Generalized dynamic principal component for monthly nonstationary stock market price in technology sector

Yusrina Andu$^1$, Muhammad Hisyam Lee$^{1,*}$ and Zakariya Yahya Algamal$^2$

$^1$Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia, 81310 Johor Bharu, Johor, Malaysia
$^2$Statistics and Informatics Department, University of Mosul, Mosul, Iraq

E-mail: *mhl@utm.my

Abstract. The majority of stock market price is nonstationary, while only few have stationary pattern. It is noted that past researches usually transformed the stock market price into stationary prior to analysis which may lead to the loss of data originality. Thus, a direct application of the nonstationary stock market price is of main interest in this study, as such generalized dynamic principal component (GDPC) performs the analysis directly without transformation. As well as, Brillinger dynamic principal component (BDPC) were also used on the nonstationary stock market price for comparison. This dataset consists of the most recent five-year monthly observations of six different regions in technology sector. Stationarity test was performed prior to the application and the analyses were carried out based on the reconstruction of lags of the time series. The results showed that the GDPC for six stock market prices have lower mean squared error compared to BDPC. Also, the percentage of explained variance in the first component were much higher in GDPC. Thus, this indicated that GDPC model is more suitable for prediction compared to its counterpart.

1. Introduction

Stock market price are the most concern issue that has been researched by different field of study[1], [2]. However, the stock market price is not always in a stationary pattern as they might be affected by interrelated factors such as political events, economic variables (i.e. interest rates, exchange rates, monetary growth rates and commodity prices) and traders’ expectations[3]–[6]. Under the assumptions of stationarity, linearity and normality, most of the statistical approach listed above may be restricted within the area of stock market price prediction.

Throughout the years, several studies have been carried out on nonstationary time series via principal component analysis[7], [8]. However, prior to analysis, most of these studies transforms the data to stationary form. Among the common approach of stationary transformation is first difference and log-transformation.

Nevertheless, a direct application to nonstationary stock market price were recently proposed in [9]. They developed a principal component method based on nonstationary which is known as generalized dynamic principal component (GDPC). In their proposed methodology, two approaches were projected of which: 1) it is entirely data-analytic and does not assume any given model; and 2) it does not assume a fixed number of factors to be identified, instead the number of components is chosen to achieve a preferred degree of accuracy in the reconstruction of the original series. Moreover, the method that was
proposed are able to take any data such as not necessarily having the combination of a linear or stationary pattern. Though in this case, GDPC is more prevail with nonstationary compared to stationary data. Furthermore, GDPC approach may perhaps achieve a possible nonstationary behaviour of the series. As for present time, the GDPC methodology proves to be a better alternative in handling the data directly without need of prior stationary transformation.

2. Literature review on generalized dynamic principal component

Over the years, principal component analysis is commonly linked to factor model although both are actually quite distinct from each other. One of the distinct feature is the way the variance is analysed, in factor model, only those that have similar variances are accounted for. Meanwhile, as for principal component, all the observed variance has to be considered in the analysis. Both factor model and principal component are able to encapsulate the main sources of variation and covariation among the number of $n$ predictor variables[10]. Due to this similarity, most of the advance method that is introduced in factor model are also adaptable in principal component analysis. Worth noting that the static factor models were then extended by [11] where later known as dynamic factor models. Meanwhile [12] has extent principal component as dynamic principal component (DPC).

In static factor models, time series and a small number of factors has been assumed contemporaneously related with each other. For instance, previous studies have assumed on two types of patterns in factor models, which is some may be stationary[13]–[16] while some are nonstationary. The covariance matrix eigenvalues of the lags have been used for principal component computation in these models. These nonstationary can be seen either as having integrated approach[17], general nonstationary[18] and locally stationary[19].

Similar assumption follows DPC, as such in dynamic factor models, one important segment of the original series can be interpreted in a dynamic way only with a comparatively small number of common factors. A very general dynamic factor model has been introduced by [20] which enables an infinite number of factor lags and low correlation between any two idiosyncratic components. When the number of series were increase up till infinity, the consistency of the common component increases as well. The projection of the data in the first DPC, $q$, which also includes both leads and lags, enables acquiring the estimation of the series. The principal components here were computed through Fourier transformation of the spectral matrix eigenvectors, as such practice in [12].

Following this model, a one-sided estimation method was later proposed in designing dynamic factor models for forecasting [21]. This model has been further improved from [15] static principal component. [22] later proposed a model that is both, static and dynamic and established estimation methods for the factor structure which has finite-dimensional factor space. A one-sided explanation for the general dynamic factor model proposed in [20] was later extended in [23]. Some of these dynamic factor models generate assumptions on the relation between the studied series and the way factors are loaded [14], [22]. In contrast, the models in [20] and [24] were developed without following this assumption. Therefore, based on the previous literature on factor models and BDPC, [9] model enables to handle nonstationarity directly without transformation.

3. Methodology

3.1. Augmented Dickey-Fuller (ADF) test

In general, a series is said to be stationary if the mean value does not vary within the sampling period [25]. Meanwhile, a series with a time varying mean is a nonstationary series. The existence of nonstationarity occasionally may not be obvious as the low signal power can remain hidden among the strong stationarity of the data [26]. Therefore, the duration of the time series dataset need to be chosen properly to discern the nonstationary in the data. There are several stationarity tests available in order to check whether a time series follows a stationary or a non-stationary process. Among these tests are Augmented Dickey-Fuller (ADF)[27], Phillips-Perron (PP) test[28] and also Kwiatkowski-Phillips-
Schmidt-Shin or KPSS test[29]. Out of these, ADF is the most widely used test. The ADF test was developed to test whether the data or series is stationary or non-stationary based on the presence of unit root. The presence of unit roots implies that the standard distribution theory is not valid. Prior to the estimations, it is important to verify its stationarity by using this test, thus the suitable method can be used for that particular series. Ordinary least squares method is used to obtain the coefficients of a model distribution. Consider an autoregressive AR (1) process

$$y_t = \rho y_{t-1} + \varepsilon_t, \quad \varepsilon \sim (0, \sigma^2)$$

If $\rho = 1$, the equation defines a random walk and $y$ is nonstationary. The null hypothesis for testing nonstationary is $H_0: \rho = 1$. One of the straightforward way to test the $H_0$ is to rewrite the AR (1) equation as

$$y_t - y_{t-1} = (\rho - 1)y_{t-1} + \varepsilon_t$$

$$\Delta y_t = \gamma y_{t-1} + \varepsilon_t$$

making the hypothesis $H_0: \rho = 1$ is equivalent to a test of $H_0: \gamma = 0$. From here, we can test against $\rho < 1 (\gamma < 0)$, with rejection would be on the left side. In this study, the stationarity test was implemented using package “tseries” in R environment.

3.2. Brillinger dynamic principal component (BDPC)

BDPC method is known to provide best estimation for stationary time series dataset. The reconstruction problem is addressed as follows. Suppose zero mean $m$ dimensional stationary process, $\{z_t\}, -\infty < t < \infty$. The dynamic principal components are defined by searching for $m \times 1$ vectors $c_h, -\infty < h < \infty$ and $\beta_j, -\infty < j < \infty$, so that if we consider as first principal component, the linear combination would be

$$f_t = \sum_{h=-\infty}^{\infty} c'_h z_{t-h},$$

then

$$E \left[ \left( z_t - \sum_{j=-\infty}^{\infty} \beta_j f_{t+j} \right)' \left( z_t - \sum_{j=-\infty}^{\infty} \beta_j f_{t+j} \right) \right]$$

is minimum.

This problem was resolved by showing that $c_h$ is the inverse Fourier transform of the principal components of the cross spectral matrices for each frequency, and $\beta_j$ is the inverse Fourier transform of the conjugates of the same principal components [30]. It is worth noted that when this procedure is adapted to finite samples, the number of lags in Equation (3.1) and in the reconstruction of the series should be truncated. Although BDPC can be used with nonstationary series, however, the mean squared error (MSE) of the reconstructed series may not be close to its possible minimum value. In addition, it is unclear how to reconstruct the principal components in BDPC to make it robust.

3.3. Generalized dynamic principal component (GDPC)

This method was developed by reconstructing the vector of time series from BDPC using a finite number of lags. Assume that we observe $z_{jt}, 1 \leq j \leq m, 1 \leq t \leq T$, and consider two integer numbers $k_1 \geq 0$ and $k_2 \geq 0$. Therefore, the first dynamic principal component with $k_1$ lags and $k_2$ leads, as a vector $f = (f_t)_{-k_1+1 \leq t \leq T+k_2}$, in which the reconstruction of series $z_{jt}, 1 \leq j \leq m$, as a linear combination of $(f_{t-k_1}, f_{t-k_1+1}, \ldots, f_{t}, f_{t+1}, \ldots, f_{t+k_2})$ is optimal with the MSE criteria. Given a possible factor $f$, a $m \times (k_1 + k_2)$ matrix of coefficients $f = (f_{ji})_{1 \leq i \leq m, -k_1 \leq i \leq k_2}$, and $a = (a_1, \ldots, a_m)$, the reconstruction of the original series $z_{jt}$ is defined as
\[
\hat{z}_{j,t} = \sum_{i=-k_1}^{k_2} \gamma_{j,i} f_{t+i} + \alpha_j
\]

Let \( k = k_1 + k_2 \) and set
\[
f_t^* = f_{t-k_1}, \quad 1 \leq t \leq T + k, \quad \beta_{j,h}^* = \gamma_{j,h-k_1-1}, \quad 1 \leq h \leq k + 1.
\]
and also define
\[
f_t^{**} = f_{t+k_1}, \quad 1 - k \leq t \leq T,
\]
\[
\beta_{j,h}^{**} = \beta_{j,k+2-h}, \quad 1 \leq h \leq k + 1.
\]

Then, the reconstructed series is obtained as
\[
\hat{z}_{j,t} = \sum_{i=-k_1}^{k_1} \beta_{j,i} f_{t+i+k_i} + \alpha_j = \sum_{h=0}^{k} \beta_{j,h+1} f_t^* + h + \alpha_j = \sum_{h=0}^{k} \beta_{j,h}^* f_t^* - h + \alpha_j
\]

Then without loss of generality, the indistinctly \( k \) lags or \( k \) leads of the principal component to reconstruct the series. Once the forward optimal solution is obtained, the backward obtained as well via Equation (3.4).

Let \( f = (f_1, ..., f_{T+k})' \), \( \beta = (\beta_{j,1})_{1 \leq j \leq m, 1 \leq k+1} \) and \( \alpha = (\alpha_1, ..., \alpha_m) \), through the reconstruction of \( m \) series using \( k \) leads, the MSE loss function is
\[
\text{MSE}(f, \beta, \alpha) = \frac{1}{Tm} \sum_{j=1}^{m} \sum_{t=1}^{T} (\hat{z}_{j,t} - z_{j,t}(f, \beta, \alpha))^2 = \frac{1}{Tm} \sum_{j=1}^{m} \sum_{t=1}^{T} (\hat{z}_{j,t} - \beta_{j,i} f_{t+i} - \alpha_j)^2
\]

(3.4)

It is worth noting that this loss function is well defined and is also acceptable in the case of nonstationary vector time series. The optimal choices of \( f = (f_1, ..., f_{T+k})' \) and \( \beta = (\beta_{j,1})_{1 \leq j \leq m, 1 \leq k+1} \), \( \alpha = (\alpha_1, ..., \alpha_m) \) are defined by
\[
(\hat{f}, \hat{\beta}, \hat{\alpha}) = \arg_{f \in \mathbb{R}^{T+k}, \beta \in \mathbb{R}^{m \times (k+1)}, \alpha \in \mathbb{R}^m} \min \text{MSE}(f, \beta, \alpha)
\]

(3.5)

If \( f \) is optimal, clearly \( \gamma f + \delta \) is optimal too. Hence, \( f \) is chosen in order that \( \sum_{t=1}^{T+k} f_t = 0 \) and \( (1/(T+k)) \sum_{t=1}^{T+k} f_t^2 = 1 \). The first GDPC of order \( k \) of the observed series \( z_1, ..., z_T \) is \( \hat{f} \). The first GDPC of order \( 0 \) corresponds to the first regular principal component of the data.

Let \( C_j(\alpha_j) = (c_{j,t,q}(\alpha_j))_{1 \leq t \leq T+k, 1 \leq q \leq k+1} \) be the \((T + k) \times (k + 1)\) matrix defined by
\[
c_{j,t,q}(\alpha_j) = \begin{cases} (z_{j,t-q+1} - \alpha_j), & 1 \vee (t - T + 1) \leq q \leq (k + 1) \wedge t \\ 0, & \text{otherwise} \end{cases}
\]

where \( a \vee b = \max(a, b) \) and \( a \wedge b = \min(a, b) \). Let \( D_j(\beta_j) = (d_{j,t,q}(\beta_j)) \) be the \((T + k) \times (T + k)\) given by
if \((t - k) \lor 1 \leq q \leq (t + k) \land (T + k)\) and 0 otherwise and

\[
d_{j,t,q}(\beta_j) = \sum_{v=(t-k)\lor 1}^{t\land T} \beta_{j,q-v+1}\beta_{j,t-v+1}
\]

Differentiating Equation (3.5) with respect to \(f_t\), the following equation is derived as

\[
f = D(\beta)^{-1} \sum_{j=1}^{m} C_j(\alpha)(\beta_j)
\]

(3.7)

The coefficients \(\beta_j, 1 \leq j \leq m\), can be obtained using least-squares estimator, that is

\[
\left(\begin{array}{c}
\beta_j \\
\alpha_j
\end{array}\right) = (F(f)^T F(f))^{-1} F(f)^T z^{(j)}
\]

(3.8)

where \(z^{(j)} = (z_{j1, \ldots, z_{jT}})^T\) and \(F(f)\) is the \(T \times (k + 2)\) matrix with \(t\) th row \((f_t, f_{t+1}, \ldots, f_{t+k}, 1)\). Then the first GDPC is determined by Equation (3.8) and Equation (3.9). The second GDPC is defined as the first GDPC of the residual \(r_{jt}(f, \beta)\). Higher order GDPC are defined in a similar manner.

4. Result analysis

Dataset that were used for this study comprise of six technology sector stock market price which were 6888.KL (Malaysia), NOK (Finland), STX(Ireland), TI-A(Italy), TLK(Indonesia) and TU(Canada) obtained from Yahoo Finance. These stock market prices were based on monthly observations of a 5-year period from September 2013 until September 2017. Prior to analysis, stationarity test was performed as to ascertain the nonstationary pattern of the series. It is worth noting that the stock market price is transform to stationary for BDPC, whereas GDPC uses the data directly without the transformation. The six nonstationary stock market price series were plotted as in Figure 1.

Several lags were considered to reconstruct the stock market price series and were tested using both Akaike information criterion (AIC) [31] and Bayesian information criterion (BIC) [32] as has been shown in Table 1, where lower value of AIC indicates a better model fit. Therefore, based on this, the model that is chosen throughout consist of the original stock market price and lags \(k = 4\).

In Table 2, the results showed the MSE and the percentage of explained variance of the stock market price series using lags \(k = 4\). For the purpose of demonstration, the percentage of explained variance is considered for one component only [33]. The reconstruction of five out of the six stock market price showed lower MSE in GDPC [9]. This showed that GDPC performed better as compared to BDPC. In addition, although STX showed higher MSE in GDPC, the percentage of variance is much higher at 85.3% in comparison of BDPC.
Figure 1. Monthly stock market price in technology between September 2013 to September 2017.

Table 1. Information criteria of the six stock market series.

| Stock Price | AIC     | BIC     |
|-------------|---------|---------|
| 6888.KL     | 476.68  | 485.88  |
| NOK         | 472.99  | 481.09  |
| STX         | 475.91  | 484.73  |
| TI-A        | 477.80  | 485.90  |
| TLK         | 478.25  | 486.35  |
| TU          | 478.41  | 486.51  |

Table 2. Mean squared error and the percentage of explained variance of BDPC and GDPC of the stock market price series.

| Stock Price | BDPC | GDPC | BDPC | GDPC |
|-------------|------|------|------|------|
| 6888.KL     | 6.09 | 0.05 | 70.4 | 93.3 |
| NOK         | 6.60 | 0.39 | 77.5 | 83.9 |
| STX         | 9.53 | 19.21| 59.5 | 85.3 |
| TI-A        | 7.07 | 0.42 | 86.1 | 81.0 |
| TLK         | 39.74| 29.25| 95.0 | 60.2 |
| TU          | 5.10 | 1.74 | 69.0 | 64.6 |

Meanwhile, in BDPC, three stock market prices showed a higher percentage of explained variance which were TI-A, TLK and TU at 86.1%, 95% and 69%, respectively. However, out of these two series, there was only a small difference of 5% each as compared to GDPC. Conversely, the overall percentage of explained variance in GDPC is much higher exhibiting almost more than 80% in the first component itself. Thus, much information can be obtained from GDPC.
5. Conclusion
GDPC tackles the nonstationary that exist in the stock market price through the reconstruction of the time series using lags. The direct applications of nonstationary stock market price through GDPC had shown that it performs much better than of the transformed series in BDPC. Furthermore, it also proves to be useful in providing higher percentage of explained variance in the first component of the nonstationary series. Therefore, the findings indicate that GDPC performs much better than BDPC.

Acknowledgements
This study was partially funded by the Ministry of Higher Education, Malaysia through the Research University Grant for Universiti Teknologi Malaysia (grant number 15H64).

References
[1] B. McElroy, T., & Monsell, “Model estimation, prediction, and signal extraction for nonstationary stock and flow time series observed at mixed frequencies,” J. Am. Stat. Assoc., pp. 1284–1303, 2015.
[2] J. Wang, J., & Wang, “Forecasting stock market indexes using principle component analysis and stochastic time effective neural networks,” Neurocomputing, vol. 156, pp. 68–78, 2015.
[3] D. Zhong, X., & Enke, “Forecasting daily stock market return using dimensionality reduction,” Expert Syst. Appl., no. 67, pp. 126–139, 2017.
[4] A. L. Cavalcante, R.C., Brasileiro, R.C., Souza, V.L., Nobrega, J.P. and Oliveira, “Computational intelligence and financial markets: A survey and future directions,” Expert Syst. Appl., no. 55, pp. 194–211, 2016.
[5] M. Göçken, M. Özçalici, A. Boru, and A. T. Dosdoʇru, “Integrating metaheuristics and Artificial Neural Networks for improved stock price prediction,” Expert Syst. Appl., vol. 44, pp. 320–331, 2016.
[6] S. P. Wang, J. Z., Wang, J. J., Zhang, Z. G., & Guo, “Forecasting stock indices with back propagation neural network,” Expert Syst. Appl., vol. 38, no. 11, pp. 14346–14355, 2011.
[7] P. S. Souza Filho, J. B., & Diniz, “A Fixed-Point Online Kernel Principal Component Extraction Algorithm,” IEEE Trans. Signal Process., 2017.
[8] P. Zhao, X., & Shang, “Principal component analysis for non-stationary time series based on detrended cross-correlation analysis,” Nonlinear Dyn., vol. 84, no. 2, pp. 1033–1044, 2016.
[9] D. Peña and V. J. Yohai, “Generalized Dynamic Principal Components,” J. Am. Stat. Assoc., vol. 111, no. 515, pp. 1121–1131, 2016.
[10] M. W. Stock, J. H., & Watson, Forecasting with many predictors., Handbook o. 2006.
[11] J. Geweke, The dynamic factor analysis of economic time series. North-Holland, Amsterdam., 1977.
[12] D. R. Brillinger, “Fourier analysis of stationary processes,” in Proceedings of the IEEE, 1974, vol. 62, no. 2, pp. 1628–1643.
[13] Q. Lam, C., & Yao, “Factor modeling for high-dimensional time series: inference for the number of factors.,” Ann. Stat., vol. 40, no. 2, pp. 694–726, 2012.
[14] S. Bai, J., and Ng, “Determining the Number of Factors in Approximate Factor Models,” Econometrica, vol. 70, pp. 191–221, 2002.
[15] M. W. Stock, J. H., & Watson, “Forecasting using principal components from a large number of predictors,” J. Am. Stat. Assoc., vol. 97, no. 460, pp. 1167–1179, 2002.
[16] G. E. Pena, D., & Box, “Identifying a simplifying structure in time series.,” J. Am. Stat. Assoc., vol. 82, no. 399, p. 836–843., 1987.
[17] D. Peña and P. Poncela, “Nonstationary dynamic factor analysis,” J. Stat. Plan. Inference, vol. 136, no. 4, pp. 1237–1257, 2006.
[18] Q. Pan, J., & Yao, “Modelling multiple time series via common factors,” Biometrika, vol. 95, no. 2, pp. 365–379, 2008.
[19] H. Motta, G., & Ombao, “Evolutionary factor analysis of replicated time series,” Biometrics, vol.
68, no. 3, pp. 825–836, 2012.

[20] L. Forni, M., Hallin, M., Lippi, M., & Reichlin, “The Generalized Dynamic Factor Model: Identification and Estimation,” *Rev. Econ. Stat.*, vol. 82, p. 540–554., 2000.

[21] L. Forni, M., Hallin, M., Lippi, M., & Reichlin, “The Generalized Dynamic Factor Model: One Sided Estimation and Forecasting,” *J. Am. Stat. Assoc.*, vol. 100, pp. 830–840, 2005.

[22] L. Forni, M., Giannone, D., Lippi, M., & Reichlin, “Opening the Black Box: Structural Factor Models with Large Cross Sections,” *Econom. Theory*, vol. 25, pp. 1319–1347, 2009.

[23] M. Forni, M., & Lippi, “The General Dynamic Factor Model: One-Sided Representation Results,” *J. Econom.*, vol. 163, pp. 23–28, 2011.

[24] P. Forni, M., Hallin, M., Lippi, M., & Zaffaroni, “Dynamic Factor Models with Infinite-Dimensional Factor Spaces: One-Sided Representations,” *J. Econom.*, vol. 185, pp. 359–371, 2015.

[25] A. F. Ayinde, O. E., Aina, V. I., Ayinde, K., & Lukman, “Drivers of rice price variation in Nigeria: A two-stage iterative ridge regression approach,” *J. Agric. Sci. Belgrade*, vol. 61, no. 1, pp. 79–92, 2016.

[26] K. Hara, S., Kawahara, Y., Washio, T., Von BüNau, P., Tokunaga, T., & Yumoto, “Separation of stationary and non-stationary sources with a generalized eigenvalue problem,” *Neural networks*, vol. 33, pp. 7–20.

[27] W. A. Dickey, D. A., & Fuller, “Distribution of the estimators for autoregressive time series with a unit root,” *J. Am. Stat. Assoc.*, vol. 74, no. 366a, pp. 427–431, 1979.

[28] P. Phillips, P. C., & Perron, “Testing for a unit root in time series regression,” *Biometrika*, vol. 75, no. 2, pp. 335–346, 1988.

[29] Y. Kwiatkowski, D., Phillips, P. C., Schmidt, P., & Shin, “Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?”, *J. Econom.*, vol. 54, no. 13, pp. 159–178.

[30] D. Brillinger, *Time Series: Data Analysis and Theory*. San Francisco.: Holden-Day, 1981.

[31] H. Akaike, “A new look at the statistical model identification,” *IEEE Trans. Automat. Contr.*, vol. 19, no. 6, pp. 716–723, 1974.

[32] H. Akaike, “A Bayesian extension of the minimum AIC procedure of autoregressive model fitting,” *Biometrika*, vol. 66, no. 2, pp. 237–242, 1979.

[33] M. Hörmann, S., Kidziński, Ł., & Hallin, “Dynamic functional principal components,” *J. R. Stat. Soc. Ser. B (Statistical Methodol.)*, vol. 77, no. 2, pp. 319–348, 2015.