Menemui Matematik (Discovering Mathematics)
Vol. 44, No. 2: 78-85 (2022)

Nonstationary Daily Healthcare Stock Market Price using Non-Transformed Dimensionality Reduction Technique

Yusrina Andu[1*], Muhammad Hisyam Lee[2] and Zakariya Yahya Algamal[3]

[1]*College of Computing, Informatics and Media, Universiti Teknologi MARA Cawangan Negeri Sembilan, 72000 Kuala Pilah, Negeri Sembilan, Malaysia*

[2]*Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia, 81310 Johor Bahru, Malaysia*

[3]*Department of Statistics and Informatics, College of Computer Science and Mathematics, Universiti of Mosul, Mosul, Iraq*

[1*]yusrinaandu@uitm.edu.my, [2]mhl@ utm.my, [3]zakariya.algamal@gmail.com

ABSTRACT

Healthcare stock market price is usually nonstationary. General practice of handling nonstationary stock market price is through transformation process, which may cause loss of data originality. To overcome this, an alternative way of direct handling of the stock market price is of interest. The dimensionality reduction of nonstationary stock market price was performed by using generalized dynamic principal component (GDPC), adapting Brillinger dynamic principal component (BDPC) concept based on the reconstruction of the stock market price. Daily observations of healthcare stock market price were considered for this study. Stationarity test was carried out and the analysis were two-based, transformed and non-transformed. Then, three principal component methods were used to reduce the dimensionality. The results shows that GDPC have a higher percentage of explained variance percentage (above 90%) and lower mean squared error among the other methods. Thus, this shows that a direct application may also achieved better result performance.

Keywords: Nonstationary, Principal Component, Stock Market Price

INTRODUCTION

Healthcare is one of the emerging sectors that has gained importance in the stock market price worldwide. Among the industries that these healthcare companies venture includes drug manufacturers, medical appliances and equipment, specialized health services, medical laboratories & research, and medical insurance services. Nonstationary are found in most stock market price pattern due to several factors such as constitutional events, economic variables (i.e. interest rates, exchange rates, commodity prices) and expectations of traders (Andu et al., 2018). Similarly, healthcare also follows this nonstationary pattern. Transforming these data to stationary is perceived as the popular approach of handling nonstationary. This includes first difference (Crump and Gospodinov, 2022), linear transformation (Sánchez et al., 2015), and log transformation techniques (Andu et al., 2019; Caparole et al., 2019).

One of the approaches of handling nonstationary healthcare stock market price is by reducing its dimension to obtain a meaningful interpretation of the data. Ordinary principal component (OPC) is a well-known dimension reduction technique. This technique can concurrently reduce the dimensionality of the stock market price and retain possible variation in the data as much as possible (Andronicean et al., 2021). OPC uses covariance matrices to describe the relationship of the variables inside the data. Meanwhile, Brillinger dynamic principal component (BDPC) (Brillinger, 1964) is also a dimension reduction technique, developed from the expansion of

principal component. Unlike OPC, BDPC used spectral density matrix instead of covariance matrices. Noteworthy that both OPC and BDPC methodology are mostly used for stationary stock market series. In the case of nonstationary data, stationary transformation needs to be done first before further analysis can be carried out. However, transformation process may cause the loss of data originality (Andu et al. 2019).

To overcome this, a non-transformed principal component was introduced known as generalized dynamic principal component (GDPC) (Peña and Yohai, 2016). This method can be directly used on the nonstationary stock market price without need for stationary transformation process (Andu et al., 2019). The highlights of this method are, 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. GDPC can also be applied to any data pattern but are more focused on nonstationary data. A possible nonstationary behaviour of the time series may also be achievable through this method. Hence, the present study will be carried out to compare the performance of all the three dimension reduction techniques using daily healthcare stock market price. It is envisaged that the non-transformed methodology will be a better alternative in handling the nonstationary data directly. This paper is organized as follows. Section 2 describes the methods used. Section 3 presents the data and result comparison between the methods. Section 4 is the conclusion.

METHODOLOGY

Augmented Dickey-Fuller (ADF) test

Data which have strong stationarity may concealed the low signal power of nonstationarity (Worden et al., 2021). Hence, it is important to wisely select the time series duration to distinguish the nonstationarity in the data. This can be achieved by performing stationarity tests using either Augmented Dickey-Fuller (ADF) (Dickey & Fuller, 1979), Phillips-Perron (PP) test (Phillips & Perron, 1988) or Kwiatkowski-Phillips-Schmidt-Shin or KPSS test (Kwiatkowski et al., 1992). In this study, ADF test is chosen as it can identify the stationarity of the data based on the unit root existence. When unit root exists, it implies that the data is nonstationary.

Ordinary least squares method is used to obtain the coefficients of a model distribution. Let autoregressive AR(1) process as

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

then, if $\rho = 1$, the equation defines a random walk and y is nonstationary. The null hypothesis for testing nonstationary is $H_0: p = 1$, where the rejection would be on the left side. Here, “tseries” package in R was used to perform the stationarity test.

Transformed Ordinary Principal Component

OPC is a popular dimension reduction technique. According to (Andu et al., 2019), the OPC covariance matrix is:

$$\begin{cases} Y_1 = \vec{\alpha}_1^T \cdot \vec{X} = \alpha_{11}X_1 + \alpha_{12}X_2 + \cdots + \alpha_{1n}X_n \\ Y_2 = \vec{\alpha}_2^T \cdot \vec{X} = \alpha_{21}X_1 + \alpha_{22}X_2 + \cdots + \alpha_{2n}X_n \\ \cdots \\ Y_n = \vec{\alpha}_n^T \cdot \vec{X} = \alpha_{n1}X_1 + \ \alpha_{n2}X_2 + \cdots + \alpha_{nn}X_n \end{cases}$$

where X_i is the original variable, Y_i is the principal component and $\vec{\alpha}_i$ is the coefficient vector, respectively. Nonstationary data need to be transformed to stationary before applying OPC. Hence, it might not be able to depict the healthcare stock market price well because of its restrict feature to direct application.

Brillinger dynamic principal component (BDPC)

Another dimension reduction approach which is BDPC works by reconstructing the time series (Brillinger, 1981). Given the zero mean m dimensional stationary process, $\{z_t\}$, $-\infty < t < \infty$, the dynamic principal components can be found for $m \times 1$ vectors $c_h, -\infty < h < \infty$ and $\beta_j, -\infty < j < \infty$. Therefore, the linear combination of the first principal component becomes

$$f_t = \sum_{h=-\infty}^{\infty} c_h' \, z_{t-h,} \tag{1}$$

subsequently

$$E\left[\left(z_t - \sum_{j=-\infty}^{\infty} \beta_j f_{t+j}\right)'\left(z_t - \sum_{j=-\infty}^{\infty} \beta_j f_{t+j}\right)\right] \tag{2}$$

is minimum.

The principal components of the cross spectral matrices are given by c_h, which is the inverse Fourier transform of the principal components for each frequency. Meanwhile, inverse Fourier transform of the conjugates, β_j can be acquire from the same principal components (Brillinger, 1981). It should be noted that BDPC is best used with stationary series and can also with nonstationary series. Despite that, a best minimum mean squared error (MSE) may be difficult to be obtained.

Generalized dynamic principal component (GDPC)

GDPC was developed by reconstructing BDPC vector of time series using a finite number of lags (Peña & Yohai, 2016). The following methods are according to (Andu et al., 2018; Peña & Yohai, 2016). Supposed that $z_{j,t}$, $1 \le j \le m$, $1 \le t \le T$, and the two integer numbers $k_1 \ge 0$ and $k_2 \ge 0$ as the lags and leads. Hence, the first dynamic principal component is vector $\mathbf{f} = (f_t)_{-k_1+1 \le t \le T+k_2}$, of which the series is reconstructed where $z_{j,t}$, $1 \le j \le m$, as a linear combination of $(f_{t-k_1}, f_{t-k_1+1,} \cdots, f_t, f_{t+1}, \cdots, f_{t+k_2})$ is optimum, given the MSE criteria. In addition, **f**, a possible factor of a m $\times$ $(k_1 + k_2)$ matrix of coefficients $\gamma = (\gamma_{j,i})_{1 \le j \le m, -k_1 \le i \le k_2}$, and $\alpha = (\alpha_1, \cdots, \alpha_m)$, the reconstruction of the original series z_j, t is defined as

$$\hat{z}_{j,t} = \sum_{i=-k_1}^{k_2} \gamma_{j,i} f_{t+i} + \alpha_j.$$

When $k = k_1 + k_2$ and fixed

$$f_t^* = f_{t-k_1}, 1 \le t \le T + k, \beta_{j,h}^* = \gamma_{j,h-k_1-1}, 1 \le h \le k + 1.$$
$$f_t^{**} = f_{t+k}^*, \quad 1 - k \le t \le T,$$
$$\beta_{j,h}^{**} = \gamma_{j,k+2-h}, 1 \le h \le k + 1. \tag{3}$$

The reconstruction can be achieved as

$$\hat{z}_{j,t} = \sum_{i=-k_1}^{k} \beta_{j,i} f_{t+i+k_1} + \alpha_j = \sum_{h=0}^{k} \beta_{j,h+1}^* f_{t+h}^* + \alpha_j = \sum_{h=0}^{k} \beta_{j,h+1}^{**} f_{t-h}^{**} + \alpha_j. \tag{4}$$

The k lags or k leads of the principal component can be used to obtain the series reconstruction. Acquiring an optimal forward solution will lead to backward solution as well as shown in Equation (5). On the other hand, MSE loss function is obtained through reconstructing the m series using k leads, by letting $\mathbf{f} = (f_1, \cdots, f_{T+k})', \beta = (\beta_{j,i})_{1 \le j \le m, 1 \le i \le k+1}$ and $\alpha = (\alpha_1, \cdots, \alpha_m)$,

$$\text{MSE}(\mathbf{f}, \beta, \alpha) == \frac{1}{Tm} \sum_{j=1}^{m} \sum_{t=1}^{T} \left(z_{j,t} - \sum_{i=0}^{k} \beta_{j,i+1} f_{t+i} - \alpha_j \right)^2 \tag{5}$$

Meanwhile, the optimal options of $\mathbf{f} = (f_1, \cdots, f_{T+k})'$ and $\beta = (\beta_{j,i})_{1 \le j \le m, 1 \le i \le k+1}$, $\alpha = (\alpha_1, \ldots, \alpha_m)$ are defined by

$$(\widehat{\mathbf{f}}, \hat{\beta}, \hat{\alpha}) = \arg_{f \epsilon \mathbb{R}^{T+k}, \beta \epsilon \mathbb{R}^{m \times (k+1)}, \alpha \epsilon \mathbb{R}^m} \min \text{MSE}(\mathbf{f}, \beta, \alpha) \tag{6}$$

It should be noted that if $\mathbf{f}$ is optimal, similarly $\gamma\mathbf{f} + \delta$ is optimal. Hence, $\mathbf{f}$ is chosen in order that $\sum_{t=1}^{T+k} f_t = 0$ and $(1/(T+k)) \sum_{t=1}^{T+k} f_t^2 = 1$. From $\mathbf{z}_1, \cdots, \mathbf{z}_t$ observations, the first GDPC of order k can be acquired as $\hat{\mathbf{f}}$. Meanwhile, GDPC of order zero represents the first regular principal component.

When $\mathbf{C}_j(\alpha_j) = (c_{j,t,q}(\alpha_j))_{1 \le t \le T+k, 1 \le q \le k+1}$ be the $(T+k) \times (k+1)$ matrix can be expressed as

$$c_{j,t,q}(\alpha_j) = \begin{cases} (z_{j,t-q+1} - \alpha_j) & \text{if } 1 \vee (\text{t} - \text{T} + 1) \le \text{q} \le (\text{k} + 1) \wedge \text{t} \\ 0 & \text{if otherwise} \end{cases} \tag{7}$$

such as $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$. Then, $\boldsymbol{D}_j(\beta_j) = \left(d_{j,t,q}(\beta_j) \right)$ is $(T+k) \times (T+k)$ becomes

$$d_{j,t,q}(\beta_j) = \sum_{v=(t-k) \vee 1}^{t \wedge T} \beta_{j,q-v+1} \beta_{j,t-v+1}$$

if $(t-k) \vee 1 \le q \le (t+k) \wedge (T+k)$ and 0 otherwise and

$$\mathbf{D}(\beta) = \sum_{j=1}^{m} D_j(\beta_j) \tag{8}$$

Differentiating Equation (7) with respect to f_t, the following equation is derived as

$$\mathbf{f} = \mathbf{D}(\beta)^{-1} \sum_{j=1}^{m} C_j(\alpha) \beta_j \tag{9}$$

The coefficients β_j and α_j, $1 \le j \le m$, can be obtained using least-squares estimator, that is

$$\binom{\beta_j}{\alpha_j} = \left(\mathbf{F}(\mathbf{f})'\mathbf{F}(\mathbf{f})\right)^{-1}\mathbf{F}(\mathbf{f})'\boldsymbol{z}^{(j)} \tag{10}$$

where $z^{(j)} = \left(z_{j,1}, \dots, z_{j,T}\right)'$ and $\mathbf{F}(\mathbf{f})$ is the $T \times (k+2)$ matrix with $t-$th row $(f_t, f_{t+1}, \dots, f_{t+k}, 1)$. Finally, Equation (9) and Equation (10) define the first GDPC.

Information Criteria Methods

To describe the methods for model comparison, it is essential to firstly defined its deviance (Kim, 2021). Therefore, two log likelihood models are used in this study, namely Akaike Information Criteria (AIC) (Akaike, 1974) and Bayesian Information Criteria (BIC) (Akaike, 1979). The model formulation which has the smallest value can be achieved using these information criteria methods. The equations that were used to calculate AIC and BIC in all three principal component techniques are shown in Equation (11).

$$\begin{aligned} \text{AIC} &= -2\log L + 2q \\ \text{BIC} &= -2\log L + \log(n)\cdot q \end{aligned} \tag{11}$$

where q is the number of estimated parameters in the model, L is the maximum values of the likelihood function for the model and n is the number of observations.

RESULT ANALYSIS

The data that were used for this study comprise of ten healthcare sector stock market price which are 5225.KL (Malaysia), ABT (U.S.), AZN (U.K.), BSL.SI (Singapore), FMS (German), KANG (China), MDT (Ireland), NVO (Denmark), SKR.BK (Thailand), and SNY (France). These stock market prices were based on daily observations of a three-year period from January 1st, 2015 to January 1st, 2018. These stock market prices are chosen as they are among the major healthcare companies in their respected countries. The companies not only provide medical treatment but also healthcare services and medical technologies. A stationarity test is conducted before analysing to verify the nonstationary pattern of the series (Andu et al., 2019). It is worth noting that the nonstationary healthcare stock market price is directly applied using GPDC. As BDPC and OPC require the data to be stationary, therefore first difference and log transformation are carried out. Figure 1 shows the ten-nonstationary healthcare stock market price series plot.

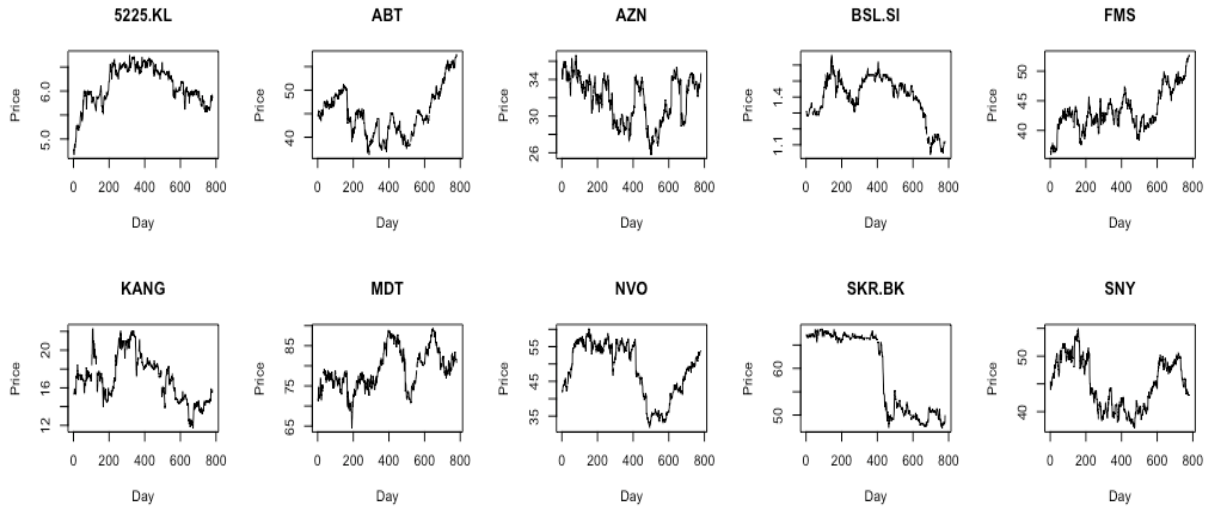

Figure 1. Daily stock market price in healthcare sector between January 1st, 2015 to January 1st, 2018.

The ADF test results show that the test statistics values were higher than the 95% critical values in all healthcare stock market prices. Moreover, the test also failed to reject H_0 indicating the presence of unit root. Therefore, the selected healthcare stock market prices can be concluded as nonstationary. Several lags were considered to reconstruct the stock market price series and were tested using both AIC (Akaike, 1974) and BIC (Akaike, 1979) as has been shown in Table 1, where lower value of AIC indicates a better model fit. The AIC model is preferred over BIC because AIC had lower values in the model indicating better fitted (Andu et al., 2018). Therefore, based on this, the model that is chosen throughout consist of the original stock market price and lags k =3**.**

Table 1. Information criteria of the healthcare stock market price.

Stock Price	**AIC**	**BIC**
5225.KL	10131.97	10150.43
ABT	10215.69	10234.18
AZN	10217.57	10236.06
BSL.SI	10167.04	10185.51
FMS	10218.41	10236.90
KANG	10218.01	10236.49
MDT	10218.22	10236.71
NVO	10218.43	10236.92
SKR.BK	9844.63	9862.98
SNY	10218.36	10236.85

Table 2 shows the comparison of MSE and the percentage of explained variance between the non-transformed method and transformed method. Noteworthy, lower MSE values indicates that the model has better performance. Hence, among the three dimension reduction techniques, GDPC

has the lowest MSE values as compared to the other two methods with the exception of ABT stock. This is followed by BDPC and finally OPC which has the highest value of MSE. These findings suggest that the non-transformed method can achieved a better model performance compared to the transformed methods.

Table 2. Mean squared error and percentage of explained variance of healthcare stock market price.

Stock Price	MSE			Percentage of Explained Variance		
	OPC	BDPC	GDPC	OPC	BDPC	GDPC
5225.KL	1.4989	1.0793	0.0020	52.8	97.8	98.7
ABT	5.4446	0.0437	0.2320	50.7	96.8	99.0
AZN	4.6357	1.0543	0.1730	52.5	93.4	97.2
BSL.SI	1.6002	0.2912	0.0002	37.9	75.0	99.1
FMS	1.8025	1.0097	0.2050	50.8	96.1	98.2
KANG	4.1370	1.1310	0.1160	52.6	85.8	98.0
MDT	2.0821	1.0461	0.5610	51.1	75.0	97.7
NVO	4.2714	1.0428	0.4780	51.4	87.4	99.2
SKR.BK	1.6707	0.9585	0.1480	50.4	87.2	99.8
SNY	2.0741	0.8731	0.2750	50.2	92.7	98.6

Here, only the first component percentage of explained variance is presented for demonstration reason (Andu et al., 2019; Peña & Yohai, 2016). The non-transformed method had higher explained percentage in all its ten healthcare stock market price (Table 2). On the other hand, five healthcare stock market prices have higher percentage of explained variance at 90% above in BDPC. Both BSL.SI and MDT stock market price shared the same percentage of explained variance at 75%. However, the transformed method of OPC has the lowest percentage of explained variance with an average of almost 50% on all of its stock market price. Therefore, having a higher first component percentage of explained variance indicates that much more information can be obtained from the non-transformed method compared to its counterpart.

CONCLUSION

A direct approach can apply to the nonstationary healthcare stock market price without the need for transformation. The non-transformed method presented that it has better result performance than the transformed techniques. Furthermore, the non-transformed methodology shows a higher percentage of explained variance in the first component of the nonstationary healthcare stock market price. Future studies can apply the non-transformed method to other stock market price sectors.

REFERENCES

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

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, *19*(6), 716-723.

Andronicеanu, A., Kinnunen, J., & Georgescu, I. (2021). Circular economy as a strategic option to promote sustainable economic growth and effective human development. *Journal of International Studies*, *14*(1).

Andu, Y., Lee, M. H., & Algamal, Z. Y. (2019). Non-transformed Principal Component Technique on Weekly Construction Stock Market Price. *MATEMATIKA: Malaysian Journal of Industrial and Applied Mathematics*, 139-147.

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

Brillinger, D. R. (1981). Time Series: Data Analysis and Theory. San Francsico.

Brillinger, D. R. (1964). The asymptotic behaviour of Tukey's general method of setting approximate confidence limits (the jackknife) when applied to maximum likelihood estimates. *Revue de l'Institut International de Statistique*, 202-206.

Caporale, G. M., Gil-Alana, L. A., & Tripathy, T. (2020). Volatility persistence in the Russian stock market. *Finance Research Letters*, *32*, 101216.

Crump, R. K., & Gospodinov, N. (2022). On the factor structure of bond returns. *Econometrica*, *90*(1), 295-314.

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

Kim, C. (2021). Deviance information criteria for mixtures of distributions. *Communications in Statistics-Simulation and Computation*, *50*(10), 2935-2948.

Kwiatkowski, D., Phillips, P. C., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?. *Journal of econometrics*, *54*(1-3), 159-178.

Peña, D., & Yohai, V. J. (2016). Generalized dynamic principal components. *Journal of the American Statistical Association*, *111*(515), 1121-1131.

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

Sánchez, M. Á., Trinidad, J. E., García, J., & Fernández, M. (2015). The effect of the underlying distribution in Hurst exponent estimation. *PLoS One*, *10*(5), e0127824.

Worden, K., Iakovidis, I., & Cross, E. J. (2021). New results for the ADF statistic in nonstationary signal analysis with a view towards structural health monitoring. Mechanical Systems and Signal Processing, 146, 106979.