International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 November 2023*

Feature Extraction using Singular Spectrum Analysis: Characterizing Dominant Modes for Time Series Forecasting

[1]Iftikhar U. Sikder
[1] Cleveland State University
[1] i.sikder@csuohio.edu

Abstract

Abstract-- This study explores the application of Singular Spectrum Analysis (SSA) for feature extraction from the dominant modes of an industry sector. These modes are hypothesized to encapsulate the underlying market trends and cycles, offering an enhanced understanding of stock price dynamics. The methodology involves identification of dominant modes of historical stock price data from leading semiconductor companies, and applying Singular Spectrum Analysis (SSA) to identify and isolate the relevant features contributing to price dynamics. Finally, the features extracted are used to forecast a new time series in the same sector using Elastic Net Regression. The forecasting evaluation metrics indicates lower error rates and high predictive accuracy.

Index Terms: Singular Spectrum Analysis, Singular Value Decomposition, Semiconductor Industry, Time Series Decomposition, Dominant Modes

I. INTRODUCTION

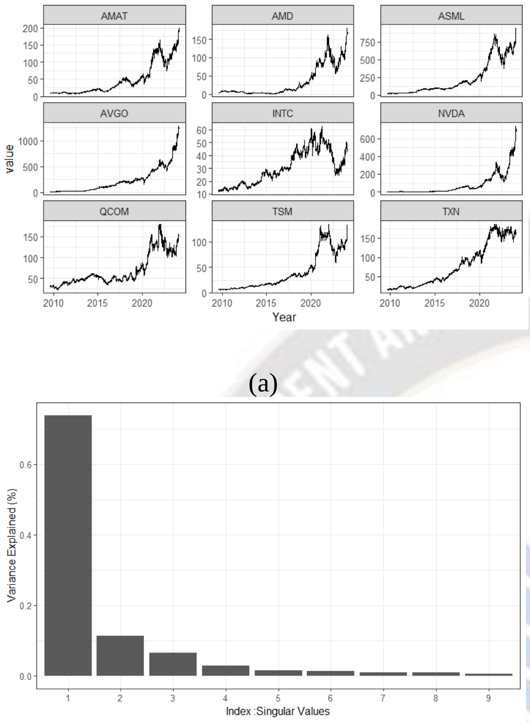
Forecasting stock market movements stands as a formidable challenge, one that is of paramount importance to many stakeholders. This challenge is particularly pronounced in the volatile semiconductor sector, where stock prices are influenced by a myriad of factors, from technological advancements and regulatory changes to global supply chain dynamics. Given the sector's susceptibility to rapid technological advancements, regulatory changes, and global supply chain dynamics, traditional linear forecasting models often fall short in capturing the complex, non-linear interactions that characterize semiconductor stock movements. The challenge is to identify and characterize the dominant patterns or mode of the sector of semiconductor. In the realm of financial time series, the "dominant modes" identified through singular value decomposition encapsulate the most influential trends and patterns driving stock movements. This paper aims to explore Singular Spectrum Analysis (SSA) in extracting meaningful features from time series data and to investigate how these features can be leveraged to improve forecasting outcomes. SSA provides a non-parametric framework that decomposes time series into constituent components, allowing for the isolation of the signal from the noise. Its ability to handle complex, non-linear, and non-stationary data makes it a valuable method in the forecaster's toolkit. We hypothesize that by identifying and characterizing the dominant modes within a time series, it is possible to construct forecasting models that are not only accurate but also more interpretable. The paper is structured as follows: section II reviews the literature on time series forecasting in financial markets, with a focus on the application of SSA in this domain. Section III describes the methodology, detailing the SSA process for extracting dominant modes and the rationale behind focusing on the significant left singular vectors. Section IV presents the empirical results, illustrating the forecasting models in the context of semiconductor stocks. Finally, Section V discusses the broader implications of our findings and conclusion.

II. LITERATURE REVIEW

Historically, time series forecasting in financial markets has relied heavily on linear models, such as ARIMA and its variants. However, the volatile nature of stock markets, characterized by non-linear dynamics and influenced by a plethora of external factors, often renders these traditional models less effective. Singular Spectrum Analysis (SSA) emerged as a non-parametric method that decomposes time series into a set of interpretable components, including trend, oscillatory modes, and noise. It was introduced in the climate sciences, demonstrating its prowess in identifying hidden periodicities and trends in climatological time series[1]. The adaptability and effectiveness of SSA in handling non-stationary and non-linear data soon garnered attention in the field of financial time series analysis[2-4]. The ability of SSA to dissect complex financial time series into dominant modes—significant left singular vectors—has been particularly noteworthy. These modes encapsulate crucial market trends and cycles, offering insights that are invaluable

***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 November 2023*

for forecasting. Recent studies have begun to explore the application of SSA in forecasting stock prices, recognizing its potential to outperform traditional models in capturing the multifaceted dynamics of financial markets[5, 6].



(a)

(b)

Fig.1 (a) Major semiconductor stocks (b) variance explained by singular values after performing SVD

III. METHODOLOGY

The semiconductor stocks used in this research include nine (9) major stocks, namely Advanced Micro Devices(AMD), Intel (INTC), Nvidia (NVDA) Taiwan Semiconductor Manufacturing company (TSM), Broadcom (AVGO), ASML holding (ASML), Applied Material (AMAT), Qualcomm (QCOM), Texas Instruments (TXN). The daily adjusted stock data were collected ranging from date 08-07-2009 to 02-22-2024 to form time series. Fig.1.a shows the time series plot of the stocks. In order to extract the dominant mode of semiconductor stocks, singular value decomposition (SVD) was perform on the time series matrix M.

$$\mathbf{M} = \mathbf{U\Sigma V^T} \quad (1)$$

Where, columns of U are orthonormal eigenvectors of $\mathrm{MM^T}$ with the equal length of the time series. The columns of V are eigenvectors of $\mathrm{M^TM}$. It is possible to approximately reconstruct matrix M using smaller number of vectors and singular values as:

$$\mathbf{M_k} = \boldsymbol{\sigma_1}\mathbf{u_1}\mathbf{v_1^T} + \boldsymbol{\sigma_2}\mathbf{u_2}\mathbf{v_2^T} + \cdots + \boldsymbol{\sigma_k}\mathbf{u_k}\mathbf{v_k^T} \quad (2)$$

For example, one can reconstruct the time series data frame using only with possible $\mathrm{u_1, u_2}$ (the first two columns or left singular eigenvectors) and $\mathrm{v_1, v_2}$ and associated singular values σ_1 and σ_2. Fig. 1.b shows the percentages of variance explained by the eigenvectors corresponding to their singular value indices. The first singular value explains more that 73.9% of variance and first two singular values jointly represents more than 85% of variance. Hence, $\mathrm{u_1}$ vector can be regarded as most dominant mode of the semiconductor stocks followed by $\mathrm{u_2}$. Characterizing these dominant modes provide significant market dynamics that can be used to extract features for forecasting. These dominant modes are used in Singular Spectrum Analysis (SSA) to characterize the features such as trends and cycles associated with each mode and remove noise. More specifically, we use log of negative $\mathrm{u_1}$ as input into SSA.

Singular Spectrum Analysis (SSA) is a non-parametric method used in time series analysis and forecasting that decomposes a time series into a sum of interpretable components such as trend, oscillatory components, and noise. It is based on the singular value decomposition (SVD) of a trajectory matrix constructed from the time series, allowing for the identification and separation of these components without prior assumptions about their form or the presence of a specific model. Given a time series $\mathrm{Y} = \{\mathrm{y_1, y_2, \ldots, y_n}\}$ the first step in SSA is to embed Y into a higher-dimensional space by forming a trajectory matrix. This is done by sliding a window of length L (where $1 < \mathrm{L} < \mathrm{N}$)) across the series, creating a set of lagged vectors. We used R package Rssa [7] for embedding, decomposition and reconstruction of time series.

A. Step 1: Embedding

The embedding involves construction of a trajectory matrix **C** as follows:

$$\mathbf{C} = \begin{bmatrix} \mathbf{y_1} & \mathbf{y_2} & \mathbf{y_3} & \cdots & \mathbf{y_K} \\ \mathbf{y_2} & \mathbf{y_3} & \mathbf{y_4} & \cdots & \mathbf{y_{K+1}} \\ \mathbf{y_3} & \mathbf{y_4} & \mathbf{y_5} & \cdots & \mathbf{y_{K+2}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{y_L} & \mathbf{y_{L+1}} & \mathbf{y_{L+2}} & \cdots & \mathbf{y_N} \end{bmatrix}$$

Where, $\mathrm{K} = \mathrm{N} - \mathrm{L} + 1$. This matrix C is of size $\mathrm{L} \times \mathrm{K}$. The selection of L depends on the length of the time series. We used Toeplitz variant of SSA with $\mathrm{L} = 100$ to accommodate both the trend and the seasonality.

B. Step2: Decomposition

The trajectory matrix C is decomposed using Singular

***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 November 2023*

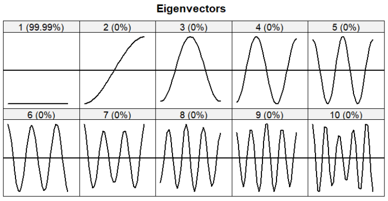
Value Decomposition(SVD), which factorizes C into the product of three matrices:

$$\mathbf{C} = \mathbf{U\Sigma V^T} \quad (3)$$

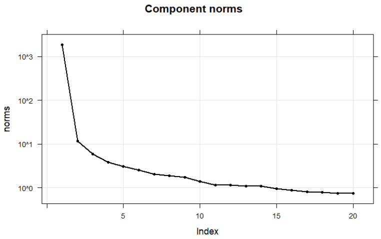
where $\mathbf{U}$ is an $\mathbf{L} \times \mathbf{L}$ orthogonal matrix whose columns are the left singular vectors of $\mathbf{C}$, $\mathbf{\Sigma}$ is an $\mathbf{L} \times \mathbf{K}$ diagonal matrix with non-negative real numbers on the diagonal, known as the singular values of $\mathbf{C}$, sorted in descending order. $\mathbf{V}$ is a $\mathbf{K} \times \mathbf{K}$ orthogonal matrix whose columns are the right singular vectors of $\mathbf{C}$. $\mathbf{V^T}$represents the transpose of $\mathbf{V}$. It is a decomposition of from $\mathbf{C} = \sum_{\mathbf{i}}^{\mathbf{d}} \sqrt{\boldsymbol{\lambda}}\mathbf{UV^T}$ where $\boldsymbol{\lambda}_\mathbf{i}(\mathbf{i} = \mathbf{1}, \dots, \mathbf{L})$ are eigenvalues of the matrix $\mathbf{CC^T}$ in decreasing order of magnitudes. The collection $\sqrt{\boldsymbol{\lambda}_\mathbf{i}}, \mathbf{U_i}, \mathbf{V_i}$ is called ith eigentriple of the matrix $\mathbf{C}$.

C. Step3: Selection of eigenvectors

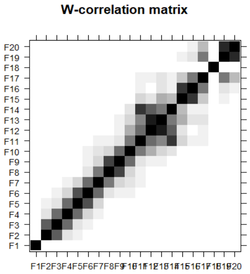
The result of the SVD is a set of triplets $\sqrt{\lambda_i}, U_i, V_i$ where each triplet corresponds to a principal component of the time series. The next step is to group these components into meaningful categories, such as trend, oscillatory components, and noise. Fig. 2a shows the first 10 eigenvectors of C corresponding to singular values along the eigenspectrum. The first component corresponds to the trend and the remaining components is associated with cyclicality. High frequencies and noises are associated with increasing index of the components. The scree plot of SSA (in Fig. 2b) shows the eigenvalues of a correlation matrix which clearly shows that first eigentriple is the most important one.



(a)



(b)



(c)

Fig. 2 (a) first 10 eigenvectors of **C** , (b) scree plot of SSA (c) W-correlation Matrix first 20 components

The W-correlation plot (Fig. 2 c) shows the first 20 weighted correlations between decomposed matrices (i.e. eigentriplets). The plot shows that the first component is w-uncorrelated with the other components, hence the first eigentriple can be used to describes the trend. The trend is reconstructed from the eigentriple (1). Fig. 3 shows the original dominant mode of semiconductor stocks along with the reconstructed trends.

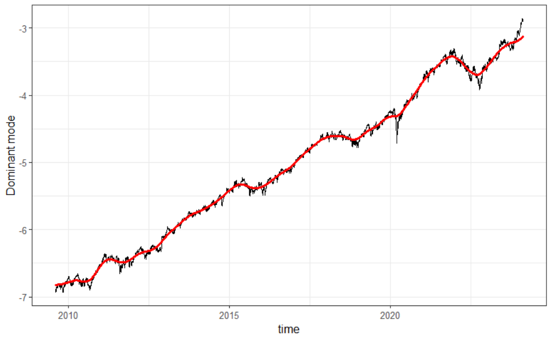


Fig. 3 Dominant mode of semiconductor stocks and reconstructed trend from SSA (in red)

D. Step 4: Diagonal Averaging

After selecting and combining the components of interest, the next step is to transform these components back into the time domain from the trajectory space. This is done through a process called diagonal averaging, which reconstructs the time series $(\tilde{Y}^{(k)} = (\tilde{y}_1^{(k)}, \dots, \tilde{y}_N^{(k)})$ from the possibly modified trajectory matrix. The reconstructed components are then summed as follows to produce the final forecast or to analyze the time series' structure:

$$\mathbf{y_n} = \sum_{\mathbf{k=1}}^{\mathbf{m}} \tilde{\mathbf{y}}_\mathbf{n}^{(\mathbf{k})} \qquad (4)$$

Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 November 2023*

By observing the diagnostics in Fig. 3, five groups of eigentriple pairs were developed. The first eigentriple is used to reconstruct the trend while the eigentriples with pairs (2,3), (4,5), (6,7), (8,9) are used for extracting harmonic components.

E. Forecasting using Features of Dominant Modes:

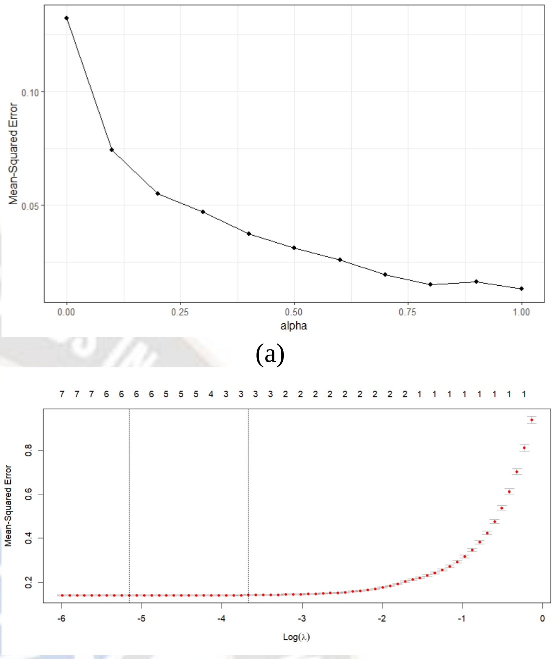
The features extracted from the dominant modes of semiconductor stocks using SSA are the trends and various harmonic components with the removal of high frequency components represent noises. Once the significant components (trend, seasonality) are identified and isolated, SSA can forecast future values by extrapolating these components into the future. This approach is particularly effective for time series with well-defined and stable patterns. In particular, the SSA features extracted from left singular vectors ($\mathbf{u_1}$ and $\mathbf{u_2}$) can be used forecast a new time series (i.e., stock) not include in the computation of SVD in the first place. As for test case, the adjusted daily stock of Micron Technology (MU) was used for forecasting. Using Elastic Net Regression, the objective function for the Gaussian family is:

$$\min_{(\beta_0,\beta)\in\mathbb{R}^{p+1}} \frac{1}{2N}\sum_{i=1}^{N}(y_i-\beta_0-x_i^T\beta)^2+\lambda[(1-\alpha)||\beta||_2^2/2+\alpha||\beta||_1] \quad (5)$$

The value of α ranges between 0 and 1; where it's a ridge regression if α = 0 and lasso regression when α = 1. The $\lambda \geq 0$ is a model complexity parameter which controls the overall strength of the penalty. The $\boldsymbol{y_i}$ is the target time series (i.e., scaled MU time series) and $\boldsymbol{x_i}$ are the features extracted from SSA. We used R package glmnet [8] to choose the best λ and α for the model, through cross-validation (e.g., using time series cross-validation techniques to preserve the temporal order of observations). By incrementally increasing α a large number of models were developed and the corresponding Mean Squared Error (MSE) were recorded from the predicted models. The α value associated with the minimum MSE is selected (Fig. 4 a). For selection of λ, the k-fold cross validation of glmnet was used to produce a plot of MSE against log(λ) with upper and lower standard deviation curves along the λ sequence (error bars). The λ is selected through regularized model so that the cross-validated error is within one standard error of the minimum.

IV. Research Results

The dominant modes derived from SVD of semiconductor stocks explains significant variance of market dynamics. The singular values associated with the first two left singular vectors explain more that 85% variance. For each dominated mode SSA is applied and corresponding trend and harmonic components are identified leaving out the noise. These features are used as covariates in Elastic Net Regression. Fig. 4 a shows response MSE associated with models corresponding to sequences of α. The minimum MSE is found at α=1, suggesting a LASSO regression. Fig. 4.b shows that cross-validation curve (red dotted line) along the λ sequence (with error bars). The λ value associated with the most regularized model was found 0.0255.

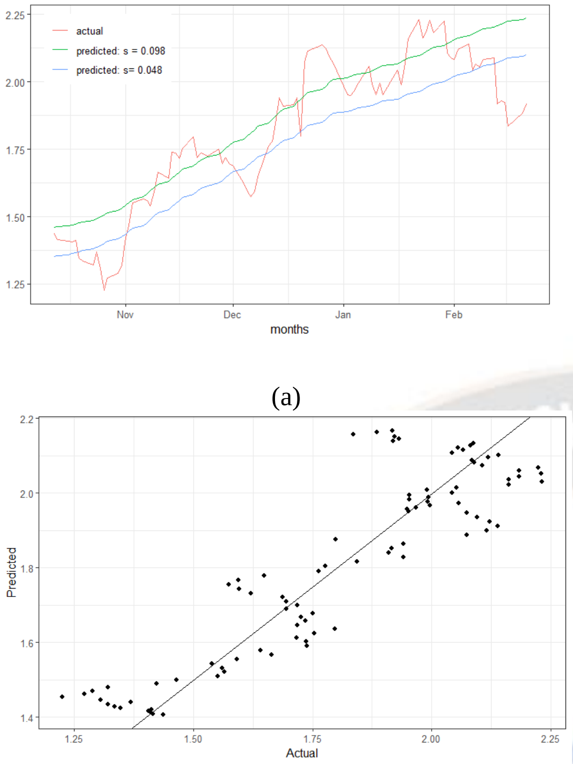


(a)

(b)

Fig. 4 (a) The MSE associated with models of sequences of α (b) The MSE associated with regularization parameters λ.

It implies that the cross-validated error is within one standard error of the minimum. The small value of λ indicates that it reduces the regularization effect, allowing more non-zero coefficients or larger absolute values of coefficients. In glmnet, after fitting the model, at the time of prediction one can specify the parameters to indicate at which λ value(s) to make predictions. It allows users to interact with the glmnet model at specific points along the λ path, enabling detailed inspection and application of the model at chosen levels of regularization. We used a vector of s score (0.048, 0.09) to represent λ. Fig. 5.a shows the actual time series of the holdout set (of scaled MU) of consecutive 90 days' period and the corresponding predicted scores of associated s parameters (i.e., specified λ). The predicted and actual values compared in Fig. 5 b.



(a)

(b)

Fig. 5 (a) The Model forecasting and comparison with test set (b) alignment with predicted and holdout set

Finally, we evaluate the forecasting performance using metrics such as MAE (Mean Absolute Error), RMSE (Root Mean Square Error), or MAPE (Mean Absolute Percentage Error). The forecasting performance shown Table I. indicates good forecasting performance with very low MAE and MSE. The MAPE is around 5.25%, signifying on average, the forecasted values deviate from the actual values by 5.25%, indicating that the model is capable of making forecasts that are close to the actual time series.

Table I. Model Evaluations

Forecast Metrics	**Formulae**	**Experimental Score**
Mean Absolute Error (MAE)	$MAE = \frac{1}{n}\sum_{i=1}^{n}\lvert Y_i - \hat{Y}_i\rvert$	0.0928
Mean Squared Error (MSE)	$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$	0.0143
Mean Absolute Percentage Error (MAPE)	$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left[\frac{Y_i - \hat{Y}_i}{Y_i}\right]$	5.25%
Root Mean Squared Error (RMSE)	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}$	0.12

V. Conclusion

The dominant modes derived from semiconductor sector are capable of generating features using SSA which separates the time series into components, making it easier to identify and isolate the signal (trend and periodic components) from the noise. This is crucial for accurate forecasting, as it allows the forecaster to focus on the underlying patterns rather than the random fluctuations. By isolating the trend component, SSA helps in understanding the long-term direction of the time series, which is vital for making long-term forecasts. SSA can identify harmonic or cyclic components, which are essential for short-term and medium-term forecasting, especially in time series with clear periodic patterns. Since SSA does not rely on a specific parametric model, it can reconstruct the series based on the identified components, making it versatile and adaptable to various types of time series data. The model evaluation metrics indicates that the dominant modes of a sector can be employed for forecasting related stocks with adequate accuracy.

REFERENCES

[1] R. Vautard, P. Yiou, and M. Ghil, "Singular-spectrum analysis: A toolkit for short, noisy chaotic signals," Physica D: Nonlinear Phenomena, vol. 58, pp. 95-126, 1992.

[2] H. Hassani and D. Thomakos, "A review on singular spectrum analysis for economic and financial time series," Statistics and its Interface, vol. 3, pp. 377-397, 2010.

[3] Q. Tang, R. Shi, T. Fan, Y. Ma, and J. Huang, "Prediction of financial time series based on LSTM using wavelet transform and singular spectrum analysis," Mathematical Problems in Engineering, vol. 2021, pp. 1-13, 2021.

[4] N. Golyandina, A. Zhigljavsky, N. Golyandina, and A. Zhigljavsky, "SSA for forecasting, interpolation, filtration and estimation," Singular Spectrum Analysis for Time Series, pp. 71-119, 2013.

[5] M. Škare and M. Porada-Rochoń, "Multi-channel singular-spectrum analysis of financial cycles in ten developed economies for 1970–2018," Journal of Business Research, vol. 112, pp. 567-575, 2020.

[6] J. Arteche and J. García-Enríquez, "Singular spectrum analysis for signal extraction in stochastic volatility models," Econometrics and statistics, vol. 1, pp. 85-98, 2017.

[7] N. Golyandina, A. Korobeynikov, A. Shlemov, and K. Usevich, "Multivariate and 2D extensions of singular spectrum analysis with the Rssa package," arXiv preprint arXiv:1309.5050, 2013.

[8] T. Hastie, J. Qian, and K. Tay, "An introduction to glmnet," CRAN R Repositary, vol. 5, pp. 1-35, 2021.