May-June 2020
ISSN: 0193-4120 Page No. 3972-3976

Forecasting Stock Indexes by Reinitiated Singular Spectrum Analysis Approach

Kong Hoong Lem*, Woan Lin Beh

Faculty of Science, Universiti Tunku Abdul Rahman, Malaysia

Article Info
Volume 83
Page Number: 3972-3976
Publication Issue:
May - June 2020

Article History
Article Received:*19 November 2019*
Revised: *27 January 2020*
Accepted: *24 February 2020*
Publication: *12 May 2020*

Abstract

This paper proposed a reinitiated singular spectrum analysis (rSSA) approach for stock index forecasting. The approach was experimented on some international stock indexes. Forecast performance was benchmarked against the conventional Autoregressive Integrated Moving Average (ARIMA) models using the agreement indexλ. In addition, other common performance metrics such as the weighted mean absolute percentage error (wMAPE) and the root mean of squared errors (RMSE) also served as reference. For the data tested, the rSSA forecast outruns those from ARIMA family. Reinitiated SSA is a potential alternative for stock market index forecasting.

Keywords**:** *Reinitiated singular spectrum analysis, stock index, forecast*

1. Introduction

Singular spectrum analysis (SSA) can be regarded as a non-parametric method of time series analysis. It is a decomposition-based method that incorporates singular value decomposition (SVD) for decomposing the time series into components according to its order of significance. A peculiar feature of SSA lies in the fact that it is model-free and hence does not rely on parametric assumptions such as stationarity and normality (Golyandina, Nekrutkin & Zhigljavsky, 2001) and this very fact turns out to be its advantage since chances for realworld time series data to obey the assumptions are low.

In literature, the initial work on SSA is commonly believed to be attributed to the works by Broomhead and King (1986). The capability of SSA in analyzing, smoothing, denoising, trend and seasonality extracting and forecasting has been popularizing its usage in various areas (such as geophysics, climatology, economics, social sciences, and market research) during the last few decades. Detailed introduction and discussion of singular spectrum analysis can be found in the monographs such as Elsner & Tsonis (1996), Golyandina & Zhigljavsky (2013), Hussain et al., 2015; Sanei & Hassani (2015) and references therein.

In general, there are two common approaches in SSA forecasting, namely the recurrent forecasting approach (RSSA) and the vector forecasting algorithm (VSSA) (Ghodsi, Hassani, Rahmani& Silva, 2017; Sanei & Hassani, 2015; Islam et al., 2017; Joarder et al,. 2015). However here, a new attempt is made to adopt the reinitiated SSA (rSSA) for stock index forecasting.

2. Methodology

The standard SSA technique comprises two stages, namely the decomposition and the reconstruction stages. At the decomposition stage, the time series data is converted into a matrix (known as the trajectory matrix) by partitioning it at a certain length (called the window length) with offset one. The trajectory matrix is then subject to SVD for decomposing into its eigen components. Later during the reconstruction stage, certain non-significant components (quantified by the SVD-level) are ignored and an "inverse-SVD" is performed to get back the smoothed time series using the reverse diagonal averaging manner.

In rSSA forecasting, an arbitrary initial value is appended to the time series data as the one-point forecast before the data undergo the standard SSA stages. At each round of SSA, an updated forecast is obtained which is then reiterated as the initial forecast for the subsequent round of SSA. This is repeated until the forecast converges up to a certain tolerance. Then, the same process is carried out for the next forecast point.

A brief rSSA algorithm is outlined in what follows. It can be divided into four stages.

Stage-1: Appending an initial guess to the time series data.

Suppose that an observed time series data of size N is available $Y_N = [y_1, y_2, \dots, y_N]$. An initial guess $\hat{z}_0$ is

TEST Engineering & Management

May-June 2020
ISSN: 0193-4120 Page No. 3972-3976

appended to the series as the prediction of the next (future) data value such that a time series data of size $(N+1)$ is formed $Y_{N+1} = [y_1, y_2, \dots, y_N, \hat{z}_0]$ where $\hat{z}_0$ is the initial guess for predicting y_{N+1}. The initial guess can be arbitrarily random, for example, the value of y_N or the average of a few preceding data is used.

Stage-2: Decomposition

Here the one dimensional time series (a single column matrix) is segmented into a rectangular matrix (called the trajectory matrix) before undergoing singular value decomposition.

***Step-1*: Embedding time series into trajectory matrix**

A window of length L is chosen $(1 < L < (N+1)/2)$. This window is used to slide over the data Y_{N+1} to form segments that have $L-1$ overlaps. Every segment is a lagged vector of length L. The lagged vectors are stacked up to form the trajectory matrix X of size $L \times M$ where $M = (N+1) - L + 1$. This process is called embedding. It maps a one dimensional time series Y_{N+1} into a multidimensional series $X = [\vec{x}_1, \vec{x}_2, \dots, \vec{x}_M] = (x)_{i,j=1}^{L,M}$ with lagged vectors $\vec{x}_i = [y_i, y_{i+1}, y_{i+2}, .., y_{i+L-1}]^T$ for $i = 1, 2, \dots, M$. If the trajectory matrix X is unraveled, we have

$$X = \begin{bmatrix} y_1 & y_2 & y_3 & \cdots & y_i & \cdots & y_M \\ y_2 & y_3 & y_4 & \cdots & y_{i+1} & \cdots & y_{M+1} \\ y_3 & y_4 & y_5 & \cdots & y_{i+2} & \cdots & y_{M+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ y_L & y_{L+1} & y_{L+2} & \vdots & y_{i+L-1} & \vdots & y_{N+1} = \hat{z}_0 \end{bmatrix} \quad \textbf{(1)}$$

The rows and columns of the trajectory matrix X are subseries of the original series. The trajectory matrix X is a Hankel matrix which means whose entries along the skew diagonal are equal. Equivalently, all the (i,j)th entries in which $(i+j) =$ constant are equal.

***Step-2*: Performing singular value decomposition**

The trajectory matrix X is subject to a singular value decomposition to obtain the U, Σ, V matrices such that

$$X = U\Sigma V^T. \quad \textbf{(2)}$$

The dimensions of the matrices U, Σ, V are $(L \times L), (L \times M), (M \times M)$ respectively. Both U, V are orthogonal i.e. $U^{-1} = U^T$; Σ is a non-negative block diagonal matrix

$$\Sigma = \left[\begin{array}{c|c} D & 0 \\ \hline 0 & 0 \end{array}\right] \quad \textbf{(3)}$$

and D is a diagonal $(r \times r)$ matrix having its main diagonal entries σ_i to be the singular values of matrix X, sorted in descending order i.e. $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r \geq 0$. The number of non-zero diagonal entries decides the rank, r of the trajectory matrix.

Stage-3: Denoising, reconstruction and iterative updating

The time series data Y_{N+1} is very likely to consist of high frequency component and low frequency component. The high frequency component may due to short-term fluctuation, trend or noise whereas the low frequency component may due to long-term trend, seasonal trend or cyclic behavior. The data will undergo certain level of noise reduction or smoothing before reconstruction. The process is repeated iteratively until certain convergence tolerance was achieved.

***Step-1*: Denoising, reconstruction**

The first k largest singular values in the matrix Σ will be retained while the rest are zeroed out such that D is reduced to a $(k \times k)$ diagonal matrix so that a reduced block diagonal matrix Σ_k is formed with $k < r$. The large singular values will carry most of the characteristic of the matrix X whereas the lower singular values contain trivial noise and hence are eliminated. The reduced trajectory matrix is then reconstructed

$$\tilde{X} = U\Sigma_k V^T \quad \textbf{(4)}$$

The last entry of the matrix $\tilde{X}$will give $\hat{z}_1$which is an updated prediction of the data point y_{N+1}. Here, the SVD-level is referred to as $\frac{k}{r} \times 100\%$.

***Step-2*: Iterative updating**

The last entry of the trajectory matrix X (in Stage-2) is replaced by this newer predictive value. The procedure is then repeated iteratively until the prediction converges toward a predetermined tolerance ϵ, that is $|\hat{z}_j - \hat{z}_{j-1}| < \epsilon$ for $j = 1, 2, \dots$ Hereafter, $\hat{z}_j$ will be accepted as the prediction for the data point y_{N+1}.

Stage-4: Proceeding to the prediction of the next data point

The process will re-loop all the way back to Stage-1 again meanwhile an initial guess for predicting the next data point, i.e. y_{N+2} will be appended to the time series data. The subsequent stages and steps then follow until y_{N+q} where q is the desired amount of data point we want to forecast.

1.1 Measuring indices for forecast accuracy

The performance of the forecast is evaluated by the agreementindex λ (Duveiller, Fasbender, & Meroni, 2016) in which

$$\lambda = 1 - \frac{n^{-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sigma_y^2 + \sigma_{\hat{y}}^2 + \left(\bar{y} - \bar{\hat{y}}\right)^2 + \kappa} \quad \textbf{(5)}$$

where

$$\kappa = \begin{cases} 0, & \text{if } r \geq 0 \\ 2\left|\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})\right|, & \text{otherwise} \end{cases} \quad \textbf{(6)}$$

TEST Engineering & Management

May-June 2020
ISSN: 0193-4120 Page No. 3972-3976

Here n is the number of forecast points; y_i and $\hat{y}_i$ are the observed and the forecasted data respectively; $\bar{y} = \sum_{i=1}^{n} y_i/n$ is the mean of the observed data; $\bar{\hat{y}} = \sum_{i=1}^{n} \hat{y}_i/n$ is the mean of the forecasted data; r is the Pearson correlation coefficient while σ_y^2, $\sigma_{\hat{y}}^2$ are the variance of the observed and the forecasted data respectively.

Other auxiliary measurements are the root mean square error (*RMSE*), the mean absolute error (*MAE*) and the weighted mean absolute percentage error (*wMAPE*) which are defined by

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}, \quad (7)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|, \quad (8)$$

$$wMAPE = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{\sum_{i=1}^{n} y_i} \times 100\% . \quad (9)$$

In principle, if the forecast is perfect, then $\lambda = 1$, $RMSE = MAE = wMAPE = 0$ will be obtained. The magnitude of *RMSE* and *MAE* are data dependent while the *wMAPE* only ranges from zero to one hundred percent.

Among the four performance indices used here, the agreement index λ was used as the main index instead of the commonly used *RMSE*. The justification is that in stock index forecasting, one is more concerned about the up and down agreement of the forecast rather than the magnitude of the variation. Moreover, *RMSE* punishes the difference between the observed and the forecasted greatly and is less sensitive to the agreement in their trend (Willmott & Matsuura, 2005).

1.2 Choosing of parameters

In deciding the values of the two tuning parameters window length and SVD-level, rSSA was repeatedly implemented onto the training data set for a range of parameters, and the parameters that bear the smallest agreement index λ would be utilized in the subsequent forecasting. In this study, the SVD-level was about 95% whereas the window length used was about 250.A common suggestion is that the window length should be a value less than half of the data length and proportional to the period of the data (Hassani & Thomakos, 2010). For our daily stock index data, 250 is equivalent to about one year, excluding weekends and holidays.

3. Result and Discussion

The data used in the study was the almost ten-year daily stock closing index. Two stocks were arbitrarily chosen in this study: the Kuala Lumpur Composite Index (KLCI) in Malaysia and the Hang Seng Index (HSI) in Hong Kong. The KLCI data ranged from 20-May 2010 till 1-Aug 2010 while the HIS data ranged from 2-Jan 2009 till 1-Aug 2019.

The temporal plots of these indexes were shown in the following figures. The dashed vertical line marked was the dividing line between the training data and the test data.

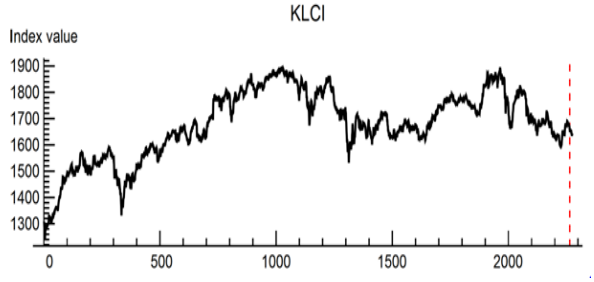

Figure 1: The temporal plot of the Kuala Lumpur Composite, KLCI.

The dotted vertical line separates the training data from the test data.

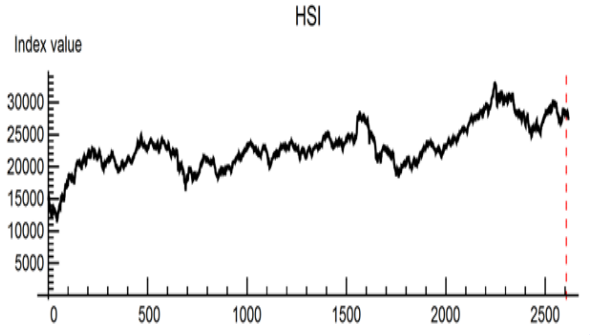

Figure 2: The temporal plot of the Hang Seng Index, HSI. The dotted vertical line separates the training data from the test data.

1.3 The 10-step-ahead forecast

A 10-step-ahead forecast (equivalent to 2 weeks) had been attempted during the period from 19-July 2019 to 1-Aug 2019. Conventional ARIMA-family (Brockwell & Davis, 2002) forecasting had also been implemented onto the same data range for comparison.

Table 1: The 10-step-ahead forecasting performance assessed using four performance indices for HIS and DJI: comparison between rSSA and ARIMA.

	Kuala Lumpur Composite Index		Hang Seng Index	
	rSSA	ARIMA	rSSA	ARIMA
λ	0.457548	0.00011326	0.401364	0.
RMSE	14.1667	10.4958	340.933	400.63
MAE	12.1652	7.6287	282.776	290.15
wMAPE (%)	0.737655	0.462576	1.00021	1.0263

The table above (Table 1.) compares the rSSA forecast performance with those from ARIMA. In terms of the agreement index λ, it is obvious that the forecasting

May-June 2020
ISSN: 0193-4120 Page No. 3972-3976

accuracy of rSSA outperform those by ARIMA in both data.

The following figures showed short segments of the observed data and their corresponding forecast. The rSSA forecasts apparently matched up better with the observed data than those from the ARIMA approach in both the stock indexes.

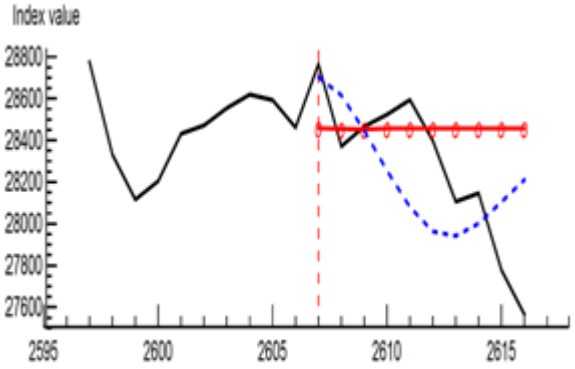

Figure 3: Temporal plots of the observed data (solid line) and the forecasted data for KLCI: comparison between rSSA (dashedline) and the ARIMA (line with circular markers) for the 10-step-ahead forecast.

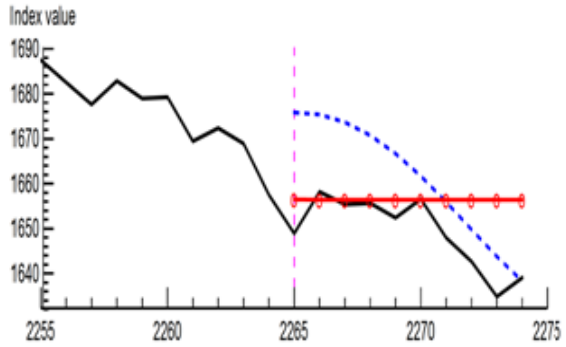

Figure 4: Temporal plots of the observed data (solid line) and the forecasted data for HSI: comparison between rSSA (dashed line) and the ARIMA (line with circular markers) for the 10-step-ahead forecast.

As can be seen from the plotting above, in both cases, the ARIMA approach tends to yield rather fixed-value forecasts which appear as horizontal lines to some extent, and those lines are visually the regional average within the forecast range. In this sense, rSSA forecasts seem more rational in virtue of exhibiting some degree of up-down wavering rather than flat straight lines.

For the KLCI data, the *RMSE* from ARIMA is smaller than the one from rSSA. With respect to *RMSE*, this is supposed to mean that ARIMA gives a better forecast than rSSA. However the graph in Fig. 3 reveals that *RMSE* is a questionable measurement for forecast performance. Failing to capture the closeness between the observed and the forecasted data is a shortcoming of *RMSE*.

4. Conclusion

This study outlined a generic algorithm for rSSA approach and implements it onto the stock indexes of Kuala Lumpur and Hang Seng. The comparison of forecasting results revealed that rSSA performs better than the ARIMA models based on the agreement index measure. The results suggest rSSA to be a feasible alternative for stock index forecasting.

Acknowledgments

The authors gratefully acknowledge the financial support of the University.

References

[1] Brockwell, P. J., & Davis R. A. (2002). *Introduction to Time Series and Forecasting* (2nd ed.). New York: Springer-Verlag.

[2] Broomhead, D., & King, G. P. (1986). Extracting qualitative dynamics from experimental data. *Physica D: Nonlinear Phenomena*, *20*(2-3), 217-236.

[3] Duveiller, G., Fasbender, D., &Meroni, M.(2016).Revisiting the concept of a symmetric index of agreement for continuous datasets. *Scientific Reports*,*6*, 19401.

[4] Elsner, J. B., & Tsonis, A. A. (1996). *Singular Spectrum Analysis: A New Tool in Time Series Analysis*. US: Springer.

[5] Ghodsi, M., Hassani, H., Rahmani, D., & Silva, E. S. (2017). Vector and recurrent singular spectrum analysis: which is better at forecasting? *Journal of Applied Statistics*, *45*(10), 1872-1899.

[6] Golyandina, N., Nekrutkin, V., & Zhigljavsky, A. (2001). *Analysis of time series structure: SSA and related techniques*. New York: Chapman & Hall/CRC.

[7] Golyandina, N., &Zhigljavsky, A. (2013).*Single spectrum analysis for time series*.Berlin Heidelberg: Springer-Verlag.

[8] Islam, R., Ghani, A.B.A., Abidin, I.S.Z., Sundari, A. (2017). Effects Of Minimum Wage Rate Towards The Unemployment Rate. Journal of Applied Economic Sciences, 12 (1), pp. 206-221.

[9] Joarder, M.H.R., Subhan, M., Ghani, A.B.A., Islam, R. (2015). Pay, Security, Support And Intention To Quit Relationship Among Academics In Developing Economy. Investment Management and Financial Innovations, 12 (3), pp. 190-199.

[10] Hassani, H., & Thomakos, D. (2010). A review on singular spectrum analysis foreconomic and financial time series. *Statistics and Its Interface*, *3*, 377–397

May-June 2020
ISSN: 0193-4120 Page No. 3972-3976

[11] Hussain, A., Mkpojiogu, E.O.C. (2015). The Effect Of Responsive Web Design On The User Experience With Laptop And Smartphone Devices.JurnalTeknologi, 77 (4), pp. 41-47.

[12] Sanei, S., & Hassani, H. (2015). *Singular Spectrum Analysis of Biomedical Signals*. Boca Raton: CRC Press.

[13] Willmott, C. J.,& Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessingaverage model performance. *Climate Research, 30*, 79–82.