E-mail: nddung@ioit.ac.vn
Institute of IT, Vietnam Academy of Science and Technology

BUILDING EXPORT FORECAST MODEL USING A KERNEL-BASED DIMENSION REDUCTION METHOD

Abstract. *The purpose of this article is to build Vietnam's export turnover forecast model at the monthly frequency on a large data set of potentially time-series predictors. The model is built based on the Dynamic factor model, where factors are extracted from the data set of non-redundant and relevant variables in potential predictors by a variable dimension reduction method using kernel tricks and based on an RMSE-best model. The results show that the percentage of absolute error of out-of-sample forecasts by the built model is less than 2%, and the out-of-sample forecast accuracy of this model is much higher than that of the model built based on the export demand model.*

Keywords*: Big data, Time series, Dimensionality reduction, Kernel trick, Factor model, PCA, Export demand model.*

JEL Classification: C51, C53, E17

1. Introduction

The export turnover forecast is one of the important macroeconomic forecasting contents. With highly open economies like Vietnam, the import and export forecast is even more important and necessary. In the field of macroeconomics, there are many methods to forecast export turnover. Classical theoretical models for forecasting exports include the Ricardian model, the technology-based model, the Heckscher-Ohlin resource factor model, the endogenous growth models, and the gravity model (Bussière et al., 2005). In these models, the factors of foreign trade, return to scale, monopolistic competition, market diversity, as well as market failure are considered to be included in the forecast model (Feenstra, 2015). Mehta and Mathur (2003) reviewed existing models for short-term Forecasting of India's exports where export turnover depends

DOI: 10.24818/18423264/56.1.22.06

on import demand for domestically produced products of the partner countries, on the exchange rate and export price index along with their lagged variables. Bussière et al. (2005) analyzed the total trade of Central and Eastern European (CEE) countries with the euro area countries using the enhanced gravity model, where total trade is the sum of total exports and total imports. The results show that the total trade of CEE countries has a highly statistically significant positive relationship with real exchange rates, with the economic size (GDP) of neighboring countries, countries with similar languages, and member countries in a trade union while geographical distance is not related to total trade.

One of the most commonly used models to forecast export turnover is the export demand model. This model assumes that supply is infinitely elastic, i.e., when demand exists, any supply can be produced. In the export demand model, most variables such as exchange rate, price indices, and the relative price of export are used, where the relative price is one of the very important determinants of export performance competitiveness, and comparative advantage are presented in detail according to the theoretical framework in the study (Siggel, 2006). Stoevsky (2009) used the export demand model to forecast total exports, where the predictors in the model include the price index of exports of goods and services (nominal or real prices), exchange rate (nominal or real), a composite index that measures external demand for domestically produced products, a price vector of the economy's key exports and their lagged variables. Lehmann (2015) has made a forecast of the import and export turnover indices of most European countries (including 18 Eastern and Central European countries, 28 EU countries) to study the economies of these countries. Variables based on the survey (called soft variables) such as business confidence index, consumer confidence index, production expectation index, were included in the import and export demand models to see if the forecast errors are lower if compared with the model consisting of only hard variables such as price variables and price indices, aggregate demand variables, exchange rates. In most countries, the results show that adding soft variables to the import and export forecast models by the demand model approach gives a much lower forecast error. However, the inclusion of both hard and soft variables into the import and export demand model makes the number of predictors increase, and in many cases, the model estimation by multivariate regression is not feasible.

The new model of international trade of OECD countries (Pain et al., 2005) detailed the re-estimates and re-characterizations of international trade volumes and price equations on a panel dataset made in the Department of Economics of OECD. It aims to analyze and forecast international trade. In which, the model for forecasting the export turnover of countries is the vector error correction model, where the predictors in the model are the trade price index of OECD countries and non-OECD countries, private investment, government investment, and exchange rate. This model has contributed to successfully monitor the global consistency of international trade forecasts for world economy prospects.

In a data-rich or big data environment, many other variables also affect an economy's export turnover, such as industrial production, inventories, and credit debt balances of manufacturing industries, domestic and international political-economic situation.

It can be seen that the more variables that are non-redundant and relevant ones to export turnover in a forecast model, the better because then by capturing more information, the forecast accuracy by the model will be higher.

Dynamic factor model is a statistical technique first proposed by Geweke (1977) and later by Stock and Watson (2002), which allows building forecast models on data sets of a large number of scalar time series predictors, in which factors are extracted from these data sets by variable dimension reduction method. Guichard and Rusticelli (2011) have built an export turnover forecast model on a large number of socio-economic predictors using the dynamic factor model, where factors are extracted from the data set of those variables using the PCA method. Many experiments have shown that the forecast accuracy of models built based on the dynamic factor model, in which the factors extracted from the data set of predictors using the PCA method is higher than that of many other Benchmark models where only a few predictors selected based on economic theories are inputted into the models (Guichard and Rusticelli, 2011; Baffigi et al., 2004; Kim and Swanson, 2018). The PCA method is evaluated to be highly effective for real-world economic-financial data analysis (García-Gallego and Mures-Quintana, 2016).

PCA is a typical unsupervised learning method for extracting factors from a large dataset of scalar time series predictors. Each principal component factor is a linear combination of the set of predictors with the weights that are elements of an eigenvector of the covariance matrix of the original data set. Then a few factors will be selected to replace the original predictors in forecast models. PCA is the effective variable dimension reduction method if the data points of the input data set are approximately hyperplane (Van Der Maaten and Postma, 2009). In the opposite case, this method may no longer be such, and then it is no longer appropriate to extract the principal component factors.

Although there are many nonlinear dimensional reduction methods to transform variables in high dimensional space into lower dimensional space (Sarveniazi, 2014), there are few methods to efficiently reduce the variable dimension for data sets of scalar time series predictors whose data points may or may not approximate a hyperplane. The kernel PCA method (or KPCA for short) is one of very few such methods (Schölkopf et al., 1998). The KPCA method is a natural extension of the PCA method. The main idea of this method is to map data points in the input space into another space with a higher dimension number (called the feature space) such that in this feature space, the mapped data points are approximately a hyperplane and then the PCA method can be used to reduce the variable dimension of the mapped dataset.

It is arduous to explicitly define the feature space as well as the mapping from the input space to the feature space, and instead of explicitly defining such

mapping and feature space, Schölkopf et al.(1998) used kernel tricks to implement the aforementioned idea. Namely, the nonlinear principal components are projections of the original data set mapped and mean-centered in feature space onto the eigenvectors of the kernel matrix of this original data set (Schölkopf et al., 1998). KPCA method has been applied to reduce the dimensionality of data sets in pattern analysis exercises (Kung, 2014), especially in exercises on image recognition (Kim et al., 2005).

In a recent study, we have shown that as a natural extension of the PCA method, the KPCA method is only used to reduce the observation dimension, not the variable dimension, and this method can be used to reduce the variable dimension, but then it is not a natural extension of the PCA method. In short, the KPCA method is not suitable for variable dimension reduction in forecasting exercises on large data sets of scalar time series predictors. In that study, we also proposed a variable dimension reduction method based on kernel tricks as another natural extension of the PCA method. It is called the KTPCA method. Unlike the KPCA method, the principal component factors in the KTPCA method are projections of the mean-centered original input data set in the input space onto the eigenvectors of the kernel matrix of that dataset. Experiments showed that the variable dimension reduction performance of the KTPCA method based on an RMSE-best model is superior to the methods of PCA, sparse PCA, randomized sparse PCA, and robust SPCA. Here the variable dimension reduction performance is measured by the RMSE of the forecast model.

This article will apply the KTPCA method based on an RMSE-best model for building a forecast model of export turnover on a large data set of time series predictors at the monthly frequency. More specifically, the export turnover forecast model is built based on the Dynamic factor model, where the factors are extracted using the KTPCA method based on an RMSE-best model. Forecasts of in-sample and out-of-sample by the built model are compared with forecasts by the export demand model, which is considered the most widely used model to forecast the export turnover of a nation (Stoevsky, 2009).

The structure of this article is as follows: Following this section, section 2 presents some preliminaries relating to the next sections. The description of the data set of potential time-series predictors used to build the export turnover forecast model is introduced in Section 3. Section 4 presents the data pre-processing and the method of building the forecast model of Vietnam's monthly export turnover. The comparison of the forecast accuracy by the built model under the proposed method and by the export demand model is presented in section 5, and finally, section 6 presents some conclusions and discussions.

2. Preliminaries

2.1. Dynamic factor model

Suppose $\mathbf{X} = [X_1, X_2, \ldots, X_N]$ is a data set of a large number of scalar time series predictors, PC_i (i =1, ...d and d≪ N) are the principal component factors extracted from **X**. Suppose Y, PC_i , $i = 1, \ldots, d$ are all stationary time series. The

dynamic factor model (Stock and Watson, 2002; Panagiotelis et al., 2009) for forecasting the dependent variable Y according to the predictors X_i ($i = 1, \dots, N$) can be defined as:

$$Y = \sum_{i=1}^{d} \sum_{h=0}^{r_i} a_{i_h} PC_i(-h) + c + u_t \quad (1)$$

where u_t is the residual assumed to be white noise; c and a_{i_h} are estimated parameters; $PC_i(-h)$ is the variable PC_i lagged h steps; r_i ($i = 1, \dots, d$) is the optimal lag of the variable PC_i.

If it is only to improve the forecast accuracy then it is necessary to add into the model (1) some lagged variables of the dependent variable, but then the assessment of the impact of some predictors on the dependent variable is restricted (Stock and Watson, 2002). With the expectation that the built forecast model has both forecast ability and the ability to evaluate the impact of some important predictors on the dependent variable, choosing the dynamic factor model (1) to develop forecast models is the more appropriate. The way to evaluate the impact of some predictors on the dependent variable by a quantitative forecast model is to compare the forecast scenario having the impact of several economic shocks and the forecast scenario without any effects of any economic shocks (Thanh, 2019).

2.2. KTPCA method based on RMSE-best model

Suppose $\mathbf{X} = [X_1, X_2, \dots, X_N]$ is the mean-centered data set of predictors, which means $\sum_{j=1}^{m} x_{i_j} = 0$, where $X_i = (x_{i_j}) \in \mathbb{R}^m$ and i =1...N. Assuming the kernel $\boldsymbol{\kappa}$ is a positive definite function, then the kernel matrix $\mathbf{K} = [< \mathbf{k}(X_i), \mathbf{k}(X_j) >]$ is a positive definite symmetric matrix, and then the eigenvalues of $\mathbf{K}$ are all positive. Assume the eigenvalues are sorted by descending value. The symbol $\mathbf{E}$ is the matrix of the eigenvectors of $\mathbf{K}$ as columns. The KTPCA variable dimension reduction method said that the set of d factors corresponding to the first d eigenvalues extracted from the original data set has the form as:

$$\mathbf{PC}_{m \times d} = \mathbf{X}_{m \times N} \mathbf{E}_{N \times d} \quad (2)$$

Similar to the KPCA method, so far, there is still no criterion to choose the best kernel function for variable dimension reduction, the variable dimension reduction using the KTPCA method must be a trial and error process with selection criteria to be the RMSE in-sample of a forecast model built by regressing the dependent variable on factors extracted by the KTPCA method. Here RMSE in-sample and out-of-sample of a forecast model, respectively, are defined by:

$$RMSE_{IN} = \sqrt{\frac{1}{m} \cdot \sum_{j=1}^{m} (Y_j - \hat{Y}_j)^2} \quad (3)$$

$$RMSE_{OUT} = \sqrt{\frac{1}{p} \cdot \sum_{j=1}^{p} (Y_{t+j} - \hat{Y}_{t+j})^2} \quad (4)$$

in which, $\hat{Y}_i$ is the fitted value by the forecast model for the dependent variable Y_t at time point i; $\hat{Y}_{t+1}, \hat{Y}_{t+2}, \dots$, and $\hat{Y}_{t+p}$ are out-of-sample forecasts of Y_t using this model at time point t. The KTPCA method based on an RMSE-best model can be briefly present as follows:

1) First, perform the extraction of factors using the PCA method on the input data set **X**, thereby determining:

- **PC** $= \{PC_i, \mathrm{i} = 1, \dots, \mathrm{d}\}$ is a set of d factors as the first principal components PC_i such that the cumulative eigenvalue percentage of the first d eigenvalues to the sum of the eigenvalues corresponding to these d factors is: 100* $\frac{\sum_{i=1}^{d} \lambda_i}{\sum_{i=1}^{N} \lambda_i} \geq \alpha$ %, normally $70 \leq \alpha \leq 90$, this implies that the chosen d factors can at least capture $\alpha\%$ of the information in the input data set **X.**
- The forecast model of the export turnover variable Y on the set of chosen factors has the form $\hat{Y} = \widehat{\boldsymbol{F}}_{\boldsymbol{1}}(PC_i), i = 1, \dots, d$ and is estimated under the model (1), here the optimal lag is determined as in (Wooldridge, 2016). Accordingly, the common optimal lag is either 1 or 2 for annual frequency data, either 4 or 8 for quarterly frequency data, either possibly as 6, 12, or even 24 for monthly frequency data depending on the number of observations.
- Save the chosen set of factors (in the variable **PC**), the estimated model (the variable $\widehat{\boldsymbol{Y}}$) and $RMSE_{IN}$ of this model ($RMSE$).

2) Repeat the following steps:

- Select a positive definite kernel function **κ** and calculate the kernel matrix **K** $= [< \boldsymbol{\kappa}(X_{\mathrm{i}}), \boldsymbol{\kappa}(X_{\mathrm{j}}) >]$, where X_{i} is the ith input predictor, or the ith column vector of the data matrix **X.**
- Perform the PCA method on the kernel matrix **K,** with the same contents as the PCA method on the covariance matrix of **X** but with a difference that the d factors chosen are projections of the data set **X** onto d eigenvectors of the kernel matrix **K** according to the formula (2). Then, identify new factors (***FPC***), the new estimated model of *Y* on newly chosen factors ($\boldsymbol{F\widehat{Y}}$) as well as $RMSE_{IN}$ of this model.
- Compare new $RMSE_{IN}$ to saved ***RMSE***, if new $RMSE_{IN}$ < saved ***RMSE***, then $\boldsymbol{PC} \leftarrow \boldsymbol{FPC}, \widehat{\boldsymbol{Y}} \leftarrow \boldsymbol{F\widehat{Y}}$, and $\boldsymbol{RMSE} \leftarrow RMSE_{IN}$ to replace sets of saved factors, the saved estimated model, and the saved RMSE, respectively.
- This process can be repeated many times depending on the user. At the end of this process, we obtain the export turnover forecast model with the highest forecast accuracy ($RMSE_{IN}$ is the smallest) as well as the most suitable kernel function among the experimented kernel functions.

In our recent study, we developed an algorithm for this process. To be able to forecast automatically using the algorithm, in this algorithm we assumed that the common optimal lag for all the variables in a forecast model in each different loop is defined the same and is predetermined under the way of the study (Wooldridge, 2016). In application practice, the most popular used kernels are polynomial ones $\boldsymbol{\kappa}(X_i, X_j) = (< X_i, X_j > + \mathrm{c})^{\mathrm{d}}$, Gaussian kernels $\boldsymbol{\kappa}(X_i, X_j) = \exp(\frac{-\|X_i - \mathrm{X_j}\|^2}{2\rho^2})$ (Schölkopf and Smola, 2003; Kim et al., 2005), where the parameter ρ^2 is taken around value $c.\frac{1}{N}\sum_{i=1}^{N} min_{i \neq j} \|\mathrm{X}_i - \mathrm{X_j}\|^2$, here c is the tuning parameter (Rathi et al., 2006).

2.3. Export demand model

Suppose variable Y_t, ED_t, ER_t, P_t are stationary variables. The general export demand model is defined by (Stoevsky, 2009):

$$Y_t = f(Y_{t-i}, ED_{t-i+1}, ER_{t-i+1}, P_{t-i+1}), \text{ i} = \{1, 2, 3, ...\} \quad (5)$$

where Y_t are the total exports (expressed in nominal or real value), ED_t is a composite measure of foreign demand, ER_t is the exchange rate (nominal or real), and P_t is a price vector, which creating price dynamics for the goods and services in the international market. The index ED_t is calculated as follows: $ED = \sum n_k * IM_k$, where n_k is the import share of the kth country in the forecasted country's total exports; IM_k is the import growth rate of the kth country.

3. Data

The data set used to build the forecast model for the monthly export turnover of Vietnam is rather large, including the data of the predictors in the export demand model and of many other predictors which are potentially related to the export turnover. Therefore, in these predictors there may be unnecessary predictors as well as may not even be related to the change in export turnover and therefore cannot be included in the export turnover forecast model. There are 161 such potential scalar time series predictors. The values of these predictors can be relative numbers (%) or absolute numbers. Data of the predictors taking absolute values are collected from January 2013 to June 2019, while the data of remaining predictors are collected from February 2014 to June 2019. So the observation number of the relative number value predictors is 65. The set of all potential predictors introduced in Table 1 below includes the name of a potential predictor, frequency, usage meaning of variable, and its data source. The set of potential predictors include all the predictors used in the export demand model, such as the nominal exchange rate between VND and USD (*ER*), the world price of Vietnam rice (*PRICE_VN*) and of crude oil (*POIL*), the variables used to build the foreign aggregate demand (*ED*) index of the Vietnamese economy include 17 variables of import turnover at current prices of 16 partner countries having the largest import share and the rest of the world.

The potential predictors reflect three main aspects of the economy: the supply and demand sides of the economy and market strength (Eskin and Gusev, 2009). Specifically, the predictors that reflect the supply side include variables on the industrial production index in some economic sectors, investment from the state budget and foreign direct investment, the inventory index in some manufacturing industries, the credit balance of several economic sectors, the export turnover of some important manufacturing and processing industries, export price index of goods and services of Vietnam, purchasing manager index (*PMI*), ... The predictors reflecting the demand side include the import turnover of some economic sectors and the whole economy, the consumption index in some manufacturing and processing industries, total retail sales of goods and services of the whole economy and some economic sectors, gold and dollar price indices, the consumer price inflation of the whole economy and some baskets of goods and services, the world

price of Vietnam's rice and Thailand's rice, the world price of robusta coffee, rubber and copper, the import turnover of the main trading partner countries of the economy and total import turnover of the world. Predictors reflecting the market strength include some domestic and international stock indices; the exchange rate between VND and the Yuan to USD; the short, medium, and long-term deposits and lending interest rates; total means of payment, and deposits from institutions and private sectors. Data of potential predictors in Table 1 includes both hard data (statistical data) and soft data (collected through surveys) such as the purchasing manager index PMI.

Table 1. Potential predictors

Predictors	Freq.	Usage meaning	Source
10 indicators of industrial production in some economic sectors	M	Reflect the supply side	GSO
02 variable state budget investment and state budget revenue variable	M		GSO
04 variables of foreign direct investment	M		Fiinpro
19 inventory index in some manufacturing and processing sector	M		Fiinpro
07 outstanding credit variables in some economic sectors	M		Fiinpro
26 variables of export turnover of some manufacturing industries and total export turnover of the economy	M		GSO
01 variable of the export price index of Vietnam	M		GSO
01 purchasing manager index.	M		Markit economics
19 variables of import turnover of 18 manufacturing industries and the whole economy;	M	Reflec.the demand side	Fiinpro
19 consumption indicators in the manufacturing and processing sector;	M		GSO
05 total retail sales of consumer goods and services in some economic sectors;	M		Fiinpro
10 general consumer price inflation indices of the economy and some goods	M		GSO
02 gold and USD price indices;	M		Fred
05 is the variable world price of rice from Vietnam, Thailand, Robusta coffee, rubber, and copper.	M		Fred
01 the export price index of the world	M		Fred
Import turnover of 16 major trading economic partners and the rest of the world	M		Europa
04 is domestic and world stock indices;	M	Reflec. the strength of the market	Cophieu68
02 is exchange rates between VND and yuan to USD;	M		Fred
05 is short, medium, and long term deposit and lending interest rates;	M		Fiinpro
03 is the total means of payment, deposits from institutions and individuals.	M		Fiinpro

Table 1 also shows data for all variables collected monthly from 6 different sources, of which the two main providers are the General Statistics Office of Vietnam (GSO) and the company FiinPro that specializes in providing financial and business data services.

3. Forecast model building methods

The method of building a forecast model of export turnover on the potential predictors in Table 1 is presented in Figure 1 below. Accordingly, it is a multi-step process and is briefly presented in Figure 1 below.

3.1. Step 1: Data pre-processing

Data pre-processing needs to be done before proceeding with other contents related to building a forecast model of export turnover. The content of data pre-processing is to add missing data and convert the data into a comparable form compared to the same month of the previous year.

- *Add missing values*: (i) If the missing value occurs in the first or last observations or both of them, the solution is to use the AR(p) autoregression model with the trend or use a regression model of this predictor on several of its high relevant predictors according to economic theory. (ii) Conversely, if the missing value does not belong to (i), the remedy used is the interpolation method or the exponential smoothing method depending on whether the missing values are more or less.

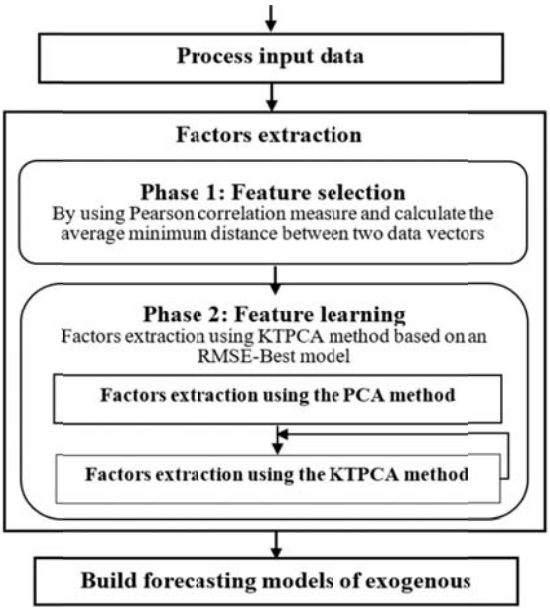

Figure 1. Schema of building the forecast model of export turnover

- *Data conversion*: most of the predictors receive relative numerical values compared to the same month of the previous year, such as industrial production,

consumption, and inventory indices in some manufacturing and processing industries, but there are also many other predictors whose values are absolute numbers at current (or nominal) prices that do not imply comparability with the same period last year such as total retail sales of consumer goods and services, import and export turnover, world prices of some goods and products, exchange rates, and some predictors take absolute numerical values implying comparability as stock indices, purchasing manager index, ...

To ensure the consistent meaning of the data, all variables that receive absolute numerical values are converted to corresponding variables that receive relative numerical values compared to the same month of the previous year according to the formula (Eskin and Gusev, 2009):

$$DLOG(X_t^*) = LOG(\frac{X_t}{X_{t-12}}) - LOG(\frac{X_{t-1}}{X_{t-13}}) \quad (6)$$

here $X_t^* = \frac{X_t}{X_{t-12}}$; X_t is variable to receive absolute numeric value; LOG: is the logarithm of base e, t is a time point.

3.2. Step 2: Factor extraction

***Phase 1**: Feature selection.* Due to the data set of potential time-series predictors may also contain redundant or noisy information. To improve the forecast accuracy of the built model, it is necessary to first remove from the input data set redundant or noisy information or both. This article will use the Pearson correlation coefficient measure to eliminate redundant or noisy information from the original data set. Suppose X, $Y \in \mathbb{R}^m$ are two scalar time-series variables, $R(X,Y)$ is the Pearson correlation coefficient between these two variables. Suppose Y, X_1, X_2 are time series predictors; Y is the dependent variable; X_i (i = 1,2) are predictors; α and β in [0, 1] are called relevant and redundant thresholds, respectively. X_1 is relevant to Y if $|R(Y,X_1)| \geq \alpha$. X_2 is called redundant if there exists X_1 such that $|R(Y,X_1)| \geq |R(Y,X_2)|$ and $|R(X_1,X_2)| \geq \beta$.

The main content of the *Feature selection* phase to eliminate redundant and noisy information in the data set of potential predictors. The elimination of such information is implemented by discarding irrelevant or redundant predictors to export turnover using the Person correlation measure. The dataset of remaining predictors is mean-centered and used as the input data set for extracting factors. The average minimum distance between two column vectors in the mean-centered input data set is calculated and used to choose a more suitable parameter ρ^2 in Gaussian kernels when using the KTPCA variable dimension reduction method.

***Phase 2**: Feature learning:* The main content of this phase is to extract factors from the input data set by using the KTPCA method based on an RMSE-best model as presented in section 2.2 above. At the end of step 2, we determine the most optimal kernel function in the experimented kernels, and the forecast model of export turnover has the lowest $RMSE_{IN}$ among the models built based on the dynamic factor model, where factors are extracted using the KTPCA method.

The variable dimension reduction using the KTPCA method based on an RMSE-best model as well as the building of the estimation model is possible in the

environment of languages such as Python or R. Based on Kernlab packages (Karatzoglou et al., 2004) and Caret (Kuhn and others, 2008), the article built a forecast model under the proposed method in the environment of R language.

3.3. Step 3: Forecast exogenous variables and perform forecasts

To receiving the future value of the export turnover, it is necessary to forecast the future values of the factors in the forecast model. The factors are often forecasted by the ARIMA model or the AR(p) model with deterministic trends. This article uses the AR(p) method with deterministic trends to following equation (6) to build forecast models of the factors.

4. Result and Evaluation

4.1. The forecast model based on the dynamic factor model

Step 1: *Data pre-processing*

At the end of this step, we receive a data set of 161 potential predictors that have been processed with missing data and converted to a form comparable to the month of the same period last year according to formula (6). The input data set includes 65 observations and is divided into 02 sets, the training set includes 62 observations from February 2014 to March 2019, and the testing set includes 03 observations from April 2019 to June 2019. Building the export turnover forecast model using the KTPCA method based on an RMSE-best model is performed on the training set.

Step 2: *Factor Extraction*

With the relevant and redundant thresholds of 0.2 and 0.9, respectively, at the end of Phase 1, we receive 63 variables that are highly relevant and non-redundant to *EX*. These variables are the input predictors to build the forecast model of export turnover. The average minimum distance between two data vectors in the input data set is 0.569 (= ρ^2).

Assume that the criterion for selecting the number of extracted factors is their cumulative eigenvalue percentage and the maximum lag of the factors in the estimation model is determined according to the experience of the article (Wooldridge, 2016) and is 6. Then, with the cumulative eigenvalue percentage threshold of 75%, the results of factor extraction using the KTPCA method are presented in Table 2, where the first line is the results using the PCA method. This line shows that the number of chosen factors is 14. Because the number of observations of the input data set is 62, so it is not possible to regress the variable *EX* on the 14 chosen factors with their common maximum lag of 6. For the Gaussian kernel functions $\kappa_3(X_i, X_j)$ and $\kappa_5(X_i, X_j)$, we also get the same results. The slow increase of cumulative eigenvalue percentage when increasing the number of principal components extracted using the PCA method, as shown in Table 2, implies that the input data set is not approximately a hyperplane (Stock and Watson, 2002).

Table 2. Factor extraction results using KTPCA method

Kernel κ	*Parameters*	*Factors*	*Cum. Eig. Percentage*	$RMSE_{IN}$
(PCA)	$\kappa_0(,)$: c = 0, d = 1	14	76.72	Not continue
Polynomial	$\kappa_1(,)$: c = 0, d = 2	5	76.02	0.0153
	$\kappa_2(,)$: c = 0, d = 3	2	81.97	0.0270
	$\boldsymbol{\kappa_3(,);\ \rho^2 = 0.569}$	10	75.56	Not continue
Gaussian	$\kappa_4(,)$: $\rho^2 = 0.833$	6	76.16	0.0104
	$\kappa_5(,)$: $\rho^2 = 0.500$	12	76.09	Not continue

Table 2 also shows that the kernel $\kappa_4(X_i, X_j)$ is the most suitable among the experimented kernels because the $RMSE_{IN}$ of export turnover forecast model on the chosen factors is the smallest and equal to 0.0104 and the ρ^2 parameter in this kernel is not average minimum distance of 2 column vectors in the input data set. At the end of Step 2, we get the export turnover forecast model built based on the dynamic factor model (the dynamic factor model for short) as follows:

$$EX = -0.111PC1^{***} + 0.023PC2^{**} - 0.029PC2(-4)^{***} - 0.017PC2(-5)^{**} + 0.030PC3(-1)^{**}$$
$$(0.015) \quad (0.010) \quad (0.008) \quad (0.007) \quad (0.013)$$
$$+0.045PC3(-2)^{***} - 0.034PC4^{***} + 0.020PC4(-3)^{**} - 0.044PC4(-6)^{***} - 0.030PC5(-3)^{***}$$
$$(0.013) \quad (0.008) \quad (0.009) \quad (0.008) \quad (0.007)$$
$$-0.029PC5(-5)^{***} + 0.026PC6(-3)^{***} + 0.018PC6(-5)^{*}$$
$$(0.009) \quad (0.010) \quad (0.010)$$
$$R^2: 0.9068 \quad \text{D-W stat: } 2.3369 \quad \text{SMPL: } 56 \text{ after adjustments} \qquad (7)$$

Asterisks indicate the statistical significance of the t-Statistic: *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Step 3*: Build forecast models of exogenous variables*

Forecast models of the 06 factors in the model (7) are estimated according to Equation (1) by the OLS regression method**.** Table 3 below presents forecasts for the 06 factors in April, May, and June of 2019.

Table 3: Forecasts for the 06 factors

Months	*PC1F*	*PC2F*	*PC3F*	*PC4F*	*PC5F*	*PC6F*
19-Apr	0.1315	-0.0287	0.1855	-0.2030	-0.0265	0.1604
19-May	0.0124	-0.0039	-0.1255	0.0209	0.0106	-0.0695
19-Jun	0.0543	0.0113	0.0849	0.0666	0.0016	-0.0419

4.2. The forecast model based on the export demand model

Vietnam's economy has deeply integrated into the world economy and the openness of this economy is very high. Due to the relative export price is an important determinant determining the export performance of the economy (Siggel,

2006), so the relative export price needs to be added to the export demand model defined by Equation (5). In this article, we determine the PEX/PWEX as the relative export price of the economy, where PEX and PWEX are the export price index of Vietnam and the world, respectively.

The unit root test of the *ER*, *ED*, *POIL*, *PRICE_VN*, *PEX/PWEX* predictors shows that they are all stationary time series. The export turnover forecast model based on the export demand model (the export demand model for short) is estimated with the common optimal maximum lag of 6 (Wooldridge, 2016). This model has the form:

$$EX = -0.99EX(-1)^{***} - 0.77EX(-2)^{***} - 0.3EX(-3)^{***} - 0.2EX(-4)^{***} + 1.91ER(-1)^{**}$$
$$(0.124) \quad (0.168) \quad (1.180) \quad (1.099) \quad (0.816)$$
$$+ 2.36ER(-3)^{***} + 1.78ER(-4)^{**} + 1.56ER(-5)^{*} - 2.55ER(-6)^{***} - 0.17ED^{***} - 0.10ED(-1)^{*}$$
$$(0.848) \quad (0.798) \quad (0.927) \quad (0.915) \quad (0.047) \quad (0.055)$$
$$- 0.13ED(-2)^{***} + 0.07ED(-4)^{*} + 0.13ED(-5)^{***} + 0.58PEX/PWEX^{***}$$
$$(0.047) \quad (0.039) \quad (0.039) \quad (0.057)$$
$$- 0.31PEX/PWEX(-1)^{***} - 0.27PEX/PWEX(-3)^{***} + 0.14*POIL^{***} + 0.16POIL(-1)^{***}$$
$$(0.093) \quad (0.075) \quad (0.078) \quad (0.041)$$
$$- 0.10POIL(-5)^{**} + 0.14PRICE\_VN(-3)^{*} + 0.34PRICE\_VN(-4)^{***} \quad (8)$$
$$(0.039) \quad (0.078) \quad (0.101)$$

R^2: *0.9367* D-W stat*: 2.0113* SMPL*: 56* after adjustments

The forecast models of predictors in the model (8) are presented in Table 4.

Table 4: Forecasts of predictors

Months	*ERF*	*EDF*	*POILF*	*PRICE_VNF*	*PEX/PWEX*
19-Apr	-0.0009	0.0011	-0.0067	-0.0169	0.9820
19-May	-0.0002	-0.0018	0.0024	-0.00463	0.9461
19-Jun	-0.0005	0.0009	0.0055	0.0045	0.9877

4.3. Evaluation of the quality of the forecast model

The Analysis of the Jarque-Bera statistical test and the Kurtosis of the EX variable, 06 factors in the dynamic factor model and the 05 predictors in the export demand model in Table 7 below shows that the probability distribution of *PC3, PC4, PC5, PC6* in the dynamic factor model and *POIL, PRICE_VN* in the export demand model can be considered as a normal distribution, while the remaining factors and predictors are not. The Skewness also shows that the tail of the probability distribution of *EX, PC4, PC5, ER, POIL,* and *PRICE_VN* is to the left, while the tail of the probability distribution of the remaining factors or predictors is on the right. On the other hand, all the exogenous variables in these models are stationary time series. This ensures that there is no spurious regression when building the export turnover forecast model based on these two different approaches.

The *EXF* and *DEXF* represent the fitted variables of *EX* by the dynamic factor model and the export demand model, respectively. Values of *EXF* and *DEXF*

at three months of April, May, and June of 2019 are calculated using the (7) and (8) models, respectively, here the values of the exogenous variables in the two models are introduced in Tables 3 and Table 4. Comparing the forecasts for the variable *EX* in April, May, and June of 2019 using the two forecast models mentioned above with the actual statistical values are presented in Table 5.

Table 5: Compare export turnover forecast results with reality

		The dynamic factor model		*The demand export model*	
Months	*EX*	*EXF*	forecast error %	*DEXF*	forecast error %
19-Apr	20439.83	20051.57	1.90	19757.77	3.34
19-May	21904.59	21603.89	1.37	21464.56	2.01
19-Jun	21427.77	21203.48	1.05	22246.80	-3.82
$RMSE_{IN}$		0.0104		0.0261	
$RMSE_{OUT}$		0.0038		0.0296	

here, the forecast error % is the percentage of (actual statistical value - fitted value) divided by the actual statistical value.

Thus, the absolute forecast error % and the $RMSE_{OUT}$ of the dynamic factor model are always smaller than the export demand model. This shows that the out-of-sample forecast accuracy of the dynamic factor model is higher than that of the export demand model. On the other hand, the fluctuation trend of the actual export turnover (*EX*) and forecasted export turnover by the dynamic factor model (*EXF*) are the same, while that by the export demand model is not so.

5. Conclusion and Discussion

Based on the dynamic factor model, this article has built the export turnover forecast model by month from the data set of 161 potential predictors, where factors are extracted from the input data set of these predictors using the KTPCA method based on an RMSE-best model. The out-of-sample forecasts by the dynamic factor model (Table 5) show that the absolute forecast error percentages by this model are less than 2%, and the forecast accuracy by the dynamic factor model is always higher than that by the export demand model. This result is not surprising because the set of predictors in the dynamic factor model includes the predictors in the export demand model. Moreover, the factors also captured most of the information in the original set of the predictors.

The article shows that if the common maximum lag of predictors at monthly frequency is determined as in (Wooldridge, 2016), then the PCA method is not suitable to reduce the variable dimension of the data set of selected input predictors, and the KTPCA method based on an RMSE-best model overcomes this limitation.

This article also shows that data pre-processing, such as overcoming missing data and dealing with the seasonality of data by converting absolute numerical values to relative numerical values compared to the same month of the

previous year, are very important. Such data conversion is a way of transforming an economic-financial time series into a stationary time series. That helps to avoid spurious regressions during forecast model building. With a very large set of potential predictors, it is difficult to avoid redundant and/or noisy information in the set of predictors, so it is necessary to eliminate such information before building forecast models. The variable selection method by eliminating redundant predictors and/or irrelevant predictors to the fluctuations of export turnover using the Pearson correlation coefficient measure is simple but effective in forecasting exercises on large datasets of economic-financial time-series predictors.

REFERENCES

[1] **Baffigi, A., Golinelli, R., Parigi, G. (2004),** ***Bridge Models to Forecast the Euro Area GDP***. *International Journal of Forecasting*, 20(3), 447–460;
[2] **Bussière, M., Fidrmuc, J., Schnatz, B. (2005),** ***Trade Integration of Central and Eastern European Countries: Lessons from a Gravity Model.*** *ECB Working Article, No.* 545;
[3] **Eskin, V., Gusev, M. (2009),** ***High-frequency Forecasting Model for the Russian Economy.*** *The Making of National Economic Forecasts* (Edited by L. R. Klein), Edward Elgar Publishing Limited, 93-120;
[4] **Feenstra, R. C. (2015),** ***Advanced International Trade: Theory and Evidence***. *Princeton University Press*;
[5] **García-Gallego, A., Mures-Quintana, M.-J. (2016),** ***Principal Components and Canonical Correlation Analyses as Complementary Tools. Application to the Processing of Financial Information***. *Economic Computation and Economic Cybernetics Studies and Research*, 50(4), 249-266; *ASE Publishing;*
[6] **Geweke, J. (1977),** ***The Dynamic Factor Analysis of Economic Time Series***. *Latent variables in socio-economic models*. *North-Holland*;
[7] **Guichard, S., Rusticelli, E. (2011),** ***A Dynamic Factor Model for World Trade Growth***. *OECD Economics Department Working Articles*, No.874. Doi:10.1787/5kg9zbvvwqq2-en;
[8] **Karatzoglou, A., Smola, A., Hornik, K., Zeileis, A. (2004),** ***Kernlab-an S4 Package for Kernel Methods in R.*** *Journal of statistical software*, 11(9), 1–20;
[9] **Kim, H. H., Swanson, N. R. (2018),** ***Methods for Backcasting, Nowcasting and Forecasting Using Factor-MIDAS: With an Application to Korean GDP***. *Journal of Forecasting*, 37(3), 281–302. Doi:10.1002/for.2499;
[10] **Kim, K. I., Franz, M. O., Scholkopf, B. (2005),** ***Iterative Kernel Principal Component Analysis for Image Modeling***. *IEEE transactions on pattern analysis and machine intelligence*, 27(9), 1351–1366;
[11] **Kuhn, M., others (2008),** ***Building Predictive Models in R Using the Caret Package***. *Journal of statistical software*, 28(5), 1–26;
[12] **Kung, S. Y. (2014),** ***Kernel Methods and Machine Learning***. *Cambridge University Press*;
[13] **Lehmann, R. (2015),** ***Survey-based Indicators vs. Hard Data: What Improves Export Forecasts in Europe?.*** *Ifo Working Article,* No. 196;

[14] **Mehta, R., Mathur, P. (2003),** ***Short-term Forecasting of India's Export: Developing a Framework by Countries and Commodities***. *Research and Information System for the Non-aligned and Other Developing Countries (RIS)*;
[15] **Pain, N., Mourougane, A., Sédillot, F., Fouler, L. (2005),** ***The New OECD International Trade Model.*** *OECD Economics Department Working Articles*, N. 440. Doi:10.1787/680050777016;
[16] **Panagiotelis A., George A., Rob J. H., Bin J., and Farshid V. (2019)**, ***Macroeconomic Forecasting for Australia Using a Large Number of Predictors***. *International Journal of Forecasting* 35 (2): 616–33;
[17] **Sarveniazi, A. (2014),** ***An Actual Survey of Dimensionality Reduction*****.** *American Journal of Computational Mathematics*, 04(02), 55–72; Doi:10.4236/ajcm.2014.42006;
[18] **Schölkopf, B., Smola, A., Müller, K.-R. (1998),** ***Nonlinear Component Analysis as a Kernel Eigenvalue Problem***. *Neural computation*, 10(5), 1299–1319;
[19] **Schölkopf, B. and Smola, A. J. (2003),** ***A Short Introduction to Learning with Kernels***. *Advanced lectures on machine learning*, Springer, 41–64;
[20] **Siggel, E. (2006),** ***International Competitiveness and Comparative Advantage: A Survey and a Proposal for Measurement*****.** *Journal of Industry, competition, and trade*, 6(2), 137–159;
[21] **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;
[22] **Stoevsky, G. (2009),** ***Econometric Forecasting of Bulgaria's Export and Import Flows*****.** *Bulgarian National Bank Discussion Articles DP*/77/2009;
[23] **Thanh, D. Van (2019),** ***Macro-econometric Model for Medium-Term Socio-Economic Development Planning in Vietnam. Part 2: Application of the Model*****.** *Journal Ekonomika Regiona*, 15(3), 695-706. Doi 10.17059/2019-3-6;
[24] **Van Der Maaten, L., Postma, E. (2009),** ***Dimensionality Reduction: A Comparative Review*****.** *Journal of Machine Learning Research*, 10, 66–71;
[25] **Wooldridge, J. M. (2016),** ***Introductory Econometrics: A Modern Approach*****.** Nelson Education.