LSTM Framework Design and Volatility Research on Intelligent Forecasting Model for Solving the Parallel Dislocation Problem

Yiran Weng¹, *, Zhiyi Wang², Longzhen Zhou³

¹School of Economics, Henan University, Kaifeng, Henan, 475000, China
²School of Middle East Studies, Beijing Language and Culture University, Beijing, 100083, China
³College of Business Administration, Dongbei University of Finance and Economics, Dalian 116000, China

*Corresponding author: wengyiran@henu.edu.cn

Abstract. The yield of treasury bonds is the benchmark interest rate in the financial market which is worth predicting and judging. Based on the Long Short-Term Memory (LSTM) neural network model in deep learning, combined with the vector autoregression method (VAR), this paper creatively constructs the VAR-LSTM framework and uses the predicted values of macroeconomic variables and lagged value of the time sequence as input factors to solve the problem of "parallel dislocation" of the fitting results of the traditional LSTM model which significantly improves the prediction accuracy. In order to meet the requirements of active quantitative investment for high precision prediction of stock market index, adaptive noise complete ensemble empirical mode decomposition (EMD) is introduced into the modeling of stock market index prediction. Combined with the efficient modeling ability of long-term and short-term memory network for medium- and long-term dependence of complex series, using the idea of "decomposition-reorganization-prediction-integration", an integrated prediction method of stock market index CEEMDAN-LSTM is proposed. CEEMDAN is used to decompose and reconstruct the index to obtain its high and low frequency components and trend items. The LSTM prediction models of each component are constructed respectively and the IMFs reorganization mode of high frequency subseries is optimized. Then the overall predicted value of the index is obtained by adding and integrating the predicted values of each component. Taking five representative stock market indexes as test data, the prediction results of CEEMDAN-LSTM and mainstream financial time series machine learning modeling methods are compared systematically. The results show that for treasury bond yield series, the prediction accuracy of ARIMA model is higher than that of general LSTM method, while VAR-LSTM model is better than ARIMA model. The prediction error in the training set and the test set is reduced by about 55% and 50% respectively, and the prediction accuracy of the change direction is improved by about 5% and 8% respectively, which has higher application value. The prediction performance of CEEMDAN-LSTM is consistently better than that of existing modeling methods, and has less prediction error and lower lag.
Keywords: Quantitative investment, Memory network, Deep learning, VAR-LSTM

1. Introduction

Stock market is one of the most important ways for listed companies to raise social funds. Stock investment has become one of the main ways for investors to maintain and increase the value of assets [1]. In the research of stock market investment, the analysis and modeling of asset price behavior is an important topic that researchers pay close attention to. For active stock investment research, the core of price behavior research is to effectively predict the trend and future value of stock prices, and then guide investors' trading decision-making behavior, so as to optimize the risk-adjusted returns of their portfolio [2]. However, faced with a rapidly changing stock market with complex state information, how to see the essence through the complex phenomena, grasp the stock market context and asset price movement status and trend to obtain ideal even excess investment income return through making contextual trading decisions, is a core subject that stock investors especially institutional investors pay high attention to and is worth in-depth study [3].

Using statistical methods to predict the rate of return on financial assets has always been a hot topic in the field of asset pricing [4]. Traditional forecasting models of bond yield can be roughly divided into macro factor model, autoregressive econometric model, term structure model, and others. The macro factor model uses the correlation between the macro economy and the national debt yield rate to predict the national debt yield rate with macro factors [5]. Autoregressive econometric models are dominated by ARIMA models [6], and N-S model is the representative of related studies on term structure models. However, the traditional econometric model is not ideal for the main reason that the traditional method has strict assumptions on data distribution and it takes the linear relationship between asset price changes and independent variables into consideration. Deep learning technology has developed rapidly since proposed, and its nonlinear modeling ability for complex data has been widely acknowledged. Among them, because the neural network algorithm only needs to fit the predicted value according to the characteristics of the input data, it is a simple mapping relationship between input and output. This mechanism can well adapt to the characteristics of the financial market with rapid changes and a complex data structure. In recent years, the Long Short-Term Memory (LSTM) neural network in deep learning has achieved great success. It has the ability to remember long-term and short-term information and is especially suitable for processing financial time series data. Therefore, there are many research literatures on the prediction of return on assets by LSTM model.

2. CEEMDAN decomposition and refactoring

CEEMDAN is a noise-assisted data analysis method for the deficiency of empirical mode decomposition (EMD) and EEMD. EMD, as an adaptive time-frequency signal processing method can be used to deal with the nonlinear and non-stationary signal analysis. Its characteristic is to stabilize and extract the signal in different scale fluctuation model, and in turn generate local characteristics of a series of data sequences with different time scales among which every sequence is an intrinsic mode function (IMF). The basic idea of EMD decomposition is to use the average value of the lower envelope to determine the "instantaneous equilibrium position" and then extract IMF. It includes the following four steps:

(1) Identify all maximum points Max and minimum points Min in $S(t)$, and draw upper and lower envolvements respectively with cubic spline interpolation method. $S(t)$ represents the current sequence to be decomposed, and its value in this paper is the closing price sequence of the market index.

(2) Calculate the local instantaneous mean of upper and lower envelope at each moment to obtain the average envelope $m(t)$. Calculate the new sequence $d(t)$ according to Equation (1).

$$d(t) = S(t) - m(t)$$ (1)

Then, the SD value is calculated according to Equation (2) to judge whether $D(t)$ is an eigenmode function.
\[ S_d = \frac{\sum_{i}^T |d_i(t) - d_{i-1}(t)|^2}{\sum_{i}^T d_i^2(t)} \] (2)

\( d_i(t) \) is the result of the ith screening. The SD threshold will be set and its value will be from 0.2 to 0.3. If the SD value is less than the threshold value, the screening process will be stopped. Otherwise, \( D(t) \) will be taken as the new sequence \( S(t) \) to be decomposed and re-execute the above iterative processing.

(3) If \( D(t) \) meets the two conditions required for the establishment of IMF, \( D(t) \) will be an IMF. If \( D(t) \) is separated from \( S(t) \), the remainder term \( R(t) = S(t) - D(t) \) will be obtained.

(4) If the remaining term \( R(t) \) has become a monotone function or a constant, or the amplitude is lower than the established threshold and IMF cannot be further extracted, the whole decomposition process is over. Otherwise, \( R(t) \) will be taken as the sequence \( S(t) \) to be decomposed. Then it will return to Step (1), and the iteration process above will be repeated.

3. Model building process

3.1. GARCH Model

The GARCH model is structured as follows:

\[ x_t = f(t, x_{t-1}, x_{t-2}, x_{t-3}, \ldots) + \varepsilon_t, \varepsilon_t = \sqrt{h_t}, h_t = \]
\[ w + \sum_{i=1}^p h_{t-i} + \sum_{j=1}^q \lambda \varepsilon^2_{t-j}, \varepsilon_t = e_t \sqrt{h_t} \] (3)

Set \( r_t \) for logarithm yield of HUSHEN 300 index in t time \((r_t = \log p_t - \log p_{t-1}, p_t)\) for the day's closing price. Set \( p_{t-1} \) for the previous day's closing price. Assume the known information of a given time \( t-1 \) set of \( F_{t-1} \) of the conditional average \( \mu_t = E(r_t | F_{t-1}) \) and conditional variances \( h_t = Var(r_t | F_{t-1}) = Var(\varepsilon_t | F_{t-1}) \). It follows the distribution of condition mean \( \mu_t \) and conditional variance \( h_t \), and the perturbation term at time \( t \) is \( \varepsilon_t \). When \( \{R_T\} \) is a stationary series, according to the GARCH model, the perturbation term \( \varepsilon_t \) meets the following conditions:

\[ \varepsilon_t = r_t - \mu_t, h_t = \omega + \sum_{i=1}^p \eta_i h_{t-i} + \sum_{j=1}^q \lambda \varepsilon^2_{t-j}, e_t = e_t \sqrt{h_t} \] (4)

3.2. The VAR-LSTM Framework of yield Forecast

The main purpose of LSTM model is to estimate parameter \( \theta \) and define the form of \( f(\cdot) \) by continuously iterating and optimizing the loss function. The research object of this paper is the treasury rate series, which is influenced by many factors. In addition to the autocorrelation effect of the series, it is also influenced by the macro economy. Therefore, the input factor of the model can be divided into lagged value of the time sequence and macro-economic indicators.
Handle general macroeconomic variables using VAR model, set up \( M_t = \{m_{t_1}, m_{t_2}, \ldots, m_{t_p}\}^T \) moment for TPD vector macroeconomic indicators and the lagged value behind the VAR model of order for:

\[
M_{t+1} = A + BM_t + \varepsilon_{t+1}
\]

(5)

However, when macro factors are used to predict the return rate, in essence, the macro indicators at time \( t \) have a strong "explanatory power" to the return rate series at time \( t \), and the macro indicators at time \( t + 1 \) should be used to predict the return rate at time \( t + 1 \). Otherwise, the phenomenon of "translation dislocation" is likely to occur. Therefore, in order to better use historical data to predict future sequence, this paper yields sequence factor and macro factor, establishes the VAR-LSTM framework, \( M_t = \{m_{t_1}, m_{t_2}, \ldots, m_{t_p}\}^T \) moment for TPD vector macroeconomic indicators and yields \( \{x_t\} \) (a time series with \( t \) as the time index). \( f(\cdot) \) is a function in the form of VAR - LSTM model, of which the model parameter is \( \theta \), and the input factors are the lag of the predicted values and the next moment yields of macro variables \( \hat{M}_{t+1} \) predicted by VAR method:

\[
\hat{x}_{t+1} = f(\hat{M}_{t+1}, x_t : \theta)
\]

(6)

Based on this VAR-LSTM framework, the existing macro variable data and national debt yield sequence can be used to accurately predict the next yield data.

4. Selection and verification of modeling data

4.1. Data selection

The framework is based on the illustration of the stock index prediction modeling in the process of selecting the csi 300 index as the modeling data base the reasons of which include: Firstly, the HUSHEN 300 index in HUSHEN two city is highly representative, on the basis of 300 stocks including most blue chips in the a-share market, the data cover a balanced and reasonable market of which the value accounts for about sixty percent of the a-share market so that it can accurately reflect the whole market of HUSHEN two city stock price’s changes and trends. Secondly, the return rate of the index is one of the important benchmarks to evaluate the performance of the stock portfolio investment, which can provide the basic conditions for the index investment in the market and the innovation of index derivatives. Therefore, the research on the prediction modeling of the CSI 300 index is also of great significance to the investment research of the derivatives market.

![Figure 1. CSI 300 index sequence](image)

Ljung-box statistics are used to test the ARCH effect of the index return series. The results show that when the lag order exceeds 4, the P value is far less than 0.05, indicating that the return series had
significant volatility and aggregation. The CSI 300 index series is non-stationary and contains a lot of noise, while the traditional econometric models such as ARIMA and GARCH are difficult to carry out high-precision prediction modeling for such complex financial timing series without efficient noise reduction processing. Therefore, it is reasonable and necessary to introduce CEEMDAN to carry out adaptive decomposition, denoising and reconstruction of the index in this paper and use the nonlinear time series modeling method LSTM to carry out prediction modeling of the index.

4.2. **CEEMDAN decomposition and recombination of exponential sequences**

CEEMDAN decomposition of exponential sequences. According to CEEMDAN mentioned above, the HUSHEN 300 index sequence adaptive decomposition get lined down on 10. The IMF and a residual term of the horizontal axis show exponential time serial number. The frequency of the longitudinal axis of each IMF, from IMF1 to IMF10 gradually declines to residual term frequency. The change model is simpler than the original sequence.

![CEEMDAN decomposition results](image)

In order to appropriately reduce the complexity of LSTM prediction modeling and avoid model overfitting, the 10 IMF generated by CEEMDAN decomposition of CSI 300 index are restructured according to the research of Li Helong et al. using the IMF restructuring method described above. The t-test with the mean value of 0 is conducted on IMF1 to IMF10 successively, and the test results show that IMF5 is the first eigenmode function with a P value less than 0.05, that is, the mean value of IMF5 is significantly different from 0. Therefore, IMF1 to IMF4 and IMF5 to IMF10 are reorganized into high frequency components of the index, and IMF5 to IMF10 are reorganized into low frequency components of the index. The residual term is taken as the trend term \( R(t) \) of the index, so as to obtain the three subsequences that depict the change pattern of the original index sequence from different frequency perspectives. The variation patterns of subsequences are relatively simple and regular, which is convenient for further fully extracting the fluctuation characteristics of each subsequence.
In essence, the factor information of the current period has a stronger explanatory power on the current rate of return, but it is unable to effectively predict the next period data. However, with the use of VAR model to predict the macro indicators, the explanatory power of its predicted value to the next period of return rate data is significantly improved, so the VAR-LSTM achieves the optimal effect.

5. Conclusions
The yield rate of national debt is the interest rate anchor of financial asset pricing, and its prediction is conducive to grasp the trend of asset prices in the financial market. Deep learning technology is developing rapidly, and its ability to construct nonlinear modeling of complex data is obvious to all, so it is believed to have a broad application prospect in the financial market. In view of this, this paper uses the deep learning LSTM model to model and predict the yield to maturity of 10-year treasury bonds. Considering that the yield of treasury bonds has autocorrelation effect and is inseparable from macroeconomic conditions, the input factors are determined as the lagged term and macroeconomic indicators of the yield series. Since macro factors have a stronger explanatory power to the current rate of return, this paper uses the VAR model to predict the macro variables forward, and puts the predicted value of macro variables and the lag term of the return series into the LSTM model for optimization, and finally constructs the VAR-LSTM framework. The experimental results confirm that the prediction performance of CEEMDAN-LSTM is consistently better than that of the existing modeling methods, and the prediction results of CEEMDAN-LSTM have less errors, higher accuracy, and lower time lag than the real index values. However, in the modeling process of CEEMDAN-LSTM in this paper, the selection of some parameters such as the number of hidden layers in the network is still subjective to some extent. At the same time, the double-layer LSTM element may not be able to find out the whole deep change pattern information contained in the nonlinear complex exponential sequence, so it is necessary to further study the optimization processing of model parameters.

References
[1] ENGLE R F. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation [J]. Econometrica: Journal of the Econometric Society, 1982, 50(4): 987-1007.
[2] BLLERSLEV T. Generalized autoregressive conditional heteroscedasticity [J]. Journal of the Econometrics, 1986, 31(3): 307-327.

[3] MIYAKOSHI T. Spillovers of stock return volatility to Asian equity markets from Japan and the US [J]. Journal of International Financial Markets, Institutions & Money, 2003, 13(4): 383-399.

[4] Estrella A, Hardouvelis G A. The Term Structure as a Predictor of Real Economic Activity [J]. The Journal of Finance, 1991, (46).

[5] Zhou L. An Empirical Study on the Correlation between the Term Structure of Chinese Government Debt Interest Rate and Macroeconomy - Based on Dynamic Nelson-Siegel Model [J]. Journal of Liaoning University (Philosophy and Social Sciences), 2019, 47(3).

[6] Roondiwala M, Patel H, Varma S. Predicting Stock Prices Using LSTM [J]. International Journal of Science and Research (IJSR), 2017, 6(4).