

A modified LSTM neural network model for stock price prediction based on auxiliary information and model average

Xiangyao Liu *

School of management, Xiamen University, Xiamen, Fujian, 361005

* Corresponding Author Email: 20420202200371@stu.xmu.edu.cn

Abstract. LSTM neural network has been widely used in stock price forecasting, but it has some problems such as difficulty in selecting superparameters, insufficient information of original price sequence, etc. Based on this, the model propose an improved method based on functional principal component analysis and distance covariance weighted model averaging. On the one hand, the method solves the problem that the original stock price information may not be sufficient to predict future price trends by using the functional principal component basis expansion coefficient of the intraday price to capture the residual sequence between the LSTM neural network predicted value and the real stock price. On the other hand, for the superparameter selection of LSTM neural network, the method uses a model averaging method based on distance covariance, which effectively balances the variance and bias of the prediction model. Furthermore, the regression method between the residuals sequence and the basis expansion coefficient of the functional principal component is model free. The actual data analysis shows that the proposed method has a greater improvement in MSE and BE than the original LSTM. Finally, the proposed method can be viewed as an optimization framework. This method can also modify ARIMA, GARCH and other models, and it can also be applied to other time series data prediction background.

Keywords: Stock price forecast, LSTM, FPCA, model average.

1. Introduction

Stock investment is an important part of our economy and an important tool for efficient allocation of social resources. Stock price forecasting is an important research direction in the field of financial forecasting. Predicting the trend of stock price accurately can help investors to carry out investment activities and stop earnings and losses in time. Stock price data is a typical financial time series data with strong seriality and periodicity. At the same time, stock price data has the characteristics of high frequency, high noise, non-stationarity, so it is difficult to predict the time series.

For the problem of stock price forecasting, the traditional forecasting method is time series analysis model. For example, on the premise of satisfying the stationarity assumption, the autoregressive model (AR) also has good prediction effect on linear time series (Zhao 2022 [1]), while the moving average model (MA) has the characteristics of processing nonlinear time series (Zhu 2022 [2]). The Autoregressive Integrated Moving Average model (ARIMA), which combines the two models, can realize short-term stock price prediction (Zhang 2021 [3]). However, the stock price data has the characteristics of volatility and aggregation, that is, when the data is stable in the long term but the short-term variance is unstable, it is necessary to use other time series prediction models on the basis of ARIMA model for correction, such as ARIMA model based on GARCH correction (yin2022 [4]). However, the traditional time series prediction method is stricter on the model assumption and is only applicable to short-term time series prediction, it cannot effectively deal with complex stochastic systems. Based on this, a time series prediction model (LSTM) called Long Short-Term Memory can effectively improve the accuracy of stock price prediction (li2022 [5]), for example, the stock price prediction of Bank of Beijing (li2022[6]) because it is suitable for processing or delaying important events (Hatami2022 [7]). However, when using LSTM neural network in predicting, there are two problems: the original time series information is not enough to predict the future trend and the selection of super parameters. For the former, information such as intraday price data or other fundamental data can be used as auxiliary information. In particular, for intraday price data, dimension reduction extraction is required during processing. Common dimension reduction methods

include principal component analysis (PCA) (Aftowicz [8]), kernel principal component analysis (KPCA) (Lu2022[9]). However, these dimension reduction methods are only applicable to multivariate data analysis and are not applicable to time series data, In consideration of the high correlation between the observation time points of the time series data and the time series can be regarded as a discrete realization of the random process, we can use B-spline basis function, Fourier basis function, Gaussian basis function, functional principal component analysis (Jarry2022 [10]) and other basis function smoothing methods to extract the functional features of the intraday price data while reducing the dimension. For the latter, the method of model averaging can be used to determine the super parameters.

In summary, this paper proposes a modified LSTM neural network model based on auxiliary time series information and distance covariance weighted model average. The innovations of the proposed method are as follows: on the one hand, the intraday price series is used as auxiliary information and the function features of the intraday price series are extracted; on the other hand, the potential model structure between the residual series and the function features is approximated by bagging framework In addition, for the super parameters of LSTM neural network, the distance covariance weighted model average method is used to determine, and the variance and deviation of the prediction model are weighed together with the bagging framework in the residual fitting stage. Specifically, for a single LSTM neural network model, we use the intraday price information as auxiliary information and reduce the dimension of the intraday price time series through functional principal component analysis to obtain the principal component basis coefficient matrix, then take the principal component basis expansion coefficient as the prediction variable, and the prediction residual sequence between the single LSTM model and the real stock price as the response variable, The fitting residual sequence is obtained by bagging regression and added with the original LSTM model prediction value to obtain a new prediction value. For the above process, we act on the LSTM models with different parameter settings to obtain the prediction matrix of multiple LSTM models. Then, the distance covariance is used to measure the fitting accuracy between the predicted value of each LSTM model and the real value, and the weight vector is normalized to obtain the weight vector. Finally, the weight vector is used to weight the prediction matrix fitted by the LSTM multi model to obtain the final predicted value. The empirical results show that this method has obvious correction effect on LSTM neural network stock price prediction.

2. Theory and method

2.1. LSTM

The Long-Short Term Memory (LSTM) is a special type of RNN model. The appearance of LSTM is to overcome the information flooding (short-term memory) of the traditional RNN neural network model and the problems of gradient disappearance and gradient explosion in training (Mao 2022 [11]). LSTM model is suitable for processing and predicting important events with long and interval in time series. It is mainly used to describe the relationship between the current data and the previous input data, and it uses its input ability to save the state information before the input network, and make it affect the subsequent exact value and development trend. The structure of LSTM neural network is shown in the figure:

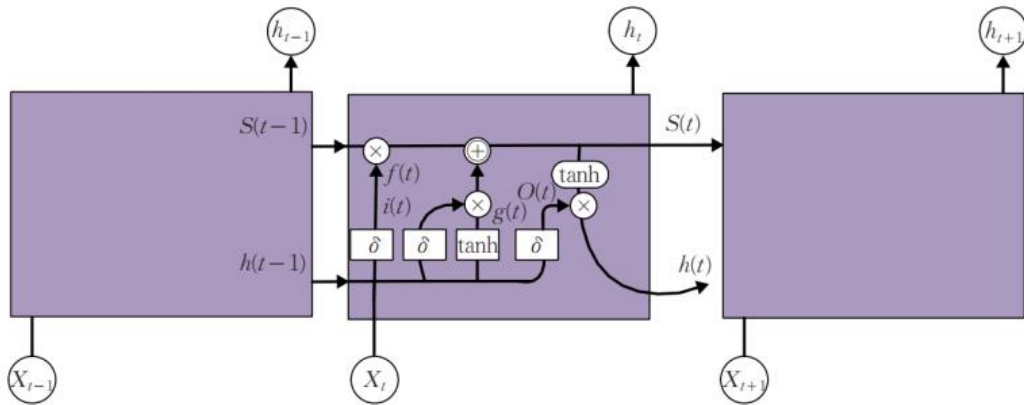


Figure 1. LSTM Neural network structure.

LSTM creatively adopts control gate mechanism, which is composed of memory cell, input gate, output gate and forgetting gate. Each gate can play the role of information selection. These gates enable the self-cycle weight to change, and control the information of the previous unit to pass through and pass to the next unit. The calculation principle of the control door is described below. First, calculate the value of the input gate i_t and the candidate state value $\tilde{C}_t$ of the cell at the time t (Gers2000 [12]). The formula is as follows:

$$i_t = \delta(W_i \times (X_t, h_{t-1}) + b_i) \quad (1)$$

$$\tilde{C}_t = \tanh(W_c \times (X_t, h_{t-1}) + b_c) \quad (2)$$

Next, the activation value of the time forgetting gate is calculated as follows:

$$f_t = \delta(W_f \times (X_t, h_{t-1}) + b_f) \quad (3)$$

The cell status update value can be calculated from the above two steps, and the formula is as follows:

$$C_t = i_t \times \tilde{C}_t + f_t \times C_{t-1} \quad (4)$$

After the cell state update value is obtained, the value of the output gate can be further calculated. The formula is as follows:

$$O_t = \delta(W_o \times (X_t, h_{t-1}) + b_o) \quad (5)$$

$$h_t = O_t \times \tanh(C_t) \quad (6)$$

Through the above process, LSTM can effectively use input to make it have long-term memory ability.

2.2. Model averaging method based on distance covariance

Model averaging refers to averaging the estimates or forecasts from different models through certain weights, which is also called model combination. How to select the combination weight is the most important issue. The basic idea of the model average method based on distance covariance is to determine the weight of the model through the prediction effect of the model. The weight of the model with good prediction effect is larger, and the weight of the model with poor prediction effect is smaller (xie2022 [13]). The specific principle is described below. Firstly, the distance covariance between the real value and the predicted value of each model is measured to obtain the correlation coefficient matrix $[I_1, I_2, \dots, I_j]$. The formula of distance covariance and distance correlation coefficient is as follows:

$$I = \hat{d}corr(u, v) = \frac{\hat{d}cov(u, v)}{\sqrt{\hat{d}cov(u, u) \cdot \hat{d}cov(v, v)}} \quad (7)$$

While

$$\hat{d}^2 cov(u, v) = \hat{s}_1 + \hat{s}_2 - 2\hat{s}_3 \quad (8)$$

$$\hat{s}_1 = \frac{1}{n^2} \sum_i \sum_j \|u_i - u_j\|_{d_u} \|u_i - u_j\|_{d_v} \quad (9)$$

$$\hat{s}_2 = \frac{1}{n^2} \sum_i \sum_j \|u_i - u_j\|_{d_u} \frac{1}{n^2} \sum_i \sum_j \|u_i - u_j\|_{d_v} \quad (10)$$

$$\hat{s}_3 = \frac{1}{n^2} \sum_i \sum_j \sum_l \|u_i - u_j\|_{d_u} \|u_i - u_j\|_{d_v} \quad (11)$$

The greater the correlation I , the greater the contribution of the model to the better prediction effect, i. e. give it a higher weight, and record the model weight w_j . The calculation method is as follows:

$$w_j = \frac{I_j}{\sum_j I_j} \quad (12)$$

2.3. Functional principal component analysis

Functional principal component analysis (FPCA) is a generalization of traditional principal component analysis. By decomposing the eigenvalues of covariance matrix, the direction with the largest constraint condition is found. The weight vector of the traditional principal component corresponds to the weight function of the principal component (Cao 2022 [14]). Assuming that the original data has been fitted into a continuous curve $X_i(s), s \in T, i = 1, 2, \dots, n$ the principal component score of each observation curve is defined as:

$$\xi_i = \int_T \beta(s) x_i(s) ds, \quad i = 1, 2, \dots, n \quad (13)$$

The principal component is converted into the following optimization problem:

$$\max \frac{1}{n} \sum_i \xi_i^2 = \max \frac{1}{n} \sum_i \left(\int_T \beta(s) x_i(s) ds \right)^2 \quad (14)$$

$$s.t \begin{cases} \|\beta\|^2 = \int_T \beta(s) \beta(s) ds = 1 \\ \int_T \beta(s) \beta^l(s) ds = 0 \quad l = 1, \dots, k-1 \end{cases} \quad (15)$$

To solve the weight function of functional principal component analysis, it is necessary to obtain the weight function of functional principal component analysis through the sample covariance matrix function. The covariance function is:

$$v(s, t) = (n-1)^{-1} \sum_i x_i(s) x_i(t) \quad (16)$$

The characteristic equation of the functional principal component weight function $\beta(s)$ can be obtained as follows:

$$\int v(s, t) \beta(t) dt = \lambda_i \beta(s) \quad (17)$$

λ is the characteristic root of the covariance function, $\beta(s)$ is the characteristic function. Therefore, after transformation, the characteristic equation can also be expressed as: $V\beta(s) = \lambda_i \beta(s)$. Finally, the number of appropriate principal components is determined according to the cumulative contribution rate. Generally, the cumulative contribution rate is greater than 85% (Assel2022 [15]).

2.4. Proposed method

First, download the relevant data of the stock, and sort out the intraday price information and closing price information of the stock after data preprocessing. For the closing price information, the LSTM model with multiple super parametric values is used for training. At the same time, after reducing the dimension of intraday price information through functional principal component analysis, bagging fitting is carried out with the obtained basis expansion coefficient and the residual value after LSTM training. Then, the predicted values of the two syntheses are averaged according to the distance covariance to obtain the final predicted value. The process of the method is shown in the flow chart below:

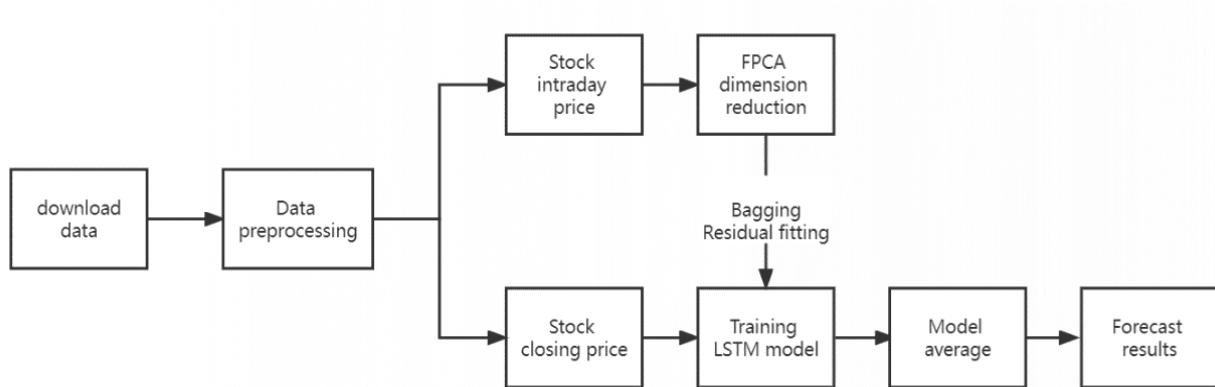


Figure 2. Flow chart of the method.

3. Data analysis and results

3.1. Data source and pre analysis

The data in this paper comes from the wind database (www.wind.com.cn), and three stocks are selected, namely, Xinchao energy (SH600777), Dongfang Group (SH600811) and Oupai furniture (SH603833). The industries they belong to are petroleum and petrochemical, commodity retail and light industry manufacturing. In order to test whether the methods used in this paper can modify the LSTM model under different volatility and complexity, these three stocks are randomly selected from all non-ST stocks of A-share listed companies in the wind database. We collated and collected their intraday price data and closing price data. It can be seen from the preprocessing of stock price data. This time series is a nonstationary time series. The autocorrelation and partial autocorrelation charts of the stock price data of the three stocks are as follows:

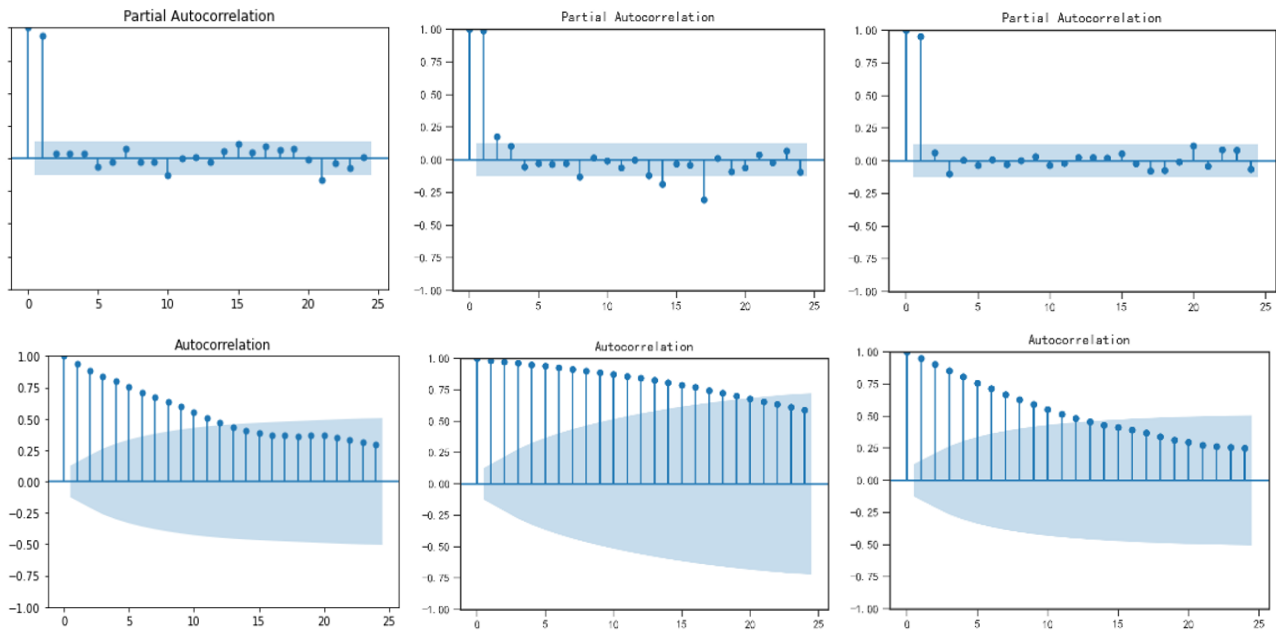


Figure 3. Autocorrelation and partial autocorrelation charts.

It can be seen from the graph analysis that although the partial autocorrelation graph is truncated, the autocorrelation graph is not truncated. At the same time, according to the calculation, the time series P values of the three stocks are all greater than 0.05. Therefore, the time series data of these three stocks are non-stationary time series.

3.2. Comparison results

This experiment is based on python programming language. In this paper, ten different training sets, with the volume of 175 to 220, are selected to test the robustness of the LSTM model. The experimental results are as follows:

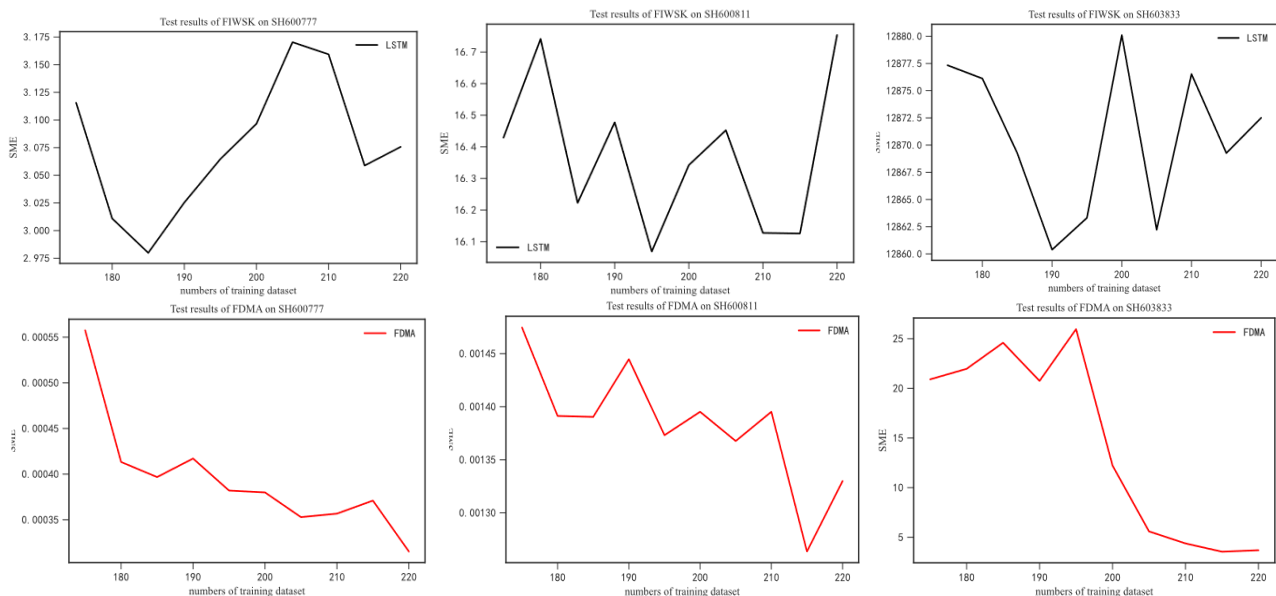


Figure 4. Experiment results.

It can be seen from the above figure that for any one of the three stocks, the prediction effect of the method used in this paper is far better than that of ordinary LSTM, which is reflected by the fact that the mean square error (MSE) predicted by this method is far less than that of ordinary LSTM. At the same time, it can be observed from the chart that the prediction result MSE of ordinary LSTM fluctuates up and down with the increase of the training set. However, the MSE of the method in this

paper has an obvious decreasing trend with the increase of the training set, and the prediction effect is getting better. The specific experimental results are shown in the following table:

Table 1. SH600777 Experiment Result.

TRAINING SET	FDMA-MSE	FDMA-MRE	FDMA-BE	LSTM-MSE	LSTM-MRE	LSTM-BE
175	0.000558	0.009602	0.156417	3.115565	0.997891	0.98579
180	0.000413	0.007925	0.135658	3.010949	0.982112	0.888805
185	0.000397	0.007598	0.133164	2.979871	0.977209	0.871813
190	0.000417	0.00727	0.136661	3.025352	0.984158	0.913124
195	0.000382	0.007209	0.130735	3.064674	0.990229	0.940746
200	0.00038	0.007032	0.130495	3.096559	0.99515	0.961186
205	0.000353	0.00681	0.126032	3.17034	1.005964	1.0409
210	0.000357	0.006833	0.126699	3.159344	1.004462	1.022643
215	0.000371	0.006789	0.12921	3.058781	0.989365	0.933312
220	0.000315	0.006541	0.119156	3.075702	0.99186	0.953089

Table 2. SH600811 Experiment Result.

TRAINING SET	FDMA-MSE	FDMA-MRE	FDMA-BE	LSTM-MSE	LSTM-MRE	LSTM-BE
175	0.001475	0.006852	0.062257	16.429	1.004269	1.012071
180	0.001391	0.006752	0.060558	16.74128	1.01273	1.043753
185	0.00139	0.006713	0.060566	16.22271	0.998506	0.993763
190	0.001445	0.006757	0.061699	16.47765	1.005539	1.018246
195	0.001373	0.00669	0.060161	16.0687	0.994188	0.979794
200	0.001395	0.00662	0.060668	16.34214	1.001835	1.004677
205	0.001368	0.006528	0.060063	16.45251	1.004816	1.016552
210	0.001395	0.00655	0.060668	16.12764	0.995812	0.985819
215	0.001264	0.006265	0.057734	16.12587	0.995754	0.985936
220	0.00133	0.006342	0.059236	16.75379	1.013221	1.041669

Table 3. SH603833 Experiment Result.

TRAINING SET	FDMA-MSE	FDMA-MRE	FDMA-BE	LSTM-MSE	LSTM-MRE	LSTM-BE
175	20.90484	0.019746	0.372507	12877.34	1.000359	1.001266
180	21.9653	0.020004	0.381603	12876.12	1.000312	1.001181
185	24.59442	0.020779	0.404416	12869.27	1.000061	1.000221
190	20.75022	0.019339	0.372613	12860.39	0.999733	0.999036
195	25.96112	0.020827	0.415978	12863.31	0.999841	0.999445
200	12.22269	0.015068	0.29241	12880.11	1.00046	1.001704
205	5.590882	0.010959	0.203076	12862.22	0.9998	0.999333
210	4.378073	0.009639	0.181778	12876.53	1.000333	1.000983
215	3.546554	0.008631	0.164157	12869.26	1.000064	1.000086
220	3.693606	0.008688	0.16754	12872.51	1.00018	1.000685

The above table tests the prediction effect of the model in this paper and the prediction effect of the general LSTM model from the three angles of mean square error (MSE), mean radial error (MRE) and posterior error (be). The test results show that the prediction effect of the model in this paper is good. The main reasons for the ideal prediction effect of the model in this paper are as follows: on the one hand, the intraday price series has played a good role in supplementing the stock price data, making up for the problem that the original time series data is not enough to predict. On the other hand, the bagging fitting between the residual sequence and the function characteristics obtained by dimension reduction is relatively successful. In addition, a variety of super parametric LSTM models combined by the model averaging method of distance covariance are combined with the bagging framework in the residual fitting stage to jointly weigh the variance and bias of the prediction model.

4. Conclusion and discussion

Based on the LSTM stock price prediction model, this paper proposes a modified LSTM neural network model based on auxiliary time series information and distance covariance weighted model average. For the original LSTM super parametric selection problem and the problem that the original time series is not enough to predict the future price trend. On the one hand, the model set several super parameters to train the LSTM model; on the other hand, the model use the daily price information as auxiliary information, and use functional principal component analysis to reduce the dimension and carry out bagging fitting with the residual of the LSTM model. Finally, the distance covariance is used to average the model to obtain the predicted value. Empirical analysis shows that the prediction effect of the method used in this paper is far better than the general LSTM model, and the correction effect is good.

In the future research, on the one hand, different function feature extraction methods can be adopted, such as B-spline basis function, Gaussian basis function, wavelet basis function, etc. On the other hand, we can compare whether different weighting measurement methods will affect the results. Thirdly, we can use similar methods to modify ARIMA and GARCH models and study their corresponding correction effects.

References

- [1] Zhao Zhiwen, Xu Shengnan. (2022). Statistical Inference of Periodic Random Coefficient Autoregressive Model. *Statistics & Decision* (4), 5.
- [2] Zhuang Li. (2022). Establishment and prediction of autoregressive integrated moving average model of monthly reported deaths of pulmonary tuberculosis in China. *Chinese Journal of Antituberculosis*, 44(4), 375-380.
- [3] Yang Yuyuan, Zhang Mei. (2021). Empirical Analysis of Stock Price Based on ARIMA Model. *Science & Technology Information*, 19(29), 4.
- [4] Lu Yin. (2022). Prediction of stock price based on ARIMA-GARCH model. *Advances in Applied Mathematics*, 11(1), 14.
- [5] Hatami, N, Cho, T. H, Mechtouff, L, Eker, O. F, Rousseau, D, & Frindel, C. (2022). Cnn-lstm based multimodal mri and clinical data fusion for predicting functional outcome in stroke patients.
- [6] LI Guicheng, XU Li, ZHANG Li. (2022). Stock price forecast analysis based on LSTM. *INTELLIGENT COMPUTER AND APPLICATIONS* (012-005).
- [7] Tao Li. (2022). Research on stock price prediction of Bank of Beijing, the leading City Commercial Bank -- Based on deep neural network model. *Trade Fair Economy* (12), 4.
- [8] Aftowicz, M, Kabin, I, Klann, D, Varabei, Y, Dyka, Z, & Langendoerfer, P. (2022). Horizontal sca attacks against kp algorithm using k-means and pca.
- [9] Lu, Y, Wu, H, & Zhou, B. (2022). Dimensionality reduction method of rotor fault data set based on KPCA-KNPE.

- [10] Jarry, G, Hassoumi, A, Delahaye, D, & Hurter, C. (2022). Interactive trajectory modification and generation with fpca. CEAS Aeronautical Journal, 13(2), 371-383.
- [11] Mao K, Zhang W, Wang D B, et al. Prediction of Depression Severity Based on the Prosodic and Semantic Features with Bidirectional LSTM and Time Distributed CNN[J]. 2022.
- [12] Gers, Felix, A, Schmidhuber, Jürgen, Cummins, & Fred. (2000). Learning to forget: continual prediction with lstm. Neural Computation.
- [13] Xie, J, Long, F, Lv, J, Wang, Q, & Li, P. (2022). Joint distribution matters: deep brownian distance covariance for few-shot classification.
- [14] Cao, X, & Liu, F. (2022). Exploring temporal patterns of urban management cases based on functional principal component data analysis. IOP Publishing Ltd.
- [15] Assel, H. V., Espinasse, T., Chiquet, J., & Picard, F. (2022). A probabilistic graph coupling view of dimension reduction.