6-2017

# Ensemble of Non-Classical Decomposition Models and Machine Learning Models for Stock Index Prediction

Dhanya Jothimani
*Department of Management Studies, Indian Institute of Technology Delhi India*, dhanyajothimani@gmail.com

Ravi Shankar
*Department of Management Studies, Indian Institute of Technology Delhi India*, ravi1@dms.iitd.ac.in

Prof. (Dr.) Surendra S. Yadav
*HOD, Department of Management Studies, Indian Institute of Technology, Delhi*, ssyadav@dms.iitd.ac.in

# Ensemble of Non-Classical Decomposition Models and Machine Learning Models for Stock Index Prediction

**Dhanya Jothimani**
Department of Management Studies
Indian Institute of Technology Delhi
dhanyajothimani@gmail.com

**Ravi Shankar**
Department of Management Studies
Indian Institute of Technology Delhi
ravi1@dms.iitd.ac.in

**Surendra S. Yadav**
Department of Management Studies
Indian Institute of Technology Delhi
ssyadav@dms.iitd.ac.in

## ABSTRACT

The paper presents an ensemble framework comprising of non-classical decomposition model and machine learning model for predicting the stock index. Firstly, the data is decomposed into various components using Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN). In the next step, the components are forecasted independently using a machine learning model, namely, Artificial Neural Network (ANN). The forecasted sub-series are aggregated to obtain the final forecasts. The framework was tested using weekly close price of Nifty. Performance measures and statistical test conclude that the performance of ensemble models is better than traditional ANN (without decomposition). The proposed ensemble framework integrates the advantages of both decomposition and machine learning models.

### Keywords

Financial time series prediction, Ensemble forecasting, CEEMDAN, ANN, Nifty

## INTRODUCTION

Characteristics of financial time series data such as non-linearity and non-stationarity attribute to complexity in predicting them. Various researchers have used statistical and computationally intelligent models to predict financial time series (Atsalakis and Valavanis 2009, 2013) but these models are not without their own limitations. Statistical methods such as Auto Regressive Moving Average (ARMA) and Auto Regressive Integrated Moving Average (ARIMA) work on the assumption that data is stationary and linear. Machine learning models such as Artificial Neural Network (ANN) and Support Vector Regression (SVR) suffer from the problem of over-fitting and are sensitive to parameter selection. Hence, prediction of financial time series, especially stock price is considered to be a tedious task. Slightest improvement in the prediction accuracy attracts researchers, investors and stock brokers equally.

The models that use multiple predictors than a single predictor are known as Ensemble Models and can improve forecasting accuracy (Opitz and Maclin 1999). There are two approaches. First approach is to determine appropriate predictor based on the characteristics of data. The second approach is to deconstruct the data into various components using decomposition model and forecast the components independently and then aggregate them back together.

The most commonly used classical model of decomposing the time series into trend, seasonal and random components, works best with linear time series. But it ignores the random component and leads to loss of information, thus, affecting the forecast accuracy (Theodosiou 2011). Recently, several signal processing techniques like Discrete Wavelet Transform (DWT) and Empirical Mode Decomposition (EMD) have been used for decomposing the series in time-frequency domain and time domain, respectively (Liu et al. 2012, Lahmiri 2014, Jothimani et al. 2015, 2016a, 2016b). These signal processing techniques are classified under non-classical decomposition models.

EMD, proposed by Huang et al. (1998), uses Huang-Hilbert Transform (HHT) to decompose a signal into a set of adaptive basis functions called Intrinsic Mode Functions (IMFs). Despite its advantages, EMD suffers from the limitation of mode-mixing problem. Mode mixing refers to a phenomenon where more than one IMF contains signals in a similar frequency

band or an IMF consists of signals spanning a wide band of frequency. To overcome this limitation, a variant of EMD was proposed, namely, Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) (Torres et al. 2011).

The paper presents a ensemble forecasting model, namely, CEEMDAN-ANN models to predict 1-step ahead forecasts for weekly Nifty price index, where the time series is first decomposed to various sub-series (namely, IMFs and residual component) using CEEMDAN. Then, the sub-series are predicted independently using ANN and are aggregated to obtain the final forecasts. The hybrid CEEMDAN-ANN model integrates the benefits of both decomposition and machine learning models.

The contributions of the paper are two-fold. Firstly, it illustrates the use of ensemble forecasting model to predict the stock index. Secondly, it demonstrates the advantages and limitations of variations of EMD as a data preprocessing technique.

## ENSEMBLE MODEL

The steps of ensemble model are enumerated below:

1. The original series is decomposed into a set of different sub-series using CEEMDAN.
2. Each sub-series is forecasted separately using ANN.
3. Forecasted sub-series are recombined to get aggregate forecast, which is then compared with the original series to calculate the error measures.
4. The proposed model is statistically tested using Wilcoxon-Signed Rank Test (WSRT).

## COMPLETE ENSEMBLE EMPIRICAL MODE DECOMPOSITION WITH ADAPTIVE NOISE (CEEMDAN)

The detailed procedure of CEEMDAN is as follows:

1. To the original series *X(t)*, add white noise $w^i[t](\sim N(0, 1))$ and decompose the combination by I realizations to obtain the first IMF of CEEMDAN using the following formula:

$$\overline{IMF_1}[t] = \frac{1}{I}\sum_{i=1}^{I} IMF_1^i[t]$$

2. Calculate first residual component by subtracting first IMF of CEEMDAN from the original series *X(t)*

$$r_1(t) = X(t) - \overline{IMF_1}[t]$$

3. Obtain second mode by decomposing realizations $r_1[t] + \varepsilon_1 E_1(w^i[t]), i = 1, 2, ..., I$

$$\overline{IMF_2}[k] = \frac{1}{I}\sum_{i=1}^{I} E_1(r_1[t]) + \varepsilon_1 E_1(w^i[t])$$

   $E_j(.)$ is an operator which produces the j-th mode of the signal obtained by EMD.

4. Calculate the $k^{th}$ residual component.

$$r_k[t] = r_{k-1}[t] - \overline{IMF}_k[t] \forall k = 2, 3, ..., K$$

5. Calculate I realizations of $r_k[t] + \varepsilon_k E_k(w^i[t]), i = 1, 2, ..., I$ to obtain $(k + 1)^{th}$ mode.

$$\overline{IMF}_{k+1}[t] = \frac{1}{I}\sum_{i=1}^{I} E_1(r_k[t]) + \varepsilon_k E_k(w^i[t])$$

6. Go to Step 4 for next k.

Repeat steps 4 to 6 till it is no longer feasible to decompose the obtained residual component.

The final residual component is defined as:

$$R(t) = X(t) - \sum_{k=1}^{K} \overline{IMF}_k$$

where, $K$ is the total number of modes.

The original series can be expressed as:

$$X(t) = \sum_{k=1}^{K} \overline{IMF}_k + R(t)$$

CEEMDAN provides complete decomposition and exact reconstruction of original data series. It is adaptive in nature.

## DATA AND METHODOLOGY

### Data

The data considered for the study consists of weekly close price of Nifty ranging from September 2007 to December 2015 covering a period of 8 years. Nifty is the benchmark index of Indian stock market and it consists of 50 companies covering 22 sectors.

### Methodology

In first phase, the data is decomposed using CEEMDAN. Both decomposition models yielded a total of seven relatively stationary IMFs along with residual component (Figure 1). In second phase, Artificial Neural Network (ANN) is used to predict each sub-series obtained (seven IMFs and one residual component) independently. The forecasts of these sub-series are then aggregated to obtain the final forecasts of the ensemble model.

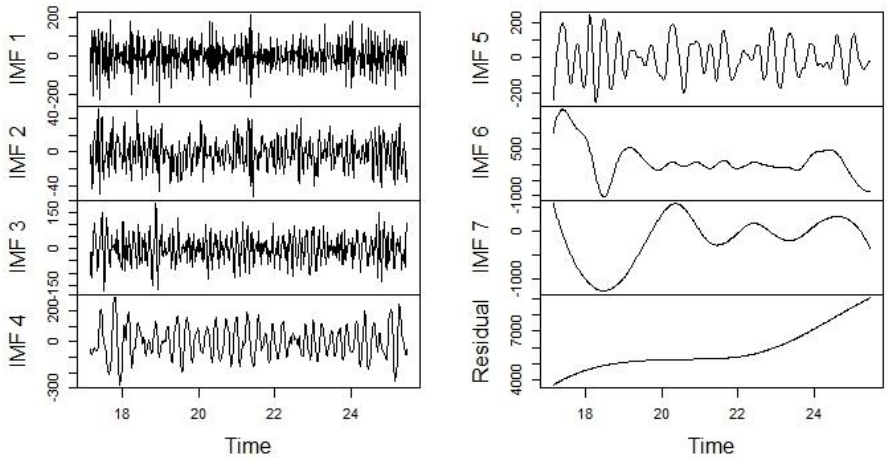


**Figure 1. Decomposed signals obtained using CEEMDAN**

A resilient three-layered feed forward neural network is used. This network consists of input layer, hidden layer and output layer. Each layer consists of certain number of neurons. The number of neurons in the input layer is determined using the relationship between data and its past values. This is also known as lag parameter, which is determined using Auto Correlation Function (ACF) and Partial Auto Correlation Function (PACF) plots. Figure 2 shows ACF and PACF plots of $IMF_3$ obtained using CEEMDAN. $IMF_3$ cuts off at lag 5, which means that at time *t*, it is dependent on its previous five values. Hence, the number of neurons in input layer is four and is expressed as:

$$X(t) = f[X(t-1), X(t-2), X(t-3), X(t-4), X(t-5)]$$

Based on performance of the neural network, the number of neurons in the hidden layer is determined iteratively. Number of neurons in the output layer is one since it is a regression problem. 70% of data is used as training dataset and remaining 30% is used as testing dataset since ANN is a supervised learning model. Resilient Backpropagation (RBP) algorithm is used for training the model as it quickens the training process (Liu et al. 2012). The data is normalized using z scores as a preprocessing step for ANN. It helps to reduce the chances of neural network getting stuck in local optima and fastens the training process (Wu and Lo 2010). To check the effectiveness of CEEMDAN-ANN model, 1-step ahead forecasts were also obtained using traditional ANN model (without decomposing the series).

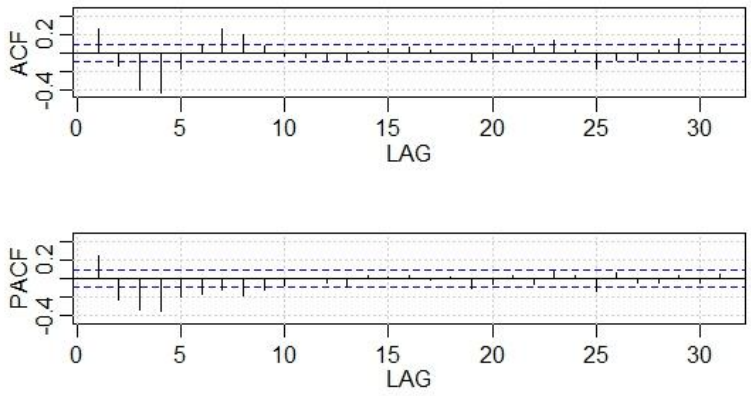


**Figure 2. ACF and PACF of $IMF_3$**

## RESULTS AND DISCUSSION

### Performance Measures

The performances of ANN and CEEMDAN-ANN models are analyzed and compared using two measures: (i) Root Mean Square Error (RMSE), and (ii) Directional Accuracy (DA). The difference between the actual value and the forecasted value is known as Error. Square root of mean of errors is known as Root Mean Square Error. Smaller the RMSE value, better is the forecast accuracy of the model. The number of times the forecasted values match the direction of original series is known as Directional Accuracy and is expressed in percentage. Higher the value of DA, better are the forecasts.

From Table 1, it can be seen that the performance of CEEMDAN-ANN is better than ANN models. It can be concluded that ensemble models produced better forecasts than traditional ANN model.

| | RMSE | DA(%) |
|---|---|---|
| CEEMDAN+ANN | 50.35 | 89 |
| ANN | 165.38 | 40 |

**Table 1. Performance Measures**

### Statistical Test

The results are statistically analyzed using Wilcoxon Signed-Rank Test (WSRT). WSRT is a non-parametric and distribution-free technique used for evaluating the predictive capabilities of two different models (Diebold and Mariano 1995).

Two-tailed WSRT was carried out on RMSE values and z value obtained was -5.780 at 99% confidence level ($\alpha = 0.01$). Since z statistics value is beyond (-1.96, 1.96), hence the null hypothesis of two models being same is not accepted. The WSRT results confirm that the ensemble forecasting model outperformed the traditional ANN model.

## CONCLUSION

An ensemble forecasting models integrating non-classical decomposition model (Complete Ensemble Empirical Mode Decomposition with Adaptive Noise) and machine learning model (Artificial Neural Network) was presented in this paper. The decomposition model was used to deconstruct the series into various components and each component was predicted independently using Artificial Neural Network. The presented model was tested on Nifty data. The presented model helped to overcome the limitations of classical decomposition and traditional computationally intelligent techniques.

Performance measures and the statistical tests confirm the superior performance of ensemble model over the base model (traditional ANN model). It can be concluded that pre-processing of data using decomposition model helped to improve the forecast accuracy of ANN model. Further, CEEMDAN overcame the limitation of mode mixing problem of EMD. As a part of future direction, the model can be tested for high frequency intraday stock index data. Performance of the presented model can be compared with CEEMDAN-SVR model.

## REFERENCES

1. Atsalakis, G. and Valavanis, K. (2009) Surveying stock market forecasting techniques- Part II: Soft computing methods. *Expert Systems with Applications*, 36,3,5932 - 5941.
2. Atsalakis, G. and Valavanis, K. (2013) Surveying stock market forecasting techniques- Part I: Conventional methods. Zopounidis C, ed., Computation Optimization in Economics and Finance Research Compendium, 49-104 (New York: Nova Science Publishers, Inc).
3. Jothimani, D., Shankar, R. and Yadav, S.S. (2015) Discrete wavelet transform-based prediction of stock index: A study on National Stock Exchange fifty index. *Journal of Financial Management and Analysis*, 28 (2), 35-49
4. Jothimani, D., Shankar, R. and Yadav, S.S. (2016a) A hybrid EMD-ANN model for stock price prediction. Swarm, Evolutionary, and Memetic Computing: SEMCCO 2015, Revised selected papers, 60-70 (Switzerland: Springer International Publishing).
5. Jothimani, D., Shankar, R. and Yadav, S.S. (2016b) A comparative study of ensemble based forecasting models for prediction of stock index. MWAIS 2016 Proceedings, Paper 5. Available at: http://aisel.aisnet.org/mwais2016/5/
6. Diebold, F.X. and Mariano, R.S. (1995) Comparing predictive accuracy. *Journal of Business and Economic Statistics*, 13,253-265.
7. Huang, N., Shen, Z., Long, S., Wu, M., Shih, H., Zheng, Q., Yen, N., Tung, C. and Liu, H. (1998) The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 454, 1971, 903- 995.
8. Lahmiri, S (2014) Wavelet low- and high-frequency components as features for predicting stock prices with backpropagation neural networks. *Journal of King Saud University - Computer and Information Sciences*, 26,2,218- 227.
9. Liu, H., Chen C., Tian, H. and Li, Y. (2012) A hybrid model for wind speed prediction using empirical mode decomposition and artificial neural networks. *Renewable Energy*, 48,545-556
10. Opitz, D. and Maclin, R. (1999) Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, 11,169-198.
11. Theodosiou, M. (2011) Forecasting monthly and quarterly time series using STL decomposition. *International Journal of Forecasting*, 27,4,1178 - 1195.
12. Torres, M.E, Colominas, M.A., Schlotthauer, G. and Flandrin, P. (2011) A complete ensemble empirical mode decomposition with adaptive noise. *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4144 - 4147.
13. Wu, G. and Lo, S. (2010) Effects of data normalization and inherent-factor on decision of optimal coagulant dosage in water treatment by artificial neural network. *Expert Systems with Applications*, 37, 7, 4974 - 4983.