Journal of Quality Measurement and Analysis *JQMA* 20(3) 2024, 89-107
e-ISSN: 2600-8602 *http://www.ukm.my/jqma*
https://doi.org/10.17576/jqma.2003.2024.07

ENHANCING STOCK PRICE DATA ANALYSIS THROUGH VARIANTS OF PRINCIPAL COMPONENT ANALYSIS

(Penambahbaikan Analisis Data Harga Saham Menerusi Varian-varian Analisis Komponen Prinsipal)

SEOW TECK XIANG, DHARINI PATHMANATHAN* & KHOO TZUNG HSUEN

ABSTRACT

This work investigates and identifies suitable dimensionality reduction approaches based on variants of principal component analysis (PCA) for various transformations of stock price data. The classical PCA, dynamic principal component analysis (DPCA) and generalised dynamic principal component analysis (GDPCA) were applied to the closing prices, simple returns and log of returns of the top 100 holdings of Standard & Poor's 500 (S&P500) from year 2020 to year 2023. The S&P 500 is a stock market index that tracks the stock performance of 500 large-cap U.S. companies. The performances of the aforementioned variants of PCA on these data for different timeframes were compared. Results showed that GDPCA works best for non-stationary time series data such as the closing prices and DPCA works best for stationary time series data such as the simple returns and the log of returns. The results obtained from the empirical analysis was further supported by simulation studies that follow, hence GDPCA and DPCA could be among the most appropriate dimensionality reduction approaches for non-stationary and stationary time series data respectively.

Keywords: stock market; non-stationary time series; geometric Brownian Motion; GJR-GARCH; GARCH

ABSTRAK

Kajian ini menyelidik dan mengenalpasti pendekatan-pendekatan pengurangan dimensi yang sesuai berpandukan varian-varian analisis komponen prinsipal untuk pelbagai penjelmaan data harga saham. Analisis komponen prinsipal statik, analisis komponen prinsipal dinamik dan analisis komponen prinsipal dinamik teritlak telah dikenakan pada harga tutup, pulangan mudah dan pulangan log harian untuk 100 saham terbesar yang tersenarai dalam Standard & Poor's 500 (S&P500) dari tahun 2020 ke tahun 2023. S&P500 merupakan indeks pasaran saham yang mengesan prestasi saham bagi 500 syarikat dengan modal pasaran yang besar di Amerika Syarikat. Prestasi varian-varian tersebut atas data yang mempunyai jangka masa berbeza juga telah dibandingkan. Keputusan menunjukkan bahawa analisis komponen prinsipal dinamik teritlak merupakan varian yang paling sesuai untuk siri-siri masa tidak pegun seperti data harga tutup saham manakala analisis komponen prinsipal dinamik adalah paling sesuai untuk siri-siri masa pegun seperti pulangan mudah dan pulangan log. Keputusan daripada simulasi-simulasi yang dijalankan seterusnya juga menyokong kesimpulan yang telah diperoleh daripada analisis empirikal, justeru analisis komponen prinsipal dinamik teritlak ialah antara pendekatan-pendekatan pengurangan dimensi siri masa tidak pegun yang paling berkesan manakala analisis komponen prinsipal dinamik ialah antara pendekatan-pendekatan pengurangan dimensi siri masa pegun yang paling sesuai.

Kata kunci: pasaran saham; siri masa tidak pegun; gerakan Brownan geometri; GJR-GARCH; GARCH

1. Introduction

Time series data is characterised by high dimensionality, large volume, and the presence of both noise and redundant features (Ashraf *et al.* 2023). Despite the notoriously volatile and unpredictable nature of financial data such as the stock price data, these data are used extensively in various finance-related endeavours such as technical analyses, quantitative research, and portfolio optimisations (Imai & Tan 2006; Wang 2006; Zhong & Enke 2017; Zhang & Wang 2023; Song *et al.* 2023). To individual and institutional investors, it is of great interest to extract as much relevant information and signals from such data to aid high-stakes decision making. Without performing any dimensionality reduction, these data in their original forms are often large and too complex to be deciphered straight away. The implementation of a suitable variant of dimensionality reduction helps reduce this complexity, which has the potential to enhance modeling and forecasting performance in the context of time series data.

Principal Component Analysis (PCA) is a widely used and powerful unsupervised statistical technique for reducing the dimensionality of high-dimensional datasets. The essence of the classical PCA is to transform a high-dimensional dataset with multiple features orthogonally to a lower dimension in the form of linearly uncorrelated variables named principal components (PCs), which are linear combinations of the features in the original dataset (Jolliffe 2002), thereby making the most salient relationships within such dataset more apparent.

The introduction of PCA led to innovations to deal with data with dynamic characteristics such as time series data. Dimensionality reduction of time series data is more challenging than that of regular multivariate data due to its dynamic behaviour and the presence of noise. To widen the scope of applications of PCA to time series analysis, Ku *et al.* (1995) proposed the application of PCA on the augmented time series observations that include values of series up to a certain number of lags. The resulting PCs are linear combinations of past and present values of the observed time series. However, with such definitions of PCs, there is no clear approach for reconstructing the observed time series (Peña & Yohai 2016). Generally, the classical PCA is unable to account for the dynamic nature of multivariate time series data. In reducing the dimension of time-dependent data, Brillinger (1981) proposed dynamic principal component analysis (DPCA), also known as the frequency domain principal component analysis, which extends the classical PCA by capturing the serial dependence between the observations and identifying the dynamic components that incorporate the variations of these observations over time.

Since the introduction of DPCA, there have been some applications of DPCA in fields that involve time series analysis. In the context of economics and finance, Mancino and Renò (2005) used DPCA as a part of the procedure to analyse multivariate volatility of stocks through Fourier analysis. Elliott *et al.* (2006) constructed some forecasting methods for time series in economics, finance and marketing based on DPCA. Donadelli and Paradiso (2014) also carried out DPCA to examine the financial integration process of emerging equity markets in Latin America, Asia and Eastern Europe.

A crucial requirement of DPCA is that the time series being analysed should be stationary (Brillinger 1981). Even though this does not forbid the application of DPCA on time series in general, the usefulness of DPCA in the real world could be limited since it is rare for observed time series to be truly stationary, considering that most real-world phenomena are affected by factors that change over time (Mader *et al.* 2006). To approach dimensionality reduction of non-stationary multivariate time series data, Peña and Yohai (2016) proposed generalised dynamic principal component analysis (GDPCA), which is an extension of DPCA. In GDPCA, the PCs need not be a linear combination of the observations, and various loss functions including robust ones could be considered. Peña and Yohai (2016) also introduced a robust version of GDPCA that could deal with the presence of outliers.

One possible use of other variants of PCA (i.e. DPCA, GDPCA), that is yet to be explored thoroughly, is its application on various transformations (e.g. simple return, log of returns) of time series data related to stock prices. The literature on dimensionality reduction of financial data using DPCA or GDPCA is limited (Elliott *et al.* 2006; Donadelli & Paradiso 2014; Mancino & Renò 2005), and there is no consensus on how the classical PCA, DPCA or GDPCA would perform on stock prices under various transformations. The nature of the transformations of stock price data might warrant the use of certain variants of PCA to minimise the loss of information. Moreover, the use of GDPCA specifically could effectively reduce the complexity of the original dataset, potentially enhancing forecasting performance. Should such variants exist, these methods could form the groundwork of dimensionality reduction of various transformations of stock price data and, in turn, make the results of PCA more useful in the settings of investment and finance.

This work serves as an exploratory study to determine the most suitable variant of PCA, whether PCA, DPCA, or GDPCA, for reducing the dimensionality of various time series data with stationary and non-stationary characteristics. This is illustrated using transformations of stock price data, such as closing prices, simple returns, and log returns, for the top 100 holdings of the Standard & Poor's 500 (S&P 100). Additionally, it assesses the performance of these PCA variants across different timeframes. It is important to note that the simulation study in this work is designed for a shorter time frame compared to the real data application because simulated data behaves according to the model from which it is generated, whereas real data accounts for 'unexpected' changes that may not be captured in the simulated data. Therefore, data-dependent behavior is studied using real data over a longer period. In Section 2, the frameworks for PCA, DPCA, GDPCA and the models used in the simulation studies are explained. Section 3 gives the results of the application of the variants of PCA on the S&P100 stock price data. Section 4 presents the simulation studies to generalise the findings and Section 5 concludes.

2. Methodology

2.1. *Variants of principal component analysis*

2.1.1. *Principal component analysis (PCA)*

Suppose a data matrix $\mathbf{Z}$ comprising T daily stock price data points (rows) of m stocks (columns) and $T > m$. Let $\mathbf{z}_j, 1 \leq j \leq m$ be the j-th column of $\mathbf{Z}$. PCA finds the set of orthonormal column vectors $\boldsymbol{u_k}, k = 1, 2, \dots, m$, in that order, that maximise $Var(\boldsymbol{u_k}'\mathbf{z}_j)$ (Jolliffe 2002) or equivalently, minimise the reconstruction criterion $\sum_{j=1}^{m} ||\mathbf{z}_j - \hat{\mathbf{z}}_j||^2$, where $\hat{\mathbf{z}}_j$ is the orthogonal projection of $\mathbf{z}_j$ onto $\boldsymbol{u_k}$ in the T-dimensional space (Pearson 1901).

Let $\mathbf{\Sigma} = \mathbf{Z}'\mathbf{Z}/\mathrm{T}$ be the sample covariance matrix of $\mathbf{z}_j$ (Peña & Yohai 2016). Under PCA, the solution for $\boldsymbol{u_k}$ is the unit eigenvector of $\mathbf{\Sigma}$ having the k-th largest eigenvalue. $\boldsymbol{u_k}$ is thus the vector of coefficients for the k-th PC (Jolliffe 2002).

2.1.2. *Dynamic principal component analysis (DPCA)*

Let $\{z_t\}$, $-\infty < t < \infty$ be a zero mean m-dimensional stationary process. DPCA aims to minimise the same reconstruction criterion in sub-subsection 2.1.1 through Fourier analysis (Brillinger 1981).

Under the original definition, DPCA finds $m \times 1$ vectors $\boldsymbol{c}_h, -\infty < h < \infty$ and $\beta_j, -\infty < j < \infty$, so that the linear combination

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

minimises the loss function $E\left[\left(z_t - \sum_{j=-\infty}^{\infty} \beta_j \, f_{t+j}\right)'\left(z_t - \sum_{j=-\infty}^{\infty} \beta_j \, f_{t+j}\right)\right]$. f_t is then considered as the first DPC (Peña & Yohai 2016).

The solutions to $\boldsymbol{c}_k$ and β_j are the inverse Fourier transforms of the PCs of the cross spectral matrices for each frequency and the inverse Fourier transforms of the conjugates of the same PCs respectively (Brillinger 1981). More commonly, DPCA is applied to stationary processes with finite number of time points. In this case, the number of lags in Eq. (1) and in the reconstruction of the time series should be replaced with the corresponding finite number (Peña & Yohai 2016).

2.1.3. *Generalised dynamic principal component analysis (GDPCA)*

The algorithm of GDPCA adopted here is the same as that presented in Peña and Yohai's work (2016). Let $\{z_{j,t}\}, 1 \leq j \leq m, 1 \leq t \leq T$ be a non-stationary time series. Let $k_1 \geq 0$ be the number of lags and $k_2 \geq 0$ be the number of leads. Let $k = k_1 + k_2$. Also let $\mathbf{f} = (f_1, \ \ldots, f_{T+k})', \beta = (\beta_{j,i})_{1\leq j\leq m, 1\leq i\leq k+1}$ and $\alpha = (\alpha_1, \ldots, \alpha_m)$ be the unknowns to be solved.

The reconstruction criterion to be minimised here is the mean squared error (MSE):

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

Some components needed for the solutions are defined in the next part. Let $\mathbf{C}_j(\alpha_j) = (c_{j,t,q}(\alpha_j))_{1\leq t\leq T+k, 1\leq q\leq k+1}$ be the $(T+k) \times (k+1)$ matrix defined by

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

where $a \vee b = \max(a,b)$ and $a \wedge b = \min(a,b)$. Let $\mathbf{D}_j(\beta_j) = (d_{j,t,q}(\beta_j))$ be the $(T+k) \times (T+k)$ matrix given by

$$d_{j,t,q}(\beta_j) = \begin{cases} \sum_{v=(t-k)\vee 1}^{t\wedge T} \beta_{j,q-v+1} \beta_{j,t-v+1} & \text{if } (t-k) \vee 1 \leq q \leq (t+k) \wedge (T+k); \\ 0 & \text{if otherwise;} \end{cases}$$

and $\mathbf{D}(\beta) = \sum_{j=1}^{m} \mathbf{D}_j(\beta_j)$. After differentiating Eq. (2) with respect to f_t, the solution is

$$\mathbf{f} = \ \mathbf{D}(\beta)^{-1} \sum_{j=1}^{m} \mathbf{C}_j(\alpha)\beta_j. \tag{3}$$

β_j and $\alpha_j, 1 \leq j \leq m$, can also be obtained using the least squares estimator

$$\binom{\beta_j}{\alpha_j} = \left(\mathbf{F}(\mathbf{f})'\mathbf{F}(\mathbf{f})\right)^{-1} \mathbf{F}(\mathbf{f})'\mathbf{z}^{(j)}, \tag{4}$$

where $\mathbf{z}^{(j)} = (z_{j,1}, \dots, z_{j,T})'$ and $\mathbf{F}(\mathbf{f})$ is the $T \times (k+2)$ matrix with t-th row $(f_t, f_{t+1}, \dots, f_{t+k}, 1)$. Eqs. (3) and (4) define GDPC1. GDPC2 is defined as PC1 of the residuals $R_{j,t}(\mathbf{f}, \beta)$. Subsequent GDPCs are also defined similarly.

2.2. *Models in simulation studies*

Geometric Brownian motion (GBM), the Glosten Jagannatan Runkle-Generalized Autoregressive Conditional Heteroscedasticity (GJR-GARCH) model and the Generalized Autoregressive Conditional Heteroscedasticity (GARCH) model are used to simulate closing prices, simple returns and log of returns of S&P100 respectively. The choices of models were based on the most preferred methods for simulating stock price data in the literature (Nugroho *et al.* 2019; Kambouroudis *et al.* 2016; Reddy & Clinton 2016; Marathe & Ryan 2005) and these models were further supported by good model fits to the S&P100 stock price data. To avoid overfitting and ensure good generalisations to unseen data, the number of parameters was minimised without compromising the model fit (Ying 2019).

2.2.1. *Geometric Brownian motion (GBM)*

Let $X(t)$ be a GBM and $Z(t)$ be a standard Brownian motion. The stochastic differential equation of GBM is $\mathrm{d}X(t) = \mu X(t)\,\mathrm{d}t + \sigma X(t)\,\mathrm{d}Z(t)$ (McDonald 2013). Since a process that follows GBM is lognormally distributed,

$$\ln[X(t)] \sim N(\ln[X(0)] + (\mu - 0.5\sigma^2)t, \sigma^2 t). \tag{5}$$

As a consequence of Eq. (5),

$$X(t) = X(0)e^{(\mu - 0.5\sigma^2)t + \sigma^2 t Z}, \tag{6}$$

where $Z \sim N(0,1)$ (McDonald 2013). For each stock, $\sigma\sqrt{t}$ is estimated using the sample standard deviations of observed log of returns and $(\mu - 0.5\sigma^2)t$ is estimated using the mean of observed log of returns (McDonald 2013).

2.2.2. *The GJR-GARCH(1,1,1) model*

The conditional variance equation of the GJR-GARCH(1,1,1) model for simulating simple returns is

$$\sigma_t^2 = \theta_0 + \delta_1 \sigma_{t-1}^2 + \theta_1 \varepsilon_{t-1}^2 + \gamma_1 \varepsilon_{t-1}^2 I_{t-1}(\varepsilon_{t-1} < 0), \tag{7}$$

where $\theta_0 > 0$ is the additive constant of the variance equation; $\theta_1 \geq 0$ is the ARCH coefficient; $\delta_1 \geq 0$ is the GARCH coefficient; $\theta_1 + \delta_1 < 1$ to ensure that $\{\sigma_t^2\}$ is defined and is weakly stationary; γ_1 is the asymmetric parameter and $I_{t-1}(\varepsilon_{t-1} < 0)$ is the indicator function that equals one when $\varepsilon_{t-1} < 0$ and zero otherwise (Glosten *et al.* 1993).

The simulated simple returns are then given by

$$r_t | \mathcal{F}_{t-1} = E(r_t | \mathcal{F}_{t-1}) + \varepsilon_t, \tag{8}$$

where $E(r_t | \mathcal{F}_{t-1})$ is the mean of returns conditional on the information set, $\mathcal{F}_{t-1}$, up until and including $t-1$ and $\varepsilon_t = \sigma_t z_t$ where $z_t \sim N(0,1)$ (Glosten *et al.* 1993).

2.2.3. *The GARCH(1,1) model*

The conditional variance equation of the GARCH(1,1) model for simulating log of returns is

$$\sigma_t^2 = \theta_0 + \delta_1\sigma_{t-1}^2 + \theta_1\varepsilon_{t-1}^2, \tag{9}$$

where $\theta_0 > 0$ is the additive constant of the variance equation; $\theta_1 \geq 0$ is the ARCH coefficient; $\delta_1 \geq 0$ is the GARCH coefficient; $\theta_1 + \delta_1 < 1$ to ensure that $\{\sigma_t^2\}$ is defined and is weakly stationary. Under the GARCH(1,1) model setting, the simulated log of returns are also given by Eq. (8) (Bollerslev 1986).

3. Application to Real Data

3.1. *Data description*

Three-year closing prices of S&P100 from 29 March 2020 to 28 March 2023 were downloaded from Yahoo Finance through the R package quantmod (Ryan *et al.* 2022). To examine the performance of PCAs on time series of different lengths, specific 6-month, 1-year, 2-year and 3-year periods were used in the subsequent analysis. The exact dates of those periods can be found in Table 1.

Table 1: Specific periods of stock market data used for analyses

Length	First trading day	Last trading day	Number of trading days
6 months	1 July 2022	30 December 2022	127
1 year	3 January 2022	30 December 2022	251
2 years	4 January 2021	30 December 2022	503
3 years	30 March 2020	27 March 2023	754

3.2. *PCAs on closing prices*

PCA, DPCA and GDPCA were applied to the closing prices of S&P100 during the specified 6-month, 1-year, 2-year and 3-year periods by using the freqdom (Hormann & Kidzinski 2022) and gdpc (Peña *et al.* 2020) R packages.

Variants of PCA studied in this work were applied to mean-centered closing prices. In this context, PCA reduces the time points to lags where the dimension is reduced to the optimal lag length based on all the time points.

Regardless of PCA variants and investigated timeframes, the proportions of variance explained by the first three PCs are approximately 90%. Hence, the first three PCs were used to approximate the mean-centered closing prices before they are back transformed using the original means. Then, the mean absolute errors (MAE), root mean square errors (RMSE) and mean absolute percentage errors (MAPE) of the approximation for each stock were calculated (see Tables 2 and 3). Figure 1(a) shows the approximation of the 6-month closing prices for Bristol-Myers Squibb Company (BMY) using the first three PCs. Figure 2, Figure 3 and Figure 4 show the boxplots of the RMSE, MAE and MAPE of the approximation of closing prices for S&P100 respectively. On the horizontal axis of these figures, “P” represents PCA; “D” represents DPCA; and “G” represents GDPCA.

Table 2: Proportion of variance explained by the first three PCs when PCAs were applied to closing prices

Periods	**Observed Closing Prices (S&P100)**		
	Variance Explained (3 PCs)		
	PCA	**DPCA**	**GDPCA**
6 months	91.72%	91.92%	**98.61%**
1 year	89.68%	89.83%	**98.06%**
2 years	89.86%	89.92%	**97.61%**
3 years	91.92%	91.97%	**97.64%**

Table 3: Averages and standard deviations of MAE, RMSE and MAPE when PCAs were applied to closing prices

Periods	**Observed Closing Prices (S&P100)**								
	MAE			**RMSE**			**MAPE**		
	PCA	**DPCA**	**GDPCA**	**PCA**	**DPCA**	**GDPCA**	**PCA**	**DPCA**	**GDPCA**
6 months	4.25	4.16	**1.77**	5.34	5.23	**2.28**	2.33	2.28	**1.00**
	(3.19)	(3.11)	**(1.20)**	(4.06)	(3.98)	**(1.55)**	(0.87)	(0.85)	**(0.34)**
1 year	7.46	7.42	**3.21**	9.32	9.25	**4.10**	3.95	3.90	**1.71**
	(5.63)	(5.60)	**(2.34)**	(7.03)	(6.98)	**(2.99)**	(1.52)	(1.50)	**(0.61)**
2 years	9.08	9.21	**4.45**	11.27	11.59	**5.67**	4.81	4.85	**2.36**
	(7.30)	(7.35)	**(3.15)**	(8.90)	(9.10)	**(4.04)**	(2.06)	(2.10)	**(0.79)**
3 years	9.89	10.19	**5.44**	12.21	12.60	**6.90**	5.74	5.87	**3.15**
	(7.75)	(7.98)	**(3.71)**	(9.53)	(9.79)	**(4.75)**	(2.70)	(2.76)	**(1.16)**

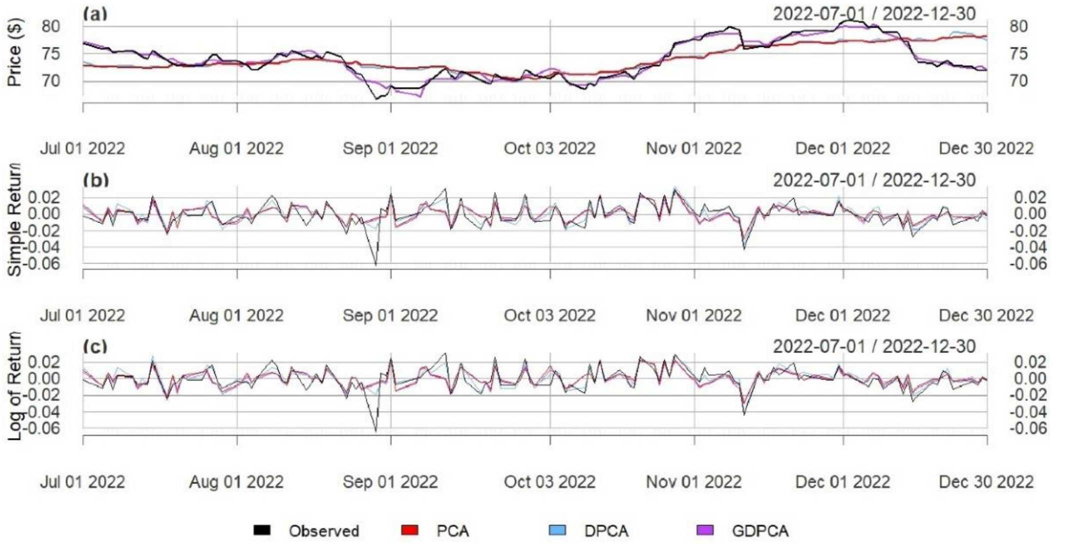

Figure 1: Approximation of 6-month (a) closing prices using the first three PCs; (b) simple returns using the first ten PCs; (c) log of returns using the first ten PCs for Bristol-Myers Squibb Company (BMY)

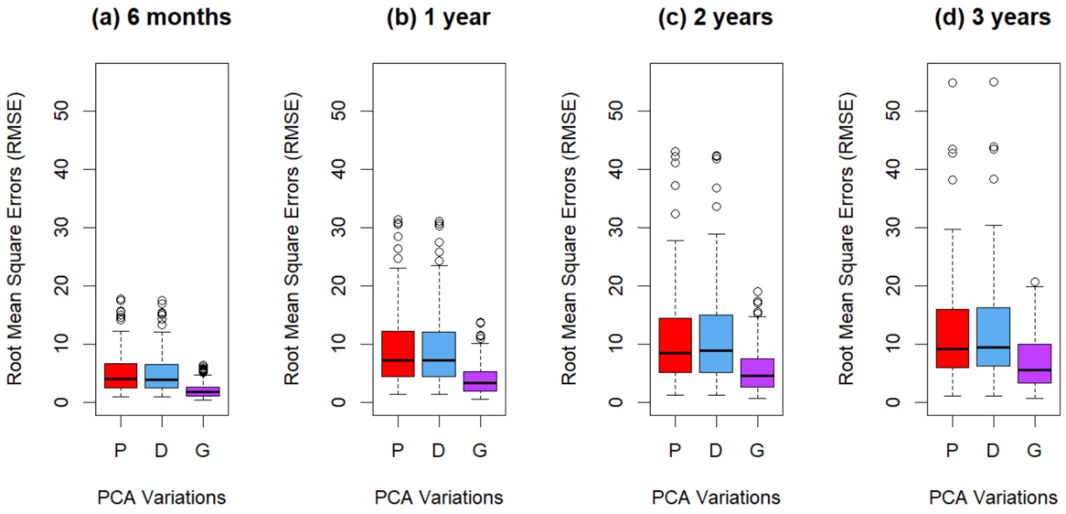

Figure 2: Boxplots of RMSE of closing price approximations using the first three PCs of various PCAs for (a) 6-month, (b) 1-year, (c) 2-year and (d) 3-year periods.

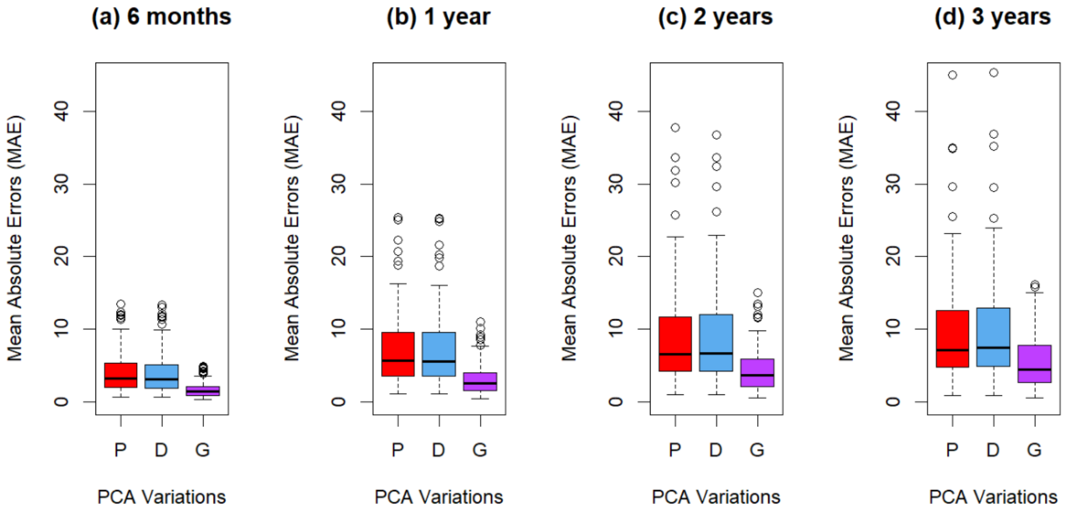

Figure 3: Boxplots of MAE of closing price approximations using the first 3 PCs of various PCAs for (a) 6-month, (b) 1-year, (c) 2-year and (d) 3-year periods

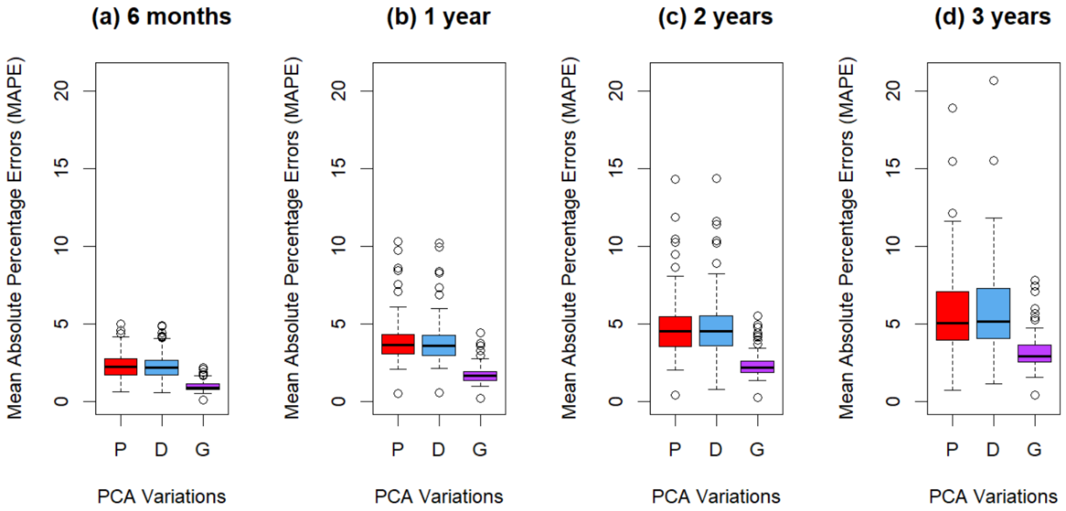

Figure 4: Boxplots of MAPE of closing price approximations using the first 3 PCs of various PCAs for (a) 6-month, (b) 1-year, (c) 2-year and (d) 3-year periods

3.3. *PCAs on simple returns*

The procedure of applying variants of PCA on simple returns data is similar to that of subsection 3.2 except that no mean-centering is required for simple returns data.

Regardless of PCA variants and investigated timeframes, the proportions of variance explained by the first ten PCs are approximately 70%. Hence, the first ten PCs were used to approximate the simple returns. Then, the MAE and RMSE of the approximation for each stock were calculated (see Tables 4 and 5). Figure 1(b) shows the approximation of 6-month simple returns of BMY using the first ten PCs. Figure 5 and Figure 6 show the boxplots of the RMSE and MAE of the approximation of simple returns for S&P100 respectively. On the horizontal axis of these figures, “P” represents PCA; “D” represents DPCA; and “G” represents GDPCA.

Table 4: Proportion of variance explained by the first ten PCs when PCAs were applied to simple returns

Periods	**Observed Simple Returns (S&P100)**		
	Variance Explained (10 PCs)		
	PCA	**DPCA**	**GDPCA**
6 months	76.77%	**83.60%**	77.93%
1 year	74.13%	**79.12%**	74.43%
2 years	68.68%	**72.66%**	71.63%
3 years	68.65%	72.18%	**72.26%**

Table 5: Averages and standard deviations of MAE and RMSE when PCAs were applied to simple returns

Periods	**Observed Simple Returns (S&P100)**					
	MAE			**RMSE**		
	PCA	**DPCA**	**GDPCA**	**PCA**	**DPCA**	**GDPCA**
6 months	0.00705	**0.00499**	0.00700	0.00980	**0.00654**	0.00958
	(0.00145)	**(0.00108)**	(0.00145)	(0.00224)	**(0.00150)**	(0.00210)
1 year	0.00786	**0.00655**	0.00785	0.01105	**0.00869**	0.01101
	(0.00156)	**(0.00114)**	(0.00148)	(0.00232)	**(0.00158)**	(0.00223)
2 years	0.00749	**0.00684**	0.00741	0.01065	**0.00925**	0.01019
	(0.00149)	**(0.00120)**	(0.00128)	(0.00225)	**(0.00167)**	(0.00185)
3 years	0.00787	**0.00731**	0.00774	0.01124	**0.00999**	0.01063
	(0.00155)	**(0.00128)**	(0.00132)	(0.00239)	**(0.00181)**	(0.00188)

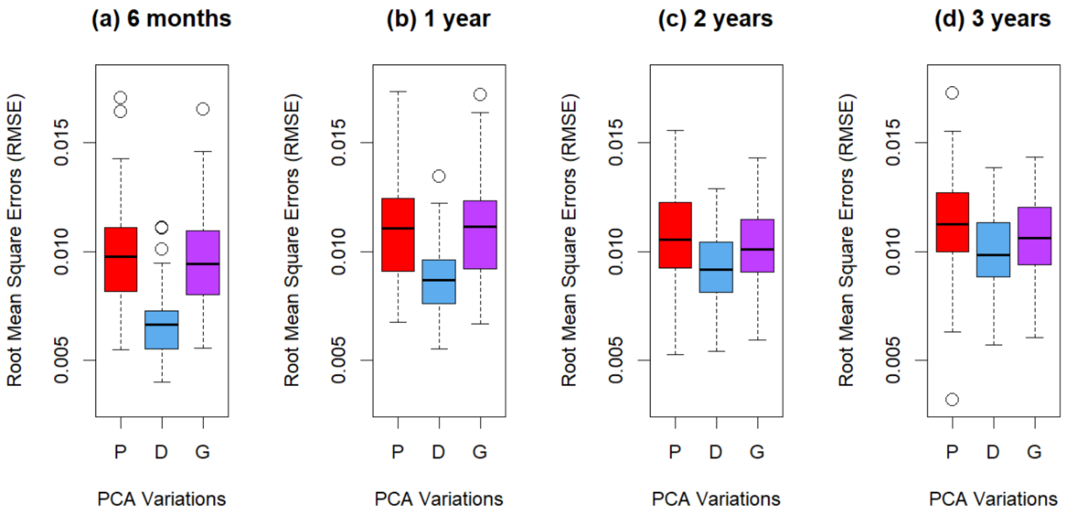

Figure 5: Boxplots of RMSE of simple return approximations using the first ten PCs of various PCAs for (a) 6-month, (b) 1-year, (c) 2-year and (d) 3-year periods

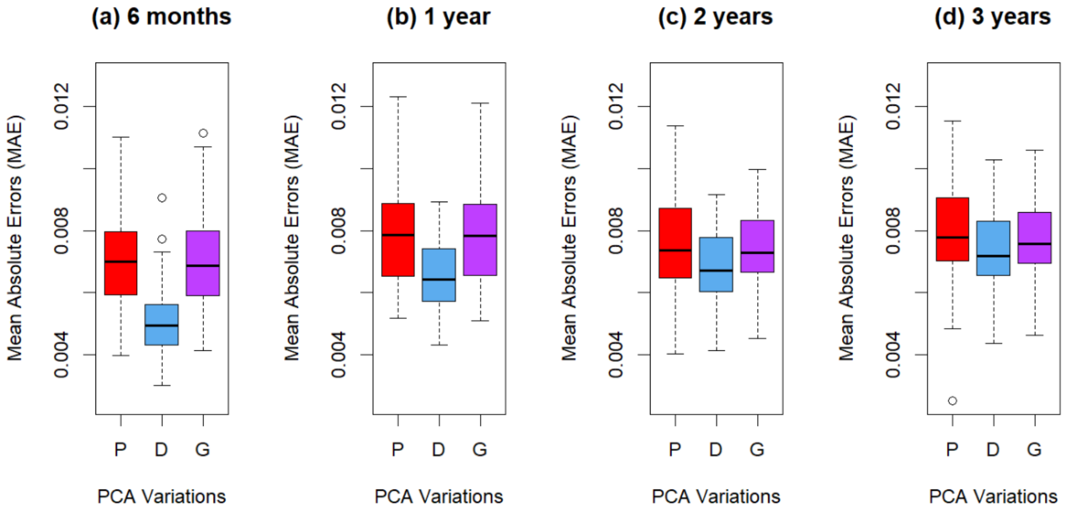

Figure 6: Boxplots of MAE of simple return approximations using the first 10 PCs of various PCAs for (a) 6-month, (b) 1-year, (c) 2-year and (d) 3-year periods

3.4. *PCAs on log of returns*

To examine the performance of PCAs on stationary time series, for each period, only the stocks among the Top 100 that are concluded to have stationary log of returns based on the Augmented Dickey-Fuller (ADF) test and the Kwiatkowski, Phillips, Schmidt, and Shin (KPSS) test are included in the analysis. The list of excluded stocks for each period analysed are available in Table 6.

Table 6: Results of the ADF test and the KPSS test

Periods	Stocks with non-stationary log of returns (ADF)	Stocks with non-stationary log of returns (KPSS)	Number of stocks with stationary log of returns
6 months	2 (GILD, HON)	1 (TSLA)	97
1 year	0	2 (GILD, SBUX)	98
2 years	0	2 (GOOG, GOOGL)	98
3 years	0	7 (PYPL, DHR, TSLA, GOOG, LOW, GOOGL, MS)	93

The procedure of applying variants of PCAs on log of returns data is similar to that of subsection 3.3. Regardless of PCA variants and investigated timeframes, the proportions of variance explained by the first ten PCs are approximately 70%. Hence, the first ten PCs were used to approximate the log of returns. Then, the MAE and RMSE of the approximation for each stock were calculated (see Tables 7 and 8). Figure 1(c) shows the approximation of 6-month log of returns of BMY using the first ten PCs. Figure 7 and Figure 8 show the boxplots of the RMSE and MAE of the approximation of log of returns respectively. On the horizontal axis of these figures, "P" represents PCA; "D" represents DPCA; and "G" represents GDPCA.

Table 7: Proportion of variance explained by the first ten PCs when PCAs were applied to stationary log of returns

Periods	**Observed Stationary Log of returns**		
	Variance Explained (10 PCs)		
	PCA	**DPCA**	**GDPCA**
6 months	76.83%	**83.83%**	77.49%
1 year	74.45%	**79.35%**	74.85%
2 years	68.70%	**72.65%**	71.19%
3 years	68.67%	72.26%	**72.43%**

Table 8: Averages and standard deviations of MAE and RMSE when PCAs were applied to stationary log of returns

Periods	**Observed Stationary Log of returns**					
	MAE			**RMSE**		
	PCA	**DPCA**	**GDPCA**	**PCA**	**DPCA**	**GDPCA**
6 months	0.00701	**0.00486**	0.00695	0.00971	**0.00640**	0.00959
	(0.00141)	**(0.00097)**	(0.00143)	(0.00213)	**(0.00137)**	(0.00212)
1 year	0.00786	**0.00651**	0.00783	0.01106	**0.00862**	0.01099
	(0.00160)	**(0.00113)**	(0.00158)	(0.00241)	**(0.00158)**	(0.00235)
2 years	0.00749	**0.00682**	0.00746	0.01066	**0.00926**	0.01030
	(0.00151)	**(0.00120)**	(0.00132)	(0.00230)	**(0.00168)**	(0.00187)
3 years	0.00774	**0.00717**	0.00755	0.01102	**0.00979**	0.01039
	(0.00142)	**(0.00119)**	(0.00122)	(0.00217)	**(0.00169)**	(0.00175)

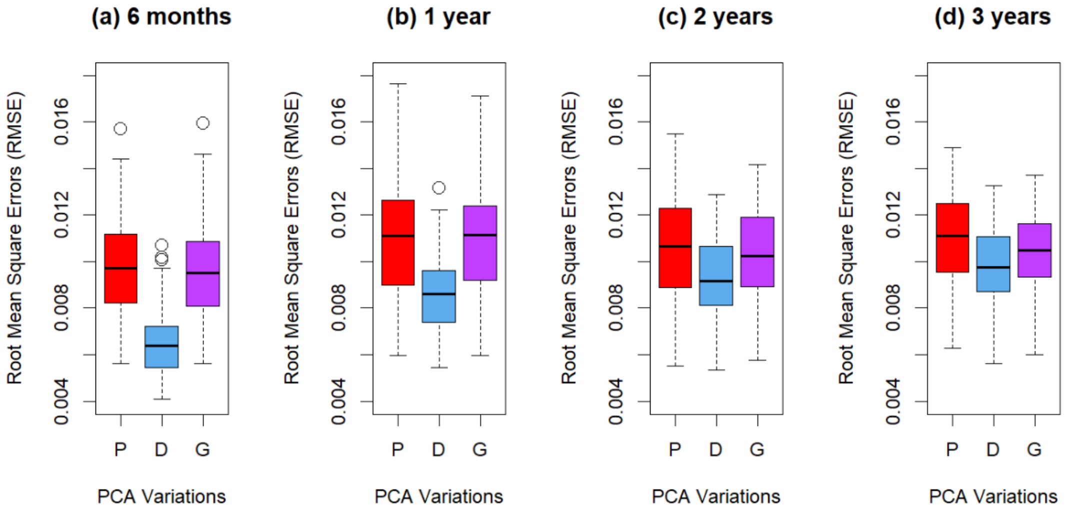

Figure 7: Boxplots of RMSE of stationary log of return approximations using the first ten PCs of various PCAs for (a) 6-month, (b) 1-year, (c) 2-year and (d) 3-year periods

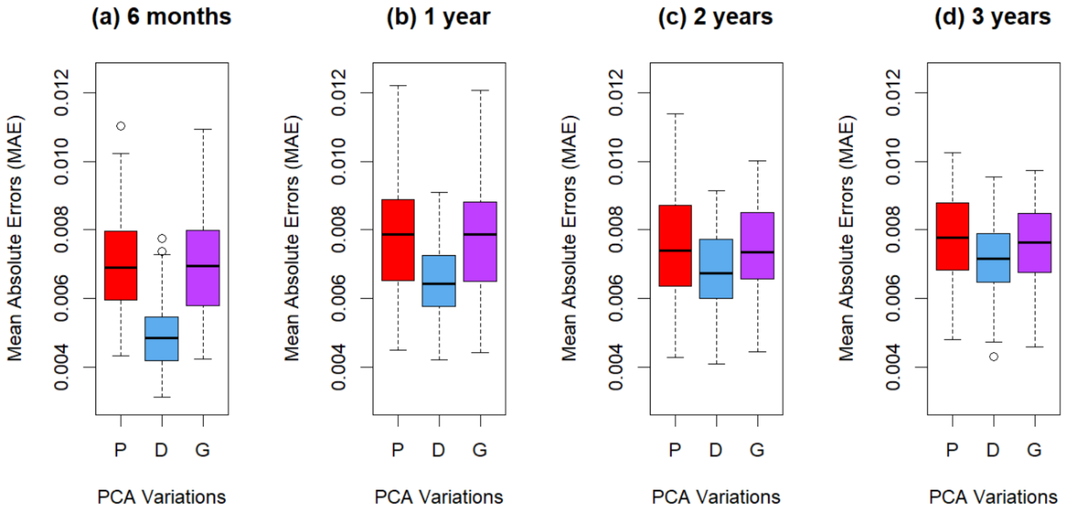

Figure 8: Boxplots of MAE of stationary log of return approximations using the first 10 PCs of various PCAs for (a) 6-month, (b) 1-year, (c) 2-year and (d) 3-year periods

3.5. *Discussion*

When applied to closing prices, GDPCA always has the highest proportions of variance explained and the lowest approximation errors. Meanwhile, for simple returns and log of returns, DPCA always has the lowest approximation errors. DPCA also always has the highest proportion of variance explained except for 3-year returns, in which its variance explained is comparable to that of GDPCA.

GDPCA is unique in its ability to handle datasets with non-stationarity features and outliers, hence it stands out when PCAs are applied to closing prices. Meanwhile, DPCA works best at capturing the dynamics of stationary time series and therefore it is the best method when it comes to stationary data such as the log of returns. This work sheds light to the effectiveness of dimensionality reduction using DPCA on simple returns as well as log of returns data, and GDPCA on closing prices data.

As the length of time of the stock price data increases, the approximation errors of PCAs increase. Stock prices and returns become more unpredictable in the long run, therefore for the same number of PCs, only a lower proportion of total signals could be captured. GDPCA's ability to capture the underlying structure of non-stationary data can be evaluated by assessing the reconstruction error, similar to the approach used for PCA. This enables a meaningful comparison between GDPCA's projection performance and the prediction performance of other models, aligning with previous work that explores PCA's role in prediction accuracy.

To validate the observations and generalise the findings, some simulation studies were carried out.

4. Simulation Studies

4.1. *Simulating closing prices using geometric Brownian motion (GBM)*

After estimating the parameters for each stock using the observed log of returns, GBM was used to simulate the closing prices of S&P100 during the specified 6-month and 1-year periods. Simulations longer than a year are not considered due to the volatility and unpredictable behaviour of the stock market so it would not make sense to expect trends in price movements to persist for long (Ma *et al.* 2022; Wei & Huang 2012). One year is considerably long enough for stock market data simulations since many unforeseen circumstances and black swan events could still occur in such timeframes (Wei & Huang 2012).

For each set of simulated closing prices (each iteration), PCAs were applied to the simulated data and the proportion of variance explained by the first three PCs, MAEs, RMSEs and MAPEs of the approximations for each stock were calculated. The number of iterations was increased until the averages and standard deviations of proportions of variance explained, mean MAEs, mean RMSEs and mean MAPEs converge. In this case, 50 iterations are sufficient to achieve such convergence. The simulation results confirm that GDPCA performs best for closing price data (see Tables 9 and 10).

Table 9: Averages and standard deviations of proportions of variance explained by the first three PCs when PCAs were applied to the simulated closing prices using 50 iterations of geometric Brownian motion model

Periods	**Simulated Closing Prices from Geometric Brownian Motion (GBM)**		
	Variance Explained (3 PCs)		
	PCA	**DPCA**	**GDPCA**
6 months	90.42%	90.72%	**98.52%**
	(3.58%)	(3.43%)	**(0.55%)**
1 year	89.85%	90.02%	**98.03%**
	(3.46%)	(3.38%)	**(0.68%)**

Table 10: Mean MAE, mean RMSE and mean MAPE when PCAs were applied to the simulated closing prices using 50 iterations of geometric Brownian motion model

Periods	**Simulated Closing Prices from Geometric Brownian Motion (GBM)**								
	MAE			**RMSE**			**MAPE**		
	PCA	**DPCA**	**GDPCA**	**PCA**	**DPCA**	**GDPCA**	**PCA**	**DPCA**	**GDPCA**
6 months	6.21 (4.92)	6.12 (4.84)	**2.54** **(1.78)**	7.70 (6.08)	7.60 (6.01)	**3.18** **(2.23)**	3.31 (1.37)	3.26 (1.34)	**1.39** **(0.49)**
1 year	10.96 (9.50)	10.94 (9.48)	**5.04** **(3.87)**	13.66 (11.81)	13.65 (11.85)	**6.34** **(4.87)**	5.11 (2.34)	5.10 (2.37)	**2.41** **(0.92)**

4.2. *Simulating simple returns using the GJR-GARCH(1,1,1) model*

The GJR-GARCH(1,1,1) model is fitted to the observed simple returns of S&P100 for the specified 6-month and 1-year periods using the R package rugarch (Ghalanos 2022). Through the weighted Ljung-Box test on standardized residuals and the weighted Ljung-Box test on standardized squared residuals that came along with the model fits, it was concluded that GJR-GARCH(1,1,1) model is a good fit to the simple returns of 92 stocks for the 6-month period and 89 stocks for the 1-year period, thus in general, GJR-GARCH(1,1,1) model is a good fit to the simple returns of S&P100. The list of stocks having simple returns that were not fitted well by GJR-GARCH(1,1,1) is available in Table 11.

Table 11: Stocks having simple returns not fitted well by GJR-GARCH(1,1,1)

Periods	**Stocks poorly fitted by GJR-GARCH(1,1,1)**
6 months	8 (LLY, ORCL, TSLA, SYK, GS, BAC, VRTX, CI)
1 year	11 (LLY, ZTS, TSLA, IBM, BKNG, MS, GS, BAC, CI, TJX, DE)

The fitted parameters of the GJR-GARCH(1,1,1) model were used to simulate simple returns for the 92 stocks during the 6-month period and the 89 stocks during the 1-year period. For each set of simulated simple returns (each iteration), PCAs were applied to the simulated returns and the proportion of variance explained by the first ten PCs, MAEs and RMSEs of the approximations for each stock were calculated.

The number of iterations was increased until the averages and standard deviations of proportions of variance explained, mean MAEs and mean RMSEs converge. In this case, 50 iterations are sufficient to achieve such convergence. The simulation results are in line with the findings in subsection 3.3 (see Tables 12 and 13).

Table 12: Averages and standard deviations of proportions of variance explained by the first ten PCs when PCAs were applied to the simulated simple returns using 50 iterations of GJR-GARCH(1,1,1) model

Periods	Simulated Simple Returns from GJR-GARCH(1,1,1)		
	Variance Explained (10 PCs)		
	PCA	DPCA	GDPCA
6 months	52.22%	**66.05%**	60.25%
	(6.94%)	**(4.74%)**	(6.55%)
1 year	36.36%	**49.51%**	47.12%
	(2.90%)	**(2.09%)**	(3.04%)

Table 13: Mean MAEs and mean RMSEs when PCAs were applied to the simulated simple returns using 50 iterations of GJR-GARCH(1,1,1) model

Periods	Simulated Simple Returns from GJR-GARCH(1,1,1)					
	MAE			RMSE		
	PCA	DPCA	GDPCA	PCA	DPCA	GDPCA
6 months	0.01232	**0.00865**	0.01139	0.01609	**0.01106**	0.01469
	(0.00514)	**(0.00315)**	(0.00471)	(0.00677)	**(0.00412)**	(0.00605)
1 year	0.01265	**0.01014**	0.01173	0.01653	**0.01295**	0.01513
	(0.00364)	**(0.00241)**	(0.00317)	(0.00478)	**(0.00308)**	(0.00403)

4.3. *Simulating log of returns using the GARCH(1,1) model*

The GARCH(1,1) model is fitted to the observed log of returns of S&P100 for the specified 6-month and 1-year periods using the R package rugarch (Ghalanos 2022). Through the weighted Ljung-Box test on standardized residuals and the weighted Ljung-Box test on standardized squared residuals that came along with the model fits, it was concluded that the GARCH(1,1) model is a good fit to the log of returns of 90 stocks for the 6-month period and 87 stocks for the 1-year period, thus in general, the GARCH(1,1) model is a good fit to the log of returns of S&P100. The list of stocks having log of returns that were not fitted well by the GARCH(1,1) model is available in Table 14.

Table 14: Stocks having log of returns not fitted well by GARCH(1,1)

Periods	Stocks failed to be fitted by GARCH(1,1)	Stocks poorly fitted by GARCH(1,1)
6 months	5 (META, WMT, COST, AAPL, TJX)	5 (TSLA, BAC, VRTX, NKE, INTC)
1 year	6 (INTU, GOOG, GOOGL, COP, NOW, MSFT)	7 (WMT, TMO, ZTS, TSLA, SO, BAC, AXP)

The procedure of simulations and applications of PCAs on simulated log of returns is the same as those in subsection 4.2. For the GARCH(1,1) model, 50 iterations are sufficient for the averages and standard deviations of variance explained and approximation errors to converge.

The outcome of the simulation study reaffirms the fact that DPCA is best at reducing the dimensionality of log of returns data (see Tables 15 and 16).

Table 15: Averages and standard deviations of proportions of variance explained by the first ten PCs when PCAs were applied to the simulated log of returns using 50 iterations of GARCH(1,1) model

Periods	**Simulated Log of returns from GARCH(1,1)**		
	Variance Explained (10 PCs)		
	PCA	**DPCA**	**GDPCA**
6 months	46.43%	**62.41%**	60.48%
	(0.95%)	**(0.57%)**	(3.20%)
1 year	37.20%	50.45%	**52.29%**
	(0.69%)	**(0.47%)**	(2.22%)

Table 16: Mean MAEs and mean RMSEs when PCAs were applied to the simulated log of returns using 50 iterations of GARCH(1,1) model

Periods	**Simulated Log of returns from GARCH(1,1)**					
	MAE			**RMSE**		
	PCA	**DPCA**	**GDPCA**	**PCA**	**DPCA**	**GDPCA**
6 months	0.01083	**0.00741**	0.00929	0.01354	**0.00929**	0.01163
	(0.00532)	**(0.00325)**	(0.00452)	(0.00669)	**(0.00417)**	(0.00565)
1 year	0.01149	**0.00903**	0.01008	0.01440	**0.01131**	0.01263
	(0.00430)	**(0.00289)**	(0.00353)	(0.00544)	**(0.00368)**	(0.00442)

5. Conclusion

Through the application of PCA variants on real stock price data and subsequent simulation studies, it can be concluded that GDPCA is suitable for closing prices, which exhibit non-stationary patterns and noise, while DPCA is more appropriate for simple returns and log returns, which exhibit stationarity. This work also indicates that GDPCA can explain the variance with fewer principal components than DPCA, which needs more components to achieve the same level of variance explained. With the suitable variants of PCA, the extracted PCs will serve as useful indicators that could encapsulate various technical aspects of the stocks and equip traders and researchers with the most essential information (Liu 2022; Bruna *et al.* 2022; Kumar 2022; Zhang & Wang 2023; Zhang 2022). For portfolio optimisations using multifactor models, the best asset allocations and portfolio management strategies could also be based upon the PCs from the right forms of PCA (Lòpez de Prado 2020).

The work by Sarıkoç and Celik (2024) introduced a hybrid model for predicting financial asset prices, utilising PCA and independent component analysis for preprocessing, followed by a long-short-term memory (LSTM) network for forecasting. Future research should explore contemporary dimensionality reduction techniques such as t-Distributed Stochastic Neighbor Embedding (t-SNE), Uniform Manifold Approximation and Projection (UMAP), Triangular Manifold Approximation and Projection (TriMAP), and Pairwise Constant Mapping (PaCMAP). These methods are adept at handling complex data and revealing patterns that traditional approaches might overlook, which could enhance model accuracy and understanding. Additionally, employing deep learning methods like autoencoders and Convolutional Neural Networks (CNNs) for dimensionality reduction could offer significant benefits. Although these techniques are not the main focus of this study, they have the potential to provide valuable insights and improve our ability to manage high-dimensional data in future research.

In our future research, we intend to replace PCA with GDPCA for dimensionality reduction and de-noising, as GDPCA is more effective at capturing the temporal dynamics in financial data. Furthermore, we plan to experiment with various deep learning architectures to enhance the accuracy and robustness of our forecasting model. On the other hand, the analysis in this study has motivated us to introduce the functional data framework of GDPCA (Khoo *et al.* 2024), a significant contribution to the field of dimensionality reduction for non-stationary time series data within a functional data framework.

Acknowledgments

This work was supported by the Universiti Malaya, Faculty Research Grant (GPF088A-2020).

References

Ashraf M., Anowar F., Setu J.H., Chowdhury A.I., Ahmed E., Islam A. & Al-Mamun A. 2023. A survey on dimensionality reduction techniques for time-series data. *IEEE Access* **11**: 42909-42923.

Bollerslev T. 1986. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* **31**(3): 307-327.

Brillinger D.R. 1981. *Time series: Data analysis and theory (Holden-Day series in time series analysis).* San Francisco, CA: Holden Day.

Bruna M.G., Loprevite S., Raucci D., Ricca B. & Rupo D. 2022. Investigating the marginal impact of ESG results on corporate financial performance. *Finance Research Letters* **47**: 102828.

Donadelli M. & Paradiso A. 2014. Is there heterogeneity in financial integration dynamics? Evidence from country and industry emerging market equity indexes. *Journal of International Financial Markets, Institutions and Money* **32**: 184-218.

Elliott G., Granger C.W.J. & Timmermann A.G. 2006. *Handbook of Economic Forecasting.* Amsterdam, The Netherlands: North-Holland.

Ghalanos A. 2022. rugarch: Univariate GARCH models. R package version 1.4-9.

Glosten L.R., Jagannathan R. & Runkle D.E. 1993. On the relation between the expected value and the volatility of the nominal excess return on stocks. *The Journal of Finance* **48**(5): 1779-1801.

Hormann S. & Kidzinski L. 2022. freqdom: Frequency Domain Based Analysis: Dynamic PCA. R package version 2.0.3.

Imai J. & Tan K.S. 2006. A general dimension reduction technique for derivative pricing. *Journal of Computational Finance* **10**(2): 129-155.

Jolliffe I.T. 2002. *Principal Component Analysis*. 2nd Ed. New York City, NY: Springer-Verlag.

Kambouroudis D.S., McMillan D.G. & Tsakou K. 2016. Forecasting stock return volatility: A comparison of GARCH, implied volatility, and realized volatility models. *Journal of Futures Markets* **36**(12): 1127-1163.

Khoo T.H., Dabo I.M., Pathmanathan D. & Dabo-Niang S. 2024. Generalized functional dynamic principal component analysis. *arXiv preprint arXiv*: 2407.16024.

Ku W., Storer R.H. & Georgakis C. 1995. Disturbance detection and isolation by dynamic principal component analysis. *Chemometrics and Intelligent Laboratory Systems* **30**(1): 179-196.

Kumar S. 2022. Effective hedging strategy for US Treasury bond portfolio using principal component analysis. *Academy of Accounting and Financial Studies Journal* **26**(1): 1-17.

Liu C. 2022. Risk prediction of digital transformation of manufacturing supply chain based on principal component analysis and backpropagation artificial neural network. *Alexandria Engineering Journal* **61**(1): 775-784.

Lòpez de Prado M.M. 2020. *Machine Learning for Asset Managers*. Cambridge, UK: Cambridge University Press.

Ma F., Guo Y., Chevallier J. & Huang D. 2022. Macroeconomic attention, economic policy uncertainty, and stock volatility predictability. *International Review of Financial Analysis* **84**: 102339.

Mader H.M., Coles S.G., Connor C.B. & Connor L.J. 2006. *Statistics in Volcanology*. London, UK: The Geological Society of London.

Mancino M.E. & Renò R. 2005. Dynamic principal component analysis of multivariate volatility via Fourier analysis. *Applied Mathematical Finance* **12**(2): 187-199.

Marathe R.R. & Ryan S.M. 2005. On the validity of the geometric Brownian motion assumption. *The Engineering Economist* **50**(2): 159-192.

McDonald R.L. 2013. *Derivatives Markets*. 3rd Ed. Upper Saddle River, NJ: Pearson Education.

Nugroho D.B., Kurniawati D., Panjaitan L.P., Kholil Z., Susanto B. & Sasongko L.R. 2019. Empirical performance of GARCH, GARCH-M, GJR-GARCH and log-GARCH models for returns volatility. *Journal of Physics: Conference Series* **1307**(1): 012003.

Pearson K. 1901. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* **2**(11): 559-572.

Peña D., Smucler E. & Yohai V.J. 2020. gdpc: An R Package for Generalized Dynamic Principal Components. *Journal of Statistical Software, Code Snippets* **92**(2): 1-23.

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

Reddy K. & Clinton V. 2016. Simulating stock prices using geometric Brownian motion: Evidence from Australian companies. *Australasian Accounting, Business and Finance Journal* **10**(3): 23-47.

Ryan J.A., Ulrich J.M., Smith E.B., Thielen W., Teetor P. & Bronder S. 2022. quantmod: Quantitative Financial Modelling Framework. R package version 0.4.20.

Sarıkoç M. & Celik M. 2024. PCA-ICA-LSTM: A hybrid deep learning model based on dimension reduction methods to predict S&P 500 index price. *Computational Economics*: 1-67.

Song Z., Gong X., Zhang C. & Yu C. 2023. Investor sentiment based on scaled PCA method: A powerful predictor of realized volatility in the Chinese stock market. *International Review of Economics & Finance* **83**: 528-545.

Wang X. 2006. On the effects of dimension reduction techniques on some high-dimensional problems in finance. *Operations Research* **54**(6): 1063-1078.

Wei J. & Huang J. 2012. An exotic long-term pattern in stock price dynamics. *PLoS One* **7**(12): e51666.

Ying X. 2019. An overview of overfitting and its solutions. *Journal of Physics: Conference Series* **1168**: 022022.

Zhang Y. & Wang Y. 2023. Forecasting crude oil futures market returns: A principal component analysis combination approach. *International Journal of Forecasting* **39**(2): 659-673.

Zhang Z. 2022. Research on stock price prediction based on PCA-LSTM model. *Academic Journal of Business & Management* **4**(3): 42-47.

Zhong X. & Enke D. 2017. Forecasting daily stock market return using dimensionality reduction. *Expert Systems with Applications* **67**: 126-139.

Institute of Mathematical Sciences
Faculty of Science
Universiti Malaya
50603 Kuala Lumpur
Wilayah Persekutuan Kuala Lumpur, MALAYSIA
E-mail: s2025219@siswa.um.edu.my, dharini@um.edu.my[*]*, 17201135@siswa.um.edu.my*

Received: 13 June 2024
Accepted: 31 July 2024

[*] Corresponding author