Financial time series prediction with optimized empirical mode decomposition

P. Syamala Rao[1], G. Parthasaradhi Varma[2]

[1] Research Scholar, Department of CSE, Acharya Nagarjuna University, Guntur, India
[2]Vice-Chancellor, KL University, Vijayawada, India.

Abstract. This paper introduce optimized empirical mode decomposition (EMD) based regression model to predict future samples of financial time series. Initially, time series data is processed through EMD to decompose the information into intrinsic mode functions (IMF's) and residual component. Later, a multi linear regression (MLR) models are developed with the help of IMF's and residual component. To improve the overall accuracy, optimal values of the MLR models are required. Therefore, population search based optimization algorithm is adopted in this paper to obtain optimal coefficients of MLR models. Results are compared with direct MLR models, wavelet transform (WT)-MLR models to show the efficacy of the proposed work.

Keywords: Empirical mode decomposition, wavelet transform, PSO, regression.
DOI Number: 10.48047/nq.2022.20.19.NQ99073 **NeuroQuantology 2022; 20(19):787-794**

1 Introduction

Application of time series forecasting found to be important in all areas to improve the mechanisms, operations, and efficiencies of the processes [1]. Several applications are found in engineering [2]-[5], agriculture [6]-[7], medicine [8]-[9] and other domains. In particular, forecasting of financial data helpful to improve future pricing and increase the revenue. Based on size, nature and structure of the data, linear and nonlinear models are suggested in earlier research [2]-[18].

Some of such models are listed in literature along with applications as follows: In [2], MLR is suggested for prediction of time series. Electrical load forecasting is done in [2] with MLR approach using EXCEL tool. The features of the time data are extended using signal processing techniques and better prediction outputs are expected from the outputs. Such method is reported in [3] applied for wind speed prediction. For multiple data sets, the relations between variables decide accuracy of the forecasting models. Such models are implemented in [4] using invasive weed optimization (IWO) and particle swarm optimization (PSO) techniques for the geotechnical engineering applications. Further, these algorithms are used in the process of weights updating mechanisms of artificial neural networks (ANN) to improve the forecasting results of the time series with volatile in nature [5].

Apart from engineering fields, these approaches are extensively applied in agriculture [6]-[7] and medical fields [8]. Support vector machine (SVM) is another powerful tool used for classification and regression purpose. In [9], SVM is used for prediction of COVID 19. Support vector regression (SVR) is used in [10] along with independent component analysis (ICA) to enhance the forecasting results of financial data. Ridge regression [11], fuzzy cognitive maps [12] and deep learning frameworks are other methods available in literature used for prediction studies. In recent years, signal processing techniques such as empirical mode decomposition (EMD), ensemble empirical mode decomposition (EEMD) and empirical wavelet transform (EWT) are used for pre-process the time data to improve the forecast results [14]-[15]. Other advanced methods used in forecasting studies are reported in [16]-[19].

EMD based optimized regression model is presented in this paper to forecast financial time series. Initially, the data is processed thorough EMD and extracted IMFs along with residual component are used for fitting model. The optimal values of the coefficients of the regression model are obtained by using PSO algorithm. This approach reduces errors between actual and predicted samples of time series.

2 Empirical mode decomposition (EMD)

EMD is a powerful signal processing technique extensively used in electrical and electronics applications. It decomposes the time domain data into several oscillatory modes called intrinsic mode functions (IMFs) and residual component. These IMFs are more helpful to predict the future samples of the time series. Let $x(t)$ is given time series data processed through EMD, produce n number of IMFs and one residual component. Mathematically,

$$x(k) = \sum_{i=1}^{n} d_i(k) + r(k) \tag{1}$$

In equation (1), d_i is the IMF of i^{th} oscillatory mode and r is the residual component. Based on the extracted IMFs and residual components, an output equation is framed given by

$$y = \sum_{i=0}^{n} \beta_i x_i + \beta_0 \tag{2}$$

In Equation (2), y is the output variable and $x_i(i = 1,2..n)$ are the IMFs. $x_i(i = 0)$ represents residual component.

3 Optimized EMD-MLR model

Suppose the time series data 'x' is processed though EMD and n IMFs are extracted in addition to residual component as described in section 2.

$$y_k = [\beta_1 \quad \beta_2 \ldots \quad \beta_n] \begin{bmatrix} IMF1(k-1) \\ IMF2(k-1) \\ \ldots \\ IMFn(k-1) \end{bmatrix} + \beta_0 r_0(k-1) + \alpha \,, k = 2,3,..N \tag{3}$$

In equation (3), $\beta_0, \beta_1, \ldots \beta_n$ are the regression model coefficients and r_0 is the residual component. N is the data size. The output value at k^{th} sample is identified by using the information at

$k-1^{th}$ sample. However, identification of optimal coefficients to fit the time series data is another challenging task in the prediction studies. Different solutions are possible at each data sample and these solutions may not be suitable for rest of the samples of the same data. By using these errors, an objective function is framed to apply optimization algorithms. The error at k^{th} sample of the data is given by

$$E(k) = (y_{actual}(k) - y_{predicted}(k))^2 \quad (4)$$

For the entire data cycle with 'N' samples, the final error function of the optimal EMD regression model is

$$J = \frac{1}{N}\sum_{k=1}^{N} E(k)^2 \quad (5)$$

The minimum value of the error function is attained at a particular solution set known as final solution achieved at the final iteration of the optimization algorithm. In this paper, PSO adopted for identification of the optimal coefficients of the regression models. According to PSO, the procedure to apply it for the proposed problem is understand with following steps.

Step-1: Initialization of PSO parameters like population size, iteration number, inertia (ω), and acceleration coefficients (c_1 *and* c_2).

Step-2: Defining dimension size of the problem and limits of the variables.

Step-3: For first iteration, solutions are randomly generated, and corresponding fitness values are calculated.

Step-4: Updating the positions of the particles (solutions) using position (x) and velocity (v) equations of the PSO given by

$$v_n^{i+1} = \omega v_n^i + c_1 r_1\left(p_{best_i} - x_n^i\right) + c_2 r_2\left(g_{best_i} - x_n^i\right) \quad (6)$$

$$x_n^{i+1} = x_n^i + v_n^{i+1} \quad (7)$$

Step-5: Calculation of fitness functions of new solutions and check whether all iterations are completed or not to display final solution of the problem. If no, then repeat from step 4

4 Results

A financial time series data set is considered to validate the proposed method whose time-data information is provided in Fig 1. The data set consist of 768 samples ($N = 768$). For identification of optimal EMD regression model, only 720 samples are considered and rest is to verify the accuracy of the method. After processing the data through EMD, IMFs and residual components are presented in Fig 2.

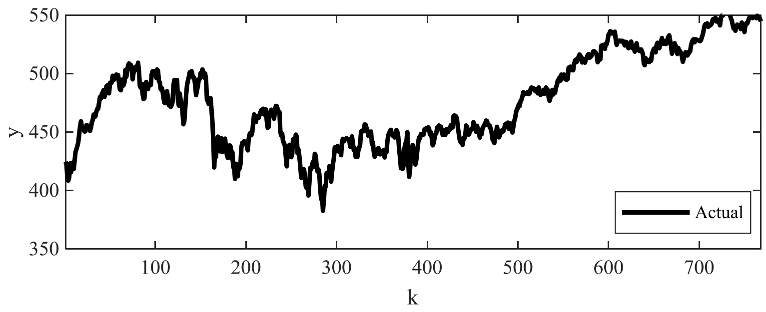

Fig. 1: Time series data (test data)

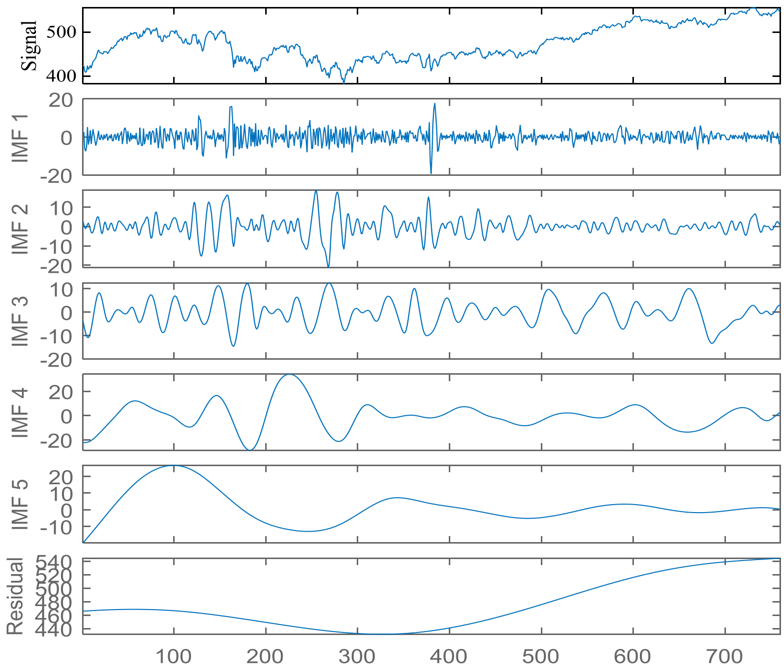

Fig. 2: EMD decomposition of the data

The time series data produces 5 IMFs and residual component. Initially, individual components usage in prediction studies are addressed in this paper later combined model is developed. For this purpose, only IMF 1 is considered as input and the regression model obtained with IMF1 is given by

$$y(k) = 0.2515\, IMF1(k-1) + 472.1485 \tag{8}$$

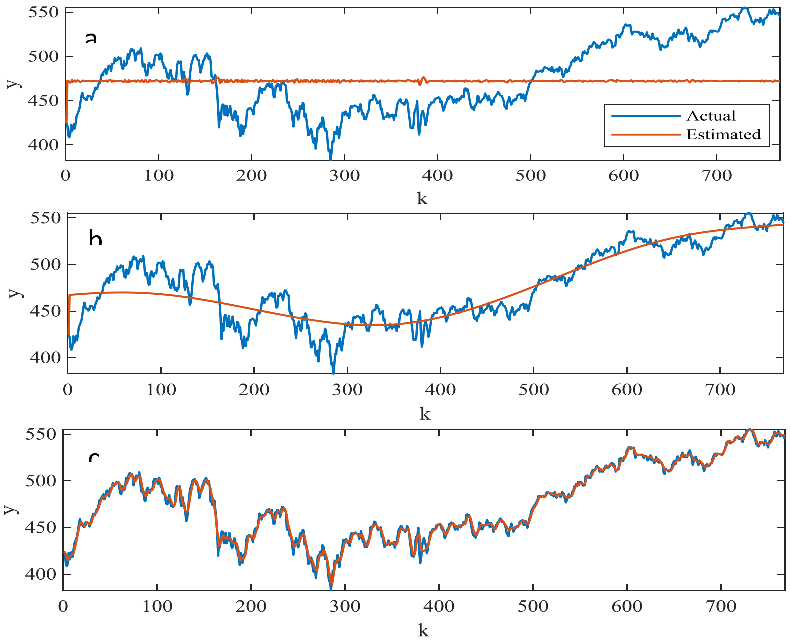

Fig. 3. Estimated time series for model a. with IMF1, b. with residual component, c. with all IMFs and residual component

This model is suitable according to the Fig 3.a since the estimated values are close to 472.1485 and the influence of unique IMF1 on prediction model is worst. This is evident from Fig 3.a. Further, similar study is carried out with residual component and the actual and estimated values of time series data are plotted in Fig 3.b. The regression eqaution of this case is

$$y(k) = 0.9538\, r(k-1) + 22.9949 \tag{9}$$

The two models mentioned in equation (8) and (9) are not providing accurate estimation of future samples. However, a model with all IMFs and residual component produces best results. The estimated values are close to actual data points as shown in Fig 3.c. The model of this equation is

$$y(k) = [0.1714 \quad 0.8904 \quad 0.9637 \quad 1.0006 \quad 0.9978\,] \begin{bmatrix} IMF1(k-1) \\ IMF2(k-1) \\ IMF3(k-1) \\ IMF4(k-1) \\ IMF5(k-1) \end{bmatrix} + 1.0033 r_0(k-1) - 1.3041$$
(10)

This final equation (10) provides reliable outputs for both the data points used in training and testing. The errors of all 768 samples are provided in Fig 4 for the model equation (10). These results in Fig 3.c and Fig 4 show the effectiveness of applying EMD and using all IMFs in the prediction models. In terms of fitness values, the minimum value is attained for third model and this value is large for other 2 models. This comparison is provided in Fig 5.

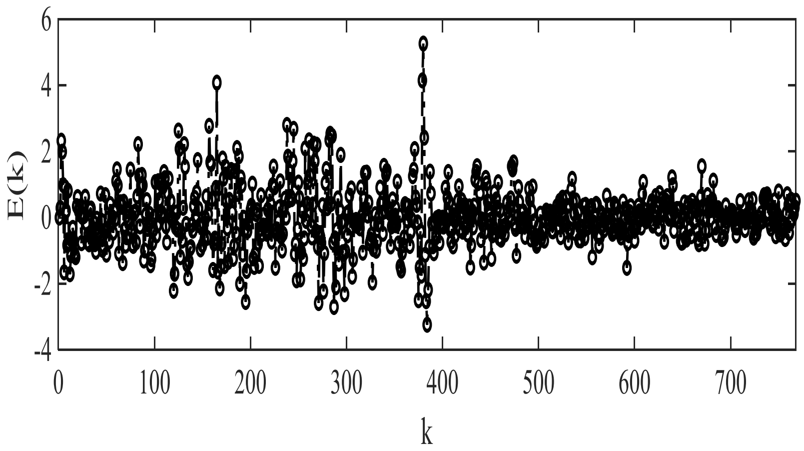

Fig. 4: Errors of all data points with final optimized EMD regression model

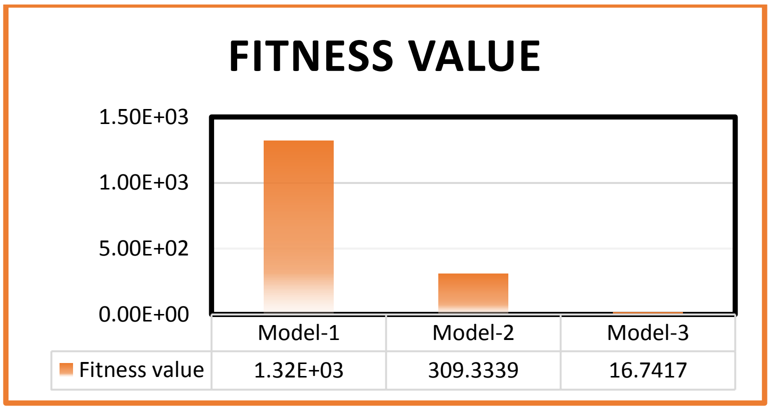

Fig. 5: Fitness function values of various EMD regression models

5 Conclusions

Optimized EMD based regression model is developed in this paper to predict financial time series. Instead of using time data directly, EMD is applied on given time series to find all IMFs and residual component. These IMFs , residual components are further used as inputs to find regression model of the time series and the optimal coefficients of the regression equation are obtained by using PSO. Results show the prediction errors are reduced with the proposed work.

References

1. Hurvich CM, Tsai CL (1989). Regression and time series model selection in small samples. Biometrika. 76(2):297-307.
2. Amral N, Ozveren CS, King D (2007). Short term load forecasting using multiple linear regression. In 2007 42nd International universities power

engineering conference. (pp. 1192-1198). IEEE.
3. Liu H, Mi X, Li Y (2018). Smart multi-step deep learning model for wind speed forecasting based on variational mode decomposition, singular spectrum analysis, LSTM network and ELM. Energy Conversion and Management. 159:54-64.
4. Nagaraju TV, Prasad CD, Murthy NG (2020). Invasive weed optimization algorithm for prediction of compression index of lime-treated expansive clays. In Soft Computing for Problem Solving (pp. 317-324). Springer, Singapore.
5. Pradeep Kumar D, Ravi V (2017). Forecasting financial time series volatility using particle swarm optimization trained quantile regression neural network. Applied Soft Computing. 58:35-52
6. Kurumatani, K. (2020). Time series forecasting of agricultural product prices based on recurrent neural networks and its evaluation method. SN Applied Sciences, 2(8), 1-17.
7. Zhang, D., Chen, S., Liwen, L., & Xia, Q. (2020). Forecasting agricultural commodity prices using model selection framework with time series features and forecast horizons. *IEEE Access*, *8*, 28197-28209.
8. Valsamis EM, Husband H, Chan GK (2019). Segmented Linear Regression Modelling of Time-Series of Binary Variables in Healthcare. Computational and Mathematical Methods in Medicine.
9. Singh V, Poonia RC, Kumar S, Dass P, Agarwal P, Bhatnagar V, Raja L (2020). Prediction of COVID-19 corona virus pandemic based on time series data using Support Vector Machine. Journal of Discrete Mathematical Sciences and Cryptography. 1-5.
10. Lu, C. J., Lee, T. S., & Chiu, C. C. (2009). Financial time series forecasting using independent component analysis and support vector regression. Decision support systems, 47(2), 115-125.
11. Shen YX, Yang JG (2012). Temperature measuring point optimization and thermal error modeling for NC machine tool based on ridge regression. Mach Tool Hydraulic. 40(5):1-3.
12. Yuan K, Liu J, Yang S, Wu K, Shen F (2020). Time series forecasting based on kernel mapping and high-order fuzzy cognitive maps. Knowledge-Based Systems. 206:106359.
13. Niu, T., Wang, J., Lu, H., Yang, W., & Du, P. (2020). Developing a deep learning framework with two-stage feature selection for multivariate financial time series forecasting. Expert Systems with Applications, 148, 113237.
14. Li, S., Liu, X., & Lin, A. (2020). Fractional frequency hybrid model based on EEMD for financial time series forecasting. *Communications in Nonlinear Science and Numerical Simulation*, *89*, 105281.
15. Yu, H., Ming, L. J., Sumei, R., & Shuping, Z. (2020). A hybrid model for financial time series forecasting—integration of EWT, ARIMA with the improved ABC optimized ELM. IEEE Access, 8, 84501-84518.

16. Sahoo BB, Jha R, Singh A, Kumar D (2019). Application of support vector regression for modeling low flow time series. KSCE Journal of Civil Engineering. 23(2):923-34.
17. Vidal A, Kristjanpoller W (2020). Gold Volatility Prediction using a CNN-LSTM approach. Expert Systems with Applications. 113481.
18. Wu, J., Zhou, T., & Li, T. (2020). A Hybrid Approach Integrating Multiple ICEEMDANs, WOA, and RVFL Networks for Economic and Financial Time Series Forecasting. Complexity, 2020.
19. Gupta D, Pratama M, Ma Z, Li J, Prasad M (2019). Financial time series forecasting using twin support vector regression. PloS one. 14(3):e0211402.
20. Rao PS, Varma GP, Prasad CD (2020). Identification of linear and non-linear curve fitting models using particle swarm optimization algorithm. In AIP Conference Proceedings. Vol. 2269, No. 1, p. 030040). AIP Publishing LLC

eISSN 1303-5150